In-Context Learning of a Linear Transformer Block: Benefits of the MLP Component and One-Step GD Initialization

Ruiqi Zhang
UC Berkeley
rqzhang@berkeley.edu

Jingfeng Wu
UC Berkeley
uuujf@berkeley.edu

Peter L. Bartlett
UC Berkeley and Google DeepMind
peter@berkeley.edu

Abstract

We study the *in-context learning* (ICL) ability of a *Linear Transformer Block* (LTB) that combines a linear attention component and a linear multi-layer perceptron (MLP) component. For ICL of linear regression with a Gaussian prior and a *non-zero mean*, we show that LTB can achieve nearly Bayes optimal ICL risk. In contrast, using only linear attention must incur an irreducible additive approximation error. Furthermore, we establish a correspondence between LTB and one-step gradient descent estimators with learnable initialization ($\mathsf{GD}\text{-}\boldsymbol{\beta}$), in the sense that every $\mathsf{GD}\text{-}\boldsymbol{\beta}$ estimator can be implemented by an LTB estimator and every optimal LTB estimator that minimizes the in-class ICL risk is effectively a $\mathsf{GD}\text{-}\boldsymbol{\beta}$ estimator. Finally, we show that $\mathsf{GD}\text{-}\boldsymbol{\beta}$ estimators can be efficiently optimized with gradient flow, despite a non-convex training objective. Our results reveal that LTB achieves ICL by implementing $\mathsf{GD}\text{-}\boldsymbol{\beta}$, and they highlight the role of MLP layers in reducing approximation error.

1 Introduction

The recent dramatic progress in natural language processing can be attributed in large part to the development of large language models based on *Transformers* [34], such as BERT [11], LLaMA [32], PaLM [8], and the GPT series [28, 29, 6, 25]. In some pioneering pre-trained large language models, a new learning paradigm known as *in-context learning* (ICL) was observed (see, for example, [6, 7, 24, 20]). ICL refers to the capability of a pre-trained model to solve a new task based on a few in-context demonstrations without updating the model parameters.

A recent line of work quantifies ICL in tractable statistical learning setups such as linear regression with a Gaussian prior (see [14, 3] and references thereafter). Specifically, an ICL model takes a sequence $(\mathbf{x}_1, y_1, ..., \mathbf{x}_M, y_M, \mathbf{x}_{\mathsf{query}})$ as input and outputs a prediction of y_{query}, where $(\mathbf{x}_i, y_i)_{i=1}^M$ and $(\mathbf{x}_{\mathsf{query}}, y_{\mathsf{query}})$ are independent samples from an unknown task-specific distribution (where the task admits a prior distribution). The ICL model is pre-trained by fitting many empirical observations of such sequence-label pairs. Experiments show that Transformers can achieve an ICL risk close to that achieved by Bayes optimal estimators in many statistical learning tasks [14, 3].

For ICL of linear regression with a Gaussian prior, the data is generated as

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0},\ \mathbf{H}), \quad y \mid \widetilde{\boldsymbol{\beta}}, \mathbf{x} \sim \mathcal{N}(\widetilde{\boldsymbol{\beta}}^\top \mathbf{x},\ \sigma^2),$$

where $\widetilde{\boldsymbol{\beta}}$ is a task parameter that satisfies a Gaussian prior, $\widetilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\mathbf{0},\ \psi^2\mathbf{I}\right)$. In this setup, [14, 3] showed in experiments that Transformers can achieve nearly optimal ICL by matching the performance of the Bayes optimal estimator, that is, an optimally tuned ridge regression. Besides, [35] showed by construction that a single *linear self-attention* (LSA) can implement *one-step gradient descent with zero initialization* ($\mathsf{GD}\text{-}0$), offering an insight into the ICL mechanism of the Transformer.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Later, theoretical works showed that optimal LSA models effectively correspond to GD-0 models [1], trained LSA models converge to GD-0 models [38], and GD-0 models (hence LSA models) can provably achieve nearly Bayes optimal ICL in certain regimes [37].

Our contributions. In this paper, we consider ICL in a more general setup, that is, linear regression with a Gaussian prior and a *non-zero mean* (that is, $\mathbb{E}[\widetilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}^* \neq 0$). The non-zero mean in task prior captures a common scenario where tasks share a signal. In this setting, we show that the LSA models considered in prior papers [1, 38, 37] incur an irreducible additive approximation error. Furthermore, we show this approximation error is mitigated by considering a *linear Transformer block* (LTB) that combines a linear multi-layer perceptron (MLP) component and an LSA component. Our results highlight the important role of MLP layers in Transformers in reducing the approximation error, and they suggest that theories about LSA [1, 38, 37] do not fully explain the power of Transformers. Motivated by this understanding, we investigate LTB in depth and obtain the following additional results:

- We show that LTB can implement *one-step gradient descent with learnable initialization*, referred to by us as GD-$\boldsymbol{\beta}$. Additionally, we show that every optimal LTB estimator that minimizes the in-class ICL risk is *effectively* a GD-$\boldsymbol{\beta}$ estimator.
- Moreover, we show that the optimal GD-$\boldsymbol{\beta}$, hence also the optimal LTB, nearly matches the performance of the Bayes optimal estimator for linear regression with a Gaussian prior and a non-zero mean, provided that the signal-to-noise ratio is upper bounded. These two results together suggest that LTB performs nearly optimal ICL by implementing GD-$\boldsymbol{\beta}$.
- Finally, we show that the GD-$\boldsymbol{\beta}$ estimator can be efficiently optimized by gradient descent with an infinitesimal stepsize (that is, gradient flow) under the population ICL risk, despite the non-convexity of the objective.

Paper organization. The remaining paper is organized as follows. We conclude this section by introducing a set of notations. We then discuss related papers in Section 2. In Section 3, we set up our ICL problems and define LTB and LSA models mathematically. In Section 4, we show a positive approximation error gap between LSA and LTB, which highlights the importance of the MLP component and motivates our subsequent efforts to study LTB. In Section 5, we connect LSA estimators to GD-$\boldsymbol{\beta}$ estimators that are more interpretable for ICL of linear regression. In Section 6, we study the in-context learning and training of GD-$\boldsymbol{\beta}$ estimators. We conclude our paper and discuss future directions in Section 7.

Notations. We use lowercase bold letters to denote vectors and uppercase bold letters to denote matrices and tensors. For a vector $\mathbf{x}$ and a positive semi-definite (PSD) matrix $\boldsymbol{A}$, we write $\|\mathbf{x}\|_{\boldsymbol{A}} := \sqrt{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}$. We write $\langle\cdot,\cdot\rangle$ for the inner product, which is defined as $\langle\mathbf{x},\mathbf{y}\rangle := \mathbf{x}^\top\mathbf{y}$ for vectors and $\langle\boldsymbol{A},\boldsymbol{B}\rangle := \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^\top)$ for matrices. We write $\boldsymbol{A}[i]$ as the i-th row of the matrix $\boldsymbol{A}$, $\boldsymbol{A}_{m,n}$ as the (m,n)-th entry, and $\boldsymbol{A}_{-1,-1}$ as the right-bottom entry. We write $\mathbf{0}_n, \mathbf{0}_{m\times n}, \mathbf{I}_n$ for the zero vector, zero matrix and identity matrix, respectively. We denote the Kronecker product as $\otimes$. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A}\otimes\boldsymbol{B}$ is a linear mapping which operates as $(\boldsymbol{A}\otimes\boldsymbol{B})\circ\boldsymbol{C} = \boldsymbol{B}\boldsymbol{C}\boldsymbol{A}^\top$. For a positive semi-definite matrix $\boldsymbol{A}$, we write $\boldsymbol{A}^{\frac{1}{2}}$ as its principle square root, which is the unique positive semi-definite matrix $\boldsymbol{B}$ such that $\boldsymbol{B}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{A}$. We also write $\boldsymbol{A}^{-\frac{1}{2}} = (\boldsymbol{A}^{\frac{1}{2}})^+$, where $(\cdot)^+$ denotes the Moore Penrose pseudo-inverse. For two sets A, B we write $A + B$ for the Minkowski sum, which is defined as $\{a + b : a \in A, b \in B\}$. We also write $a + B = \{a\} + B$ for an element a and a set B. We write $\mathsf{null}\,(\cdot)$ for the null set of a matrix or a tensor.

2 Related Works

Empirical results for ICL in controlled settings. The work by [14] first considered ICL in controlled statistical learning setups. For noiseless linear regression, [14] showed in experiments that Transformers match the performance of the optimal estimator (that is, *Ordinary Least Square*). Subsequent works by [3, 18] extended their result to noisy linear regression with a Gaussian prior and showed that Transformers can match the performance of the Bayesian optimal estimator (that is, an optimally tuned ridge regression). Besides, [30] showed that the above holds even when Transformers are pretrained on a limited number of linear regression tasks. These papers only considered a Gaussian prior with a zero mean. In contrast, we consider a Gaussian prior with a non-zero mean. Our setup better captures the common scenarios where the tasks share a signal.

The empirical investigation of ICL in tractable statistical learning setups goes beyond linear regression settings. For more examples, researchers empirically studied the ICL ability of Transformers for decision trees (see [14], they also considered two-layer networks), algorithm selection [4], linear mixture models [26], learning Fourier series [2], discrete boolean functions [5], representation learning [13, 16], and reinforcement learning [17, 19]. Among all these settings, Transformers can either compete with the Bayes optimal estimators or expert-designed strong benchmarks. These works are not directly comparable with our paper.

Transformer implements gradient descent. A line of work interpreted the ICL of Transformers by their abilities to implement *gradient descent* (GD) [35, 3, 10, 1, 38, 4, 37]. In experiments, [35, 10] showed that (multi-layer) Transformer outputs are close to (multi-step) GD outputs. When specialized to linear regression tasks, [35] constructed a single *linear self-attention* (LSA) that implements *one-step gradient descent with zero initialization* (GD-0). Subsequently, [1] showed that optimal LSA models effectively correspond to GD-0 models, [38] proved that trained LSA models converge to GD-0 models, [37] showed that GD-0 models (hence LSA models) can provably achieve nearly Bayes optimal ICL in certain regimes and provided a sharp task complexity analysis of the pre-training. These papers focused on LSA models. Instead, we consider a linear Transformer block that also utilizes the MLP layer.

From an approximation theory perspective, [3, 4] showed that Transformers can implement multi-step GD under general losses. In comparison, we consider a limited setting of linear regression and show the LSA models can implement one-step GD with learnable initialization (GD-$\boldsymbol{\beta}$), moreover, every optimal LTB model is effectively an GD-$\boldsymbol{\beta}$ model. Both of our constructions utilize MLP layers in Transformers, highlighting its importance in reducing approximation error. Different from their results, we also show a negative approximation result that reveals the limitation of LSA models.

3 Preliminaries

Model input. We use $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ to denote a feature vector and its label, respectively. Throughout the paper, we assume a fixed number of context examples, denoted by $M > 0$. We denote the context examples by $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{M\times d} \times \mathbb{R}^M$, where each row represents a context example, denoted by $(\mathbf{x}_i^\top, y_i)$, $i = 1, \dots, M$. To formalize an ICL problem, the input of a model is a *token matrix* given by [1, 38]

$$\mathbf{E} := \begin{pmatrix} \mathbf{X}^\top & \mathbf{x} \\ \mathbf{y}^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(M+1)}. \tag{3.1}$$

The output of a model corresponds to a prediction of y.

A Transformer block. Modern large language models are often made by stacking basic Transformer blocks (see, for example, [34]). A basic Transformer block consists of a *self-attention* layer and a *multi-layer perceptron* (MLP) layer [34], $\mathbf{E} \mapsto \mathsf{MLP}\left[\mathsf{ATTN}\left(\mathbf{E}\right)\right]$. Here, the MLP layer is defined as $\mathsf{MLP}\left(\mathbf{E}\right) := \mathbf{W}_2^\top \mathsf{ReLU}\left(\mathbf{W}_1 \mathbf{E}\right)$, where $\mathsf{ReLU}(\cdot)$ refers to the entrywise *rectified linear unit* (ReLU) activation function, and $\mathbf{W}_1,\ \mathbf{W}_2 \in \mathbb{R}^{d_f\times(d+1)}$ are two weight matrices. The self-attention layer (we focus on the single-head version in this paper) is defined as $\mathsf{ATTN}(\mathbf{E}) := \mathbf{E} + \mathbf{W}_P^\top \mathbf{W}_V \mathbf{E}\mathbf{M} \cdot \mathsf{sfmx}\big((\mathbf{W}_K\mathbf{E})^\top \mathbf{W}_Q \mathbf{E}\big)$, where $\mathsf{sfmx}(\cdot)$ refers to the row-wise softmax operator, $\mathbf{W}_K,\ \mathbf{W}_Q \in \mathbb{R}^{d_k\times(d+1)}$ are the key and query matrices, respectively, $\mathbf{W}_P,\ \mathbf{W}_V \in \mathbb{R}^{d_v\times(d+1)}$ are the projection and value matrices, respectively, and $\mathbf{M}$ is a fixed masking matrix given by

$$\mathbf{M} := \begin{pmatrix} \mathbf{I}_M & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{(M+1)\times(M+1)}. \tag{3.2}$$

This mask matrix is included to reflect the asymmetric structure of a prompt since the label of query input $\mathbf{x}$ is not included in the token matrix [1, 21]. In the above formulation, d_k, d_v, d_f are three hyperparameters controlling the key size, value size, and width of the MLP layer, respectively. In the single-head case, it is common to set $d_k = d_v = d + 1$ and $d_f = 4(d + 1)$, where $d + 1$ corresponds to the embedding size (see, for example, [34]). Our formulation of a basic transformer block ignores all bias parameters and some popular techniques (such as layer normalization and dropout) to focus solely on the benefits brought by the model structure.

A linear Transformer block. To facilitate theoretical analysis, we ignore the non-linearities in the Transformer block (specifically, sfmx and ReLU) and work with a *linear Transformer block* (LTB) defined as

$$f_{\mathsf{LTB}} : \mathbb{R}^{(d+1)\times(M+1)} \to \mathbb{R}, \quad \mathbf{E} \mapsto \left[\mathbf{W}_2^\top \mathbf{W}_1 \left(\mathbf{E} + \mathbf{W}_P^\top \mathbf{W}_V \mathbf{E}\mathbf{M}\frac{\mathbf{E}^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{E}}{M}\right)\right]_{-1,-1}, \tag{3.3}$$

where $\mathbf{E}$ is the token matrix given by (3.1), $[\ \cdot\ ]_{-1,-1}$ refers to the bottom right entry of a matrix, $\mathbf{M}$ is the fixed masking matrix given by (3.2). Here, the trainable parameters are $\mathbf{W}_P, \mathbf{W}_V \in \mathbb{R}^{d_v\times(d+1)}$, $\mathbf{W}_K, \mathbf{W}_Q \in \mathbb{R}^{d_k\times(d+1)}$, and $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_f\times(d+1)}$. We use the bottom right entry of the transformed token matrix as the model output to form a prediction of the label y. The $1/M$ factor is a normalization factor in linear attention and can be absorbed into trainable parameters. Finally, we denote the hypothesis class formed by LTB models as $\mathcal{F}_{\mathsf{LTB}} := \{f_{\mathsf{LTB}} : \ \mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V, \mathbf{W}_P, \mathbf{W}_1, \mathbf{W}_2, d_k \ge d,\ d_v \ge d+1,\ d_f \ge 1\}$, where f_{LTB} is defined in (3.3).

A linear self-attention. We will also consider a *linear self-attention* (LSA) defined as

$$f_{\mathsf{LSA}} : \mathbb{R}^{(d+1)\times(M+1)} \to \mathbb{R}, \quad \mathbf{E} \mapsto \left[\mathbf{E} + \mathbf{W}_P^\top \mathbf{W}_V \mathbf{E}\mathbf{M}\frac{\mathbf{E}^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{E}}{M}\right]_{-1,-1}, \tag{3.4}$$

where $\mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_P, \mathbf{W}_V$ are trainable parameters. An LSA model can be viewed as an LTB model without the MLP layer (setting $d_f = d+1$ and $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{I}$). We remark that, unlike LTB, the residual connection in LSA plays no role because the bottom right entry of the prompt $\mathbf{E}$ is zero (see (3.1)). A variant of the LSA model has been studied by [1, 38, 37], where $\mathbf{W}_K^\top \mathbf{W}_Q$ and $\mathbf{W}_P^\top \mathbf{W}_V$ are respectively merged into one matrix parameter. Similarly, we denote the hypothesis class formed by LSA models as $\mathcal{F}_{\mathsf{LSA}} := \{f_{\mathsf{LSA}} : \mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V, \mathbf{W}_P, d_k \ge d, d_v \ge 1\}$, where f_{LSA} is defined in (3.4).

Linear regression tasks with a shared signal. Assume that data and context examples are generated as follows.

Assumption 3.1 (Distributional conditions)**.** Assume that $(\mathbf{X}, \mathbf{y}, \mathbf{x}, y)$ are generated by:

- First, a task parameter is independently generated by $\widetilde{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}^*, \boldsymbol{\Psi})$.
- The feature vectors are independently generated by $\mathbf{x}, \mathbf{x}_1, \dots \mathbf{x}_M \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{H})$.
- Then, the labels are generated by $y = \langle\widetilde{\boldsymbol{\beta}}, \mathbf{x}\rangle + \varepsilon,\ \ y_i = \langle\widetilde{\boldsymbol{\beta}}, \mathbf{x}_i\rangle + \varepsilon_i,\ i = 1, \dots, M$, where ε and ε_i's are independently generated by $\varepsilon, \varepsilon_1, \dots, \varepsilon_M \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.

Here, $\sigma^2 \ge 0$, $\mathbf{H} \succeq \mathbf{0}$, $\boldsymbol{\Psi} \succeq \mathbf{0}$, and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ are fixed but unknown quantities that govern the data distribution. We denote $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_M)^\top$.

We emphasize the importance of the mean of the task parameter $\boldsymbol{\beta}^*$ in Assumption 3.1. A non-zero $\boldsymbol{\beta}^*$ represents a shared signal across tasks, which is arguably common in practice. This assumption is implicitly used in [12] where they assumed task parameters are close to a meta parameter. In comparison, the prior works for ICL of linear regression [1, 38, 37] only considered a special case where $\boldsymbol{\beta}^* = 0$. In this special case, they showed that a single LSA layer can achieve nearly optimal ICL by approximating one-step gradient descent from zero initialization (GD-0). In more general cases where $\boldsymbol{\beta}^*$ is non-zero, we will show in Section 4 that LSA is insufficient to learn the shared signal and must incur an irreducible approximation error compared to LTB models. This sets our results apart from the prior papers.

ICL risk. We measure the ICL risk of a model f by the mean squared error,

$$\mathcal{R}(f) := \mathbb{E}(f(\mathbf{E}) - y)^2, \tag{3.5}$$

where $\mathbf{E}$ is defined in (3.1) and the expectation is over $\mathbf{E}$ (equivalent to over $\mathbf{X}$, $\mathbf{y}$, and $\mathbf{x}$) and y.

4 Benefits of the MLP Component

Our first main result separates the approximation abilities of the LTB and LSA models for ICL. Recall that an LSA model can be viewed as a special case of the LTB model with $\mathbf{W}_2 = \mathbf{W}_1 = \mathbf{I}$. So the best ICL risk achieved by LTB models is no larger than the best ICL risk achieved by LSA models. However, our next theorem shows a strictly positive gap between the best ICL risks achieved by those two model classes. This result highlights the benefits of the MLP layer for reducing approximation error in Transformer.

Theorem 4.1 (Approximation gap)**.** *Consider the ICL risk defined by* (3.5) *and the two hypothesis classes $\mathcal{F}_{\mathsf{LTB}}$ and $\mathcal{F}_{\mathsf{LSA}}$. Suppose that Assumption 3.1 holds. Then we have*

- $\inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}\left(f\right)$ *is independent of $\boldsymbol{\beta}^*$.*
- $\inf_{f\in\mathcal{F}_{\mathsf{LSA}}}\mathcal{R}\left(f\right)$ *is a function of $\boldsymbol{\beta}^*$. Moreover,*

$$\inf_{f\in\mathcal{F}_{\mathsf{LSA}}}\mathcal{R}\left(f\right)-\inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}\left(f\right)\geq\max\left\{\frac{2}{3\left(M+1\right)},\frac{\left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)+\sigma^2\right)^2}{\left(M+1\right)^2\operatorname{tr}\left(\left(\mathbf{H}\boldsymbol{\Psi}\right)^2\right)}\right\}\|\boldsymbol{\beta}^*\|_{\mathbf{H}}^2.$$

The proof of Theorem 4.1 is deferred to Appendix B. Theorem 4.1 reveals a gap in terms of the approximation abilities between the hypothesis set of LTB models and that of LSA models. Specifically, the best ICL performance achieved by LTB models is independent of $\boldsymbol{\beta}^*$, while the best ICL performance achieved by LSA models is sensitive to the norm of $\boldsymbol{\beta}^*$. In particular, when $\|\boldsymbol{\beta}^*\|_{\mathbf{H}}^2=\Omega(M)$, the best ICL risk of the former is smaller than that of the latter by at least a *constant* additive term. So when $\boldsymbol{\beta}^*$ is large, the hypothesis class formed by LSA models is restricted in its ability to perform effective ICL. We will show in Section 5 that the hypothesis class formed by LTB models can achieve nearly optimal ICL in this case.

We also remark that the $\Theta(1/M)$ factor in the lower bound in Theorem 4.1 is not improvable. This is because LSA models can implement a one-step GD algorithm that is *consistent* for linear regression tasks (that is, the risk converges to the Bayes risk as the number of context examples goes to infinity), with an excess risk bound of $\Theta(1/M)$ [38, 37]. So the approximation error gap is at most $\Theta(1/M)$. Nonetheless, the $\Omega(\|\boldsymbol{\beta}^*\|_{\mathbf{H}}^2/M)$ approximation gap between LSA models and LTB models shown by Theorem 4.1 suggests that $\boldsymbol{\beta}^*$ is not learnable by LSA models during pre-training. In contrast, we will show in Section 6 that LTB models can learn $\boldsymbol{\beta}^*$ during pre-training.

We emphasize that the ability of LTB to learn non-zero mean is a joint effect of an MLP component and a skip connection. Note that LSA also has a skip connection, but its skip connection is inactive. In comparison, the MLP component in LTB activates the skip connection. Therefore, we attribute the ability to learn non-zero mean to the MLP component, which is the only difference between LTB and LSA. Nonetheless, one can attribute the ability to learn non-zero mean to the skip connection: without a skip connection, LTB reduces to LSA with a potential rank constraint on the parameter, which cannot learn non-zero mean as we have proved. The above two explanations take different perspectives to interpret the same phenomenon. Finally, we remark that Theorem 4.1 holds even when $\mathbf{H}$ and $\boldsymbol{\Psi}$ in Assumption 3.1 are not full rank.

Does scratchpad help? We have demonstrated that employing a single-layer LSA introduces an additional approximation error in the in-context learning problem for the linear regression task defined in Assumption 3.1. Nonetheless, are there alternative structures, apart from MLPs, that could potentially reduce this approximation error? One plausible strategy is to include a "scratchpad" in the input token. Specifically, we construct the token matrix as follows:

$$\mathbf{E}:=\begin{pmatrix}\mathbf{X}^\top & \mathbf{x}\\ \mathbf{1}_M^\top & 1\\ \mathbf{y}^\top & 0\end{pmatrix}\in\mathbb{R}^{(d+2)\times(M+1)}\tag{4.1}$$

which we then input into the LSA layer. We discuss the limitation of this scheme in Appendix J. Notably, this method does not successfully recover the GD-$\boldsymbol{\beta}$ estimator defined in Section 5. We leave it as future work to see whether the token matrix with scratchpad could implement other types of estimators that more effectively address the linear regression tasks defined in Assumption 3.1, as well as whether additional structures could help alleviate this approximation error.

Experiments on GPT2. Theorem 4.1 shows the importance of the MLP component in LSA models for reducing approximation error. We also empirically validate this result by training a more complex GPT2 model [14] for the ICL tasks specified by Assumption 3.1. In the experiments, we use a GPT2-small model (with or without the MLP component) with 6 layers and 4 heads in each layer. The experiments follow the setting in [14], except that we train the model using a token matrix defined

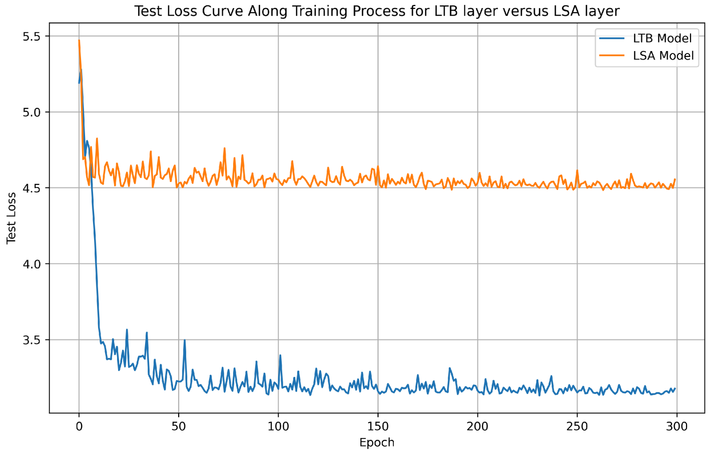

Figure 1: The test loss along the training process for LTB and LSA layer.

in (3.1). We considere two ICL settings, which instantiates Assumption 3.1 with $\boldsymbol{\beta}^* = (0, 0, ..., 0)^\top$ and $\boldsymbol{\beta}^* = (10, 10, ..., 10)^\top$, respectively. We set $d = 20, M = 40, \sigma = 0, \boldsymbol{\Psi} = \mathbf{H} = \mathbf{I}_d$ in both settings. More experimental details are in Appendix I. In each setting, we train and test the model using the same data distribution. The experimental results are presented in Table 1. From Table 1, we observe that both models (with or without the MLP component) achieve a nearly zero loss when the task mean is zero. However, when the task mean is set away from zero, the GPT2 model with MLP component still performs relatively well while the GPT2 model without MLP component incurs a significantly larger loss. These empirical observations are consistent with our Theorem 4.1, indicating the benefits of MLP layers in reducing approximation error for ICL of linear regression with a shared signal.

Experiments on LSA and LTB. We trained both the LSA and LTB layers for $\boldsymbol{\beta}^* = (1, 1, ..., 1)^\top$, and found that the trained LTB layer consistently achieved a significantly lower ICL risk than the LSA layer (see Figure 1). In these experiments, we adhered strictly to the previously defined LSA and LTB structures, setting the parameters as follows: $d = 5$, $M = 5$, $\sigma = 0$, and $\boldsymbol{\Psi} = \mathbf{H} = \mathbf{I}_d$. At each training step, we sampled $B = 128$ new linear regression tasks.

Table 1: Losses of GPT2 with or without MLP component for linear regression with a shared signal.

Model	GPT2	GPT2-noMLP
$\boldsymbol{\beta}^* = 0 \times \mathbf{1}$	0.003	0.024
$\boldsymbol{\beta}^* = 10 \times \mathbf{1}$	0.013	8.871

In this part, we have shown that LSA models, the primary focus in previous ICL theory literature (see, e.g., [1, 38, 37] and references therein), are not sufficiently expressive for ICL of linear regression with a shared signal. In what follows, we will show LTB models are sufficient for this ICL problem.

5 LTB Implements One-Step GD with Learnable Initialization

To understand the expressive power of the LTB models, we build a connection between $\mathcal{F}_{\mathsf{LTB}}$ and its subset of models which we call *one-step GD with learnable initialization* (GD-$\boldsymbol{\beta}$). We will first introduce GD-$\boldsymbol{\beta}$ models and then show that the best LTB models that minimize the ICL risk effectively belong to GD-$\boldsymbol{\beta}$ models.

The GD-$\boldsymbol{\beta}$ models. A GD-$\boldsymbol{\beta}$ model is defined as

$$f_{\mathsf{GD}\text{-}\boldsymbol{\beta}} : \mathbb{R}^{(d+1)\times(M+1)} \to \mathbb{R}, \quad \mathbf{E} \mapsto \left\langle \boldsymbol{\beta} - \boldsymbol{\Gamma} \cdot \frac{1}{M} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} - \mathbf{y}), \mathbf{x} \right\rangle, \tag{5.1}$$

where $\mathbf{E}$ is the token matrix given by (3.1), $\boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$ are trainable parameters. Similarly, we define the function class formed by GD-$\boldsymbol{\beta}$ models as $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}} := \{f_{\mathsf{GD}\text{-}\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathbb{R}^d, \boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}\}$.

A $\mathsf{GD}\text{-}\boldsymbol{\beta}$ model computes a parameter that fits the context examples $(\mathbf{X}, \mathbf{y})$ and uses that parameter to make a linear prediction of label y on feature $\mathbf{x}$. More specifically, the first step is by using one gradient descent step with a *matrix stepsize* $\boldsymbol{\Gamma}$ and an *initialization* $\boldsymbol{\beta}$ on the least square objective formed by context examples $(\mathbf{X}, \mathbf{y})$.

LTB implements $\mathsf{GD}\text{-}\boldsymbol{\beta}$**.** A $\mathsf{GD}\text{-}\boldsymbol{\beta}$ model (5.1) is a special case of an LTB model (3.3) by setting

$$\mathbf{W}_2^\top \mathbf{W}_1 = \begin{pmatrix} * & * \\ \boldsymbol{\beta}^\top & 1 \end{pmatrix}, \quad \mathbf{W}_P^\top \mathbf{W}_V = \begin{pmatrix} -\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d^\top & 1 \end{pmatrix}, \quad \mathbf{W}_K^\top \mathbf{W}_Q = \begin{pmatrix} \boldsymbol{\Gamma} & * \\ \mathbf{0}_d^\top & * \end{pmatrix},$$

where $*$ denotes the entries that do not affect the model output (hence can be set to anything). Note that $\boldsymbol{\Gamma}$ and $\boldsymbol{\beta}$ are *free* provided that $d_v \geq d+1$, $d_k \geq d$ and $d_f \geq 1$ (as required in $\mathcal{F}_{\mathsf{LTB}}$). In sum, we have proved the following lemma showing that the set of $\mathsf{GD}\text{-}\boldsymbol{\beta}$ models belongs to the set of LTB models.

Lemma 5.1. *We have $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}} \subseteq \mathcal{F}_{\mathsf{LTB}}$. Therefore, $\inf_{f \in \mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) \geq \inf_{f \in \mathcal{F}_{\mathsf{LTB}}} \mathcal{R}(f)$.*

Optimal $\mathsf{GD}\text{-}\boldsymbol{\beta}$ **models for ICL.** We now consider $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ and its optimal ICL risk. We have the following theorem that computes the globally minimal ICL risk over $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ and specifies the sufficient and necessary conditions for a global minimizer. The proof is deferred to Appendix C.

Theorem 5.2 (Optimal $\mathsf{GD}\text{-}\boldsymbol{\beta}$ models)**.** *Consider the ICL risk defined by* (3.5)*. Suppose that Assumption 3.1 holds. Then we have*

- *The minimal ICL risk of $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ is*

$$\inf_{f \in \mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) = \sigma^2 + \operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}} \boldsymbol{\Psi} \mathbf{H}^{\frac{1}{2}} \left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \boldsymbol{\Psi} \mathbf{H}^{\frac{1}{2}} \boldsymbol{\Omega}^{-1} \right) \right), \tag{5.2}$$

  where $\boldsymbol{\Omega} := [(M+1)\mathbf{H}^{\frac{1}{2}} \boldsymbol{\Psi} \mathbf{H}^{\frac{1}{2}} + (\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2)\mathbf{I}_d]/M$.

- *The global optimal parameters for $f_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ that attain the minimum* (5.2) *take the following form:*

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* + \mathsf{null}(\mathbf{H}), \quad \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^* + \mathsf{null}\left(\mathbf{H}^{\otimes 2}\right), \quad \textit{where} \quad \boldsymbol{\Gamma}^* := \boldsymbol{\Psi} \mathbf{H}^{\frac{1}{2}} \boldsymbol{\Omega}^{-1} \mathbf{H}^{-\frac{1}{2}} \tag{5.3}$$

  and $\boldsymbol{\beta}^*$ is the mean of the task parameter in Assumption 3.1 Here, $\mathsf{null}(\mathbf{H}) := \left\{ \boldsymbol{z} \in \mathbb{R}^d : \mathbf{H}\boldsymbol{z} = \mathbf{0} \right\}$ is the null space of $\mathbf{H}$, and $\mathsf{null}\left(\mathbf{H}^{\otimes 2}\right) = \left\{ \boldsymbol{Z} \in \mathbb{R}^{d \times d} : \mathbf{H}\boldsymbol{Z}\mathbf{H} = \mathbf{0} \right\}$ is the null space of $\mathbf{H}^{\otimes 2}$. In particular, when $\mathbf{H}$ is positive definite, the global optimal parameter is unique, $(\boldsymbol{\beta}, \boldsymbol{\Gamma}) = (\boldsymbol{\beta}^*, \boldsymbol{\Gamma}^*)$.

- *Under Assumption 3.1, the global optimal $f_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ that attains the minimum* (5.2) *is unique as a function of $\mathbf{E}$ (given by* (3.1)*) and takes the following form:*

$$f^*(\mathbf{E}) = \left\langle \boldsymbol{\beta}^* - \frac{1}{M} \boldsymbol{\Gamma}^* \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}), \mathbf{x} \right\rangle. \tag{5.4}$$

Theorem 5.2 characterizes the optimal $\mathsf{GD}\text{-}\boldsymbol{\beta}$ models for ICL. In the above theorem, $\boldsymbol{\Gamma}^* \to \mathbf{H}^{-1}$ as the context length M goes to infinity (assuming that $\mathbf{H}$ is positive definite). In this case, the optimal $\mathsf{GD}\text{-}\boldsymbol{\beta}$ function (5.2) implements one Newton step from initialization $\boldsymbol{\beta}^*$. With a finite context length M, (5.2) implements one regularized Newton step from initialization $\boldsymbol{\beta}^*$.

Optimal LTB models for ICL. Lemma 5.1 shows that the best ICL risk achieved by an $\mathsf{GD}\text{-}\boldsymbol{\beta}$ model is no smaller than the best ICL risk achieved by an LTB model. Surprisingly, our next theorem shows that the best ICL risk achieved by a $\mathsf{GD}\text{-}\boldsymbol{\beta}$ is *equal* to that achieved by an LTB model. Therefore, the hypothesis set $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ is diverse enough to match the approximation ability of the larger hypothesis set $\mathcal{F}_{\mathsf{LTB}}$ for ICL.

Theorem 5.3 (Optimal LTB models)**.** *Consider the ICL risk defined by* (3.5)*. Suppose that Assumption 3.1 holds and that $\mathsf{rank}\big(\mathbf{H}^{\frac{1}{2}} \boldsymbol{\Psi}^{\frac{1}{2}}\big) \geq 2$. Then we have*

- *The minimal ICL risk of $\mathcal{F}_{\mathsf{LTB}}$ and of $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ are equal,*

$$\inf_{f \in \mathcal{F}_{\mathsf{LTB}}} \mathcal{R}(f) = \inf_{f \in \mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) = \textit{ RHS of } (5.2)$$

- *Rewrite an LTB model as f_{LTB} in* (3.3) *with parameters ($*$ denotes parameters that do not affect the output)*
$$\mathbf{W}_2^\top \mathbf{W}_1 = \begin{pmatrix} * & * \\ \boldsymbol{\gamma}^\top & * \end{pmatrix}, \quad \mathbf{W}_K^\top \mathbf{W}_Q = \begin{pmatrix} \mathbf{V}_{11} & * \\ \mathbf{v}_{12}^\top & * \end{pmatrix}, \quad \mathbf{W}_2^\top \mathbf{W}_1 \mathbf{W}_P^\top \mathbf{W}_V = \begin{pmatrix} * & * \\ \mathbf{v}_{21}^\top & v_{-1} \end{pmatrix}.$$
Then the sufficient and necessary conditions for $f_{\mathsf{LTB}} \in \arg\min_{f \in \mathcal{F}_{\mathsf{LTB}}} \mathcal{R}(f)$ are
$$\begin{aligned} & v_{-1} \neq 0, \quad v_{-1}\mathbf{v}_{12} \in \mathsf{null}\left(\mathbf{H}\right), \quad \mathbf{v}_{21} \in -v_{-1}\boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}\right), \quad \boldsymbol{\gamma} \in \boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}\right), \\ & v_{-1}\mathbf{V}_{11} \in \boldsymbol{\Gamma}^{*\top} - v_{-1}\boldsymbol{\beta}^*\mathbf{v}_{12}^\top + \mathsf{null}\left(\mathbf{H}^{\otimes 2}\right), \end{aligned}$$
where $\boldsymbol{\beta}^*$ is defined in Assumption 3.1 and $\boldsymbol{\Gamma}^*$ is defined in (5.3)*. In particular, when $\mathbf{H}$ is positive definite, the globally optimal parameter represented by $(\mathbf{V}_{11}, \mathbf{v}_{12}, \mathbf{v}_{21}, v_{-1}, \boldsymbol{\gamma})$ is unique up to a rescaling of v_{-1}.*
- *Under Assumption 3.1, the globally optimal LTB model (that is, a function in $\arg\min_{f \in \mathcal{F}_{\mathsf{LTB}}}$) is unique as a function of $\mathbf{E}$ (given by* (3.1)*) and takes the form of* (5.4) *almost surely.*

Lemma 5.1 and Theorem 5.3 together show that $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ is a representative subset of $\mathcal{F}_{\mathsf{LTB}}$ that does not incur additional approximation error. In addition, every optimal LTB model is effectively an optimal GD-$\boldsymbol{\beta}$ model when restricted to all possible token matrices. Note that the optimal model parameters for LTB or GD-$\boldsymbol{\beta}$ are not unique because of redundant parameterization. But the optimal LTB and GD-$\boldsymbol{\beta}$ models are unique as a function of all possible token matrices. The above holds even when $\mathbf{H}$ and $\boldsymbol{\Psi}$ are potentially rank deficient.

Comparison with prior works. A line of papers considers LSA models for ICL under Assumption 3.1 with $\boldsymbol{\beta}^* = 0$ [35, 1, 38, 37]. They show that LSA models can (effectively) implement all possible GD-0 models that specialize GD-$\boldsymbol{\beta}$ models by fixing $\boldsymbol{\beta} = \mathbf{0}$. In addition, they show that every optimal LSA model is (effectively) a GD-0 model for ICL under Assumption 3.1 with $\boldsymbol{\beta}^* = \mathbf{0}$ (see, for example, Theorem 1 in [1]). In comparison, we consider a harder ICL problem that allows a large shared signal in tasks (that is, a large $\boldsymbol{\beta}^*$ in Assumption 3.1). In this setting, our Theorem 4.1 shows that $\mathcal{F}_{\mathsf{LSA}}$ (hence its subset formed by GD-0 models), as a subet of $\mathcal{F}_{\mathsf{LTB}}$, incurs an additional approximation error propotional to $\|\boldsymbol{\beta}^*\|_{\mathbf{H}}^2$ compared with $\mathcal{F}_{\mathsf{LTB}}$. In contrast, $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$, as a subset of $\mathcal{F}_{\mathsf{LTB}}$, does not incur additional approximation error according to our Theorem 5.3. Thus the LSA and GD-0 models considered by [35, 1, 38, 37] are not capable of learning the shared signal $\boldsymbol{\beta}^*$, while an LTB model can learn $\boldsymbol{\beta}^*$ through implementing GD-$\boldsymbol{\beta}$ and encoding $\boldsymbol{\beta}^*$ in the initialization parameter.

6 Training and In-Context Learning of GD-beta

We have shown that $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ is a representative subset of $\mathcal{F}_{\mathsf{LTB}}$ that effectively contains every optimal LTB model. We now examine the ICL and training of GD-$\boldsymbol{\beta}$ models.

Nearly optimal ICL with GD-$\boldsymbol{\beta}$. We will compare the best ICL risk achieved by GD-$\boldsymbol{\beta}$ with the best ICL risk achieved by any estimator. The following lemma is an extension of Proposition 5.1 and Corollary 5.2 in [37] (which is based on [33]) that characterizes the Bayes optimal ICL risk among all estimators.

Lemma 6.1 (Bayes optimal ICL)**.** *Given a task-specific dataset $(\mathbf{X}, \mathbf{y}, \mathbf{x}, y)$ sampled according to Assumption 3.1, let $g\left(\mathbf{X}, \mathbf{y}, \mathbf{x}\right)$ be an arbitrary estimator for y and measure the average linear regression risk by $\mathcal{L}\left(g; \mathbf{X}\right) := \mathbb{E}[(g(\mathbf{X}, \mathbf{y}, \mathbf{x}) - y)^2 \mid \mathbf{X}]$. It is clear that $\mathbb{E}\mathcal{L}\left(g; \mathbf{X}\right) = \mathcal{R}(g)$. Then,*

- *The optimal estimator that minimizes the average linear regression risk $\mathcal{L}(\cdot; \mathbf{X})$ is $g^*(\mathbf{X}, \mathbf{y}, \mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}^* + \mathbf{x}^\top \boldsymbol{\Psi}^{\frac{1}{2}} \left(\boldsymbol{\Psi}^{\frac{1}{2}} \mathbf{X}^\top \mathbf{X} \boldsymbol{\Psi}^{\frac{1}{2}} + \sigma^2 \mathbf{I}_d\right)^{-1} \boldsymbol{\Psi}^{\frac{1}{2}} \mathbf{X}^\top \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\right).$*
- *Assume the signal-to-noise ratio is upper bounded, that is, $\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) \lesssim \sigma^2$, then with probability at least $1 - \exp\left(-\Omega\left(M\right)\right)$ over the randomness of $\mathbf{X}$, it holds that $\mathcal{L}\left(g^*; \mathbf{X}\right) - \sigma^2 \simeq \sum_{i=1}^d \min\{\bar{\phi}, \phi_i\}$, where $\bar{\phi} \eqsim \frac{\sigma^2}{M}$, and $(\phi_i)_{i \geq 1}$ are the eigenvalues of $\boldsymbol{\Psi}^{\frac{1}{2}} \mathbf{H} \boldsymbol{\Psi}^{\frac{1}{2}}$.*

The proof is deferred to Appendix F. Lemma 6.1 shows that the Bayes optimal estimator is a ridge regression estimator centered at $\boldsymbol{\beta}^*$. This is consistent with [37] where the Bayes optimal estimator is a ridge regression estimator as they assumed $\boldsymbol{\beta}^* = 0$. The following corollary of Theorem 5.2 computes the rate of the ICL risk achieved by the optimal $\mathsf{GD}\text{-}\boldsymbol{\beta}$ model.

Corollary 6.2. *Under the setup of Theorem 5.2, additionaly assume the signal-to-noise ratio is upper bounded, that is,* $\operatorname{tr}(\boldsymbol{\Psi}\mathbf{H}) \lesssim \sigma^2$*. Then we have* $\inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) - \sigma^2 \eqsim \sum_{i=1}^{d} \min\left\{\bar{\phi}, \phi_i\right\}$ *, where* $(\phi_i)_{i\geq 0}$ *are the eigenvalues of* $\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}$ *and* $\bar{\phi} := [\operatorname{tr}(\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}) + \sigma^2]/M \eqsim \sigma^2/M$*.*

The optimal (expected) ICL risk achieved by $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ in Corollary 6.2 matches the (high probability) Bayes optimal ICL risk in Lemma 6.1 ignoring constant factors, provided that the signal-to-noise ratio is upper bounded. Therefore $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ achieves nearly Bayes optimal ICL risk. As a consequence, the larger hypothesis set $\mathcal{F}_{\mathsf{LTB}}$ also achieves nearly Bayes optimal ICL of linear regression under Assumption 3.1.

For the simplicity of discussion, we assume a fixed context length during pretraining and inference. Our discussions can be extended to allow a different context length during pretraining and inference using techniques in [37]. However, this is not the main focus of this work.

Optimization of $\mathsf{GD}\text{-}\boldsymbol{\beta}$ with infinite tasks. We have shown that $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ is a representative subset of $\mathcal{F}_{\mathsf{LTB}}$ that covers the optimal LTB models and achieves nearly optimal ICL risk. We now consider the optimization in the parameter space specified by $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$. For simplicity, we follow [38] and consider gradient descent with an infinitesimal stepsize on the ICL objective with an infinite number of tasks. That is, we consider the optimization of gradient flow on the population ICL risk under the parameterization of $\mathsf{GD}\text{-}\boldsymbol{\beta}$,

$$\frac{\mathrm{d}\boldsymbol{\beta}(t)}{\mathrm{d}t} = -\frac{\partial}{\partial\boldsymbol{\beta}}\mathcal{R}(f_{\mathsf{GD}\text{-}\boldsymbol{\beta}}), \quad \frac{\mathrm{d}\boldsymbol{\Gamma}(t)}{\mathrm{d}t} = -\frac{\partial}{\partial\boldsymbol{\Gamma}}\mathcal{R}(f_{\mathsf{GD}\text{-}\boldsymbol{\beta}}), \tag{6.1}$$

where $\mathcal{R}$ is defined by (3.5) and $f_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ is defined by (5.1).

The following theorem guarantees the global convergence of gradient flow. We introduce some notation to accommodate cases when $\mathbf{H}$ is rank deficient. Let $\mathcal{P}_{\mathcal{S}}$ be the orthogonal projection operator onto a subspace $\mathcal{S}$. Let $\mathcal{H} = \mathsf{Im}(\mathbf{H})$ be the image space of matrix $\mathbf{H}$ (viewing $\mathbf{H}$ as a linear map) and $\mathcal{H}^{\perp} := \mathsf{null}(\mathbf{H})$ be its orthogonal complement. Similarly, let $\mathcal{Z} := \mathsf{Im}\left(\mathbf{H}^{\otimes 2}\right) = \{\mathbf{H}\boldsymbol{Z}\mathbf{H}, \boldsymbol{Z} \in \mathbb{R}^{d\times d}\}$ be the image space of the operator $\mathbf{H}^{\otimes 2}$ and $\mathcal{Z}^{\perp}$ be its orthogonal complement. Then we have the following theorem.

Theorem 6.3. *Consider the gradient flow defined by* (6.1) *with initialization* $\boldsymbol{\beta}(0), \boldsymbol{\Gamma}(0)$*. We have,*

$$\mathcal{P}_{\mathcal{H}}(\boldsymbol{\beta}(t)) \to \mathcal{P}_{\mathcal{H}}(\boldsymbol{\beta}^*), \quad \mathcal{P}_{\mathcal{H}^{\perp}}(\boldsymbol{\beta}(t)) = \mathcal{P}_{\mathcal{H}^{\perp}}(\boldsymbol{\beta}(0)),$$
$$\mathcal{P}_{\mathcal{Z}}(\boldsymbol{\Gamma}(t)) \to \mathcal{P}_{\mathcal{Z}}(\boldsymbol{\Gamma}^*), \quad \mathcal{P}_{\mathcal{Z}^{\perp}}(\boldsymbol{\Gamma}(t)) = \mathcal{P}_{\mathcal{Z}^{\perp}}(\boldsymbol{\Gamma}(0))$$

as $t \to \infty$*. In particular, if* $\mathbf{H}$ *is positive definite, the gradient flow converges to the unique global minimizer of ICL risk over* $\mathsf{GD}\text{-}\boldsymbol{\beta}$ *class, that is,* $\boldsymbol{\beta}(t) \to \boldsymbol{\beta}^*$ *and* $\boldsymbol{\Gamma}(t) \to \boldsymbol{\Gamma}^*$ *as* $t \to \infty$*.*

The proof, as well as the convergence rate, is deferred to Appendix G. We remark that (6.1) is a complex dynamical system on a non-convex potential function of $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$. We briefly discuss our proof techniques assuming that $\mathbf{H}$ is full rank. The rank-deficient cases can be handled in the same way by applying appropriate project operators. To conquer the non-convex optimization issue, we observe that for every fixed $\boldsymbol{\Gamma}$, the potential as a function of $\boldsymbol{\beta}$ is smooth and strongly convex with a uniformly bounded condition number. This observation allows us to establish a uniform convergence for $\boldsymbol{\beta}$. When $\boldsymbol{\beta}$ is sufficiently close to $\boldsymbol{\beta}^*$, the potential as a function of $\boldsymbol{\Gamma}$ is approximately convex, allowing us to track the convergence of $\boldsymbol{\Gamma}$.

Theorem 6.3 shows that optimization of $\mathsf{GD}\text{-}\boldsymbol{\beta}$ can be done efficiently by gradient flow without suffering from non-convexity. However, as we have shown in previous sections, LTB utilizes a more complex parameterization than $\mathsf{GD}\text{-}\boldsymbol{\beta}$. So Theorem 6.3 does not imply optimization of LTB is easy. We leave it as future work to study the optimization and statistical complexity for directly learning LTB models.

7 Concluding Remarks

In this paper, we study the in-context learning of linear regression with a shared signal represented by a Gaussian prior with a non-zero mean. We show that although the linear self-attention layer discussed in prior works is consistent for this more complex task, its risk has an inevitable gap compared to that of the linear Transformer block (LTB), which is a linear self-attention layer followed by a linear multi-layer perception (MLP) layer. Next, we show that the effectiveness of the LTB arises because it can implement the one-step gradient descent estimator with learnable initialization (GD-$\boldsymbol{\beta}$). Moreover, all global minimizers in the LTB class are equivalent to the unique global minimizer in the GD-$\boldsymbol{\beta}$ class, which can achieve nearly Bayes optimal in-context learning risk. Finally, we consider training on in-context examples and prove global convergence over the GD-$\boldsymbol{\beta}$ class of gradient flow on the population loss. Several future directions are worth discussing.

Optimization and statistical complexity. This paper provides an approximation theory of LTB and an optimization theory of GD-$\boldsymbol{\beta}$. However, the statistical complexity of learning LTB or GD-$\boldsymbol{\beta}$ is not considered. The work by [37] provided techniques for analyzing the statistical task complexity for pre-training GD-0. An interesting direction is to extend their method to study the statistical complexity of learning GD-$\boldsymbol{\beta}$. However, their method crucially relies on the convexity of the risk induced GD-0, while we have shown that the risk induced by GD-$\boldsymbol{\beta}$ is non-convex. New ideas for dealing with non-convexity are needed here.

From LTB to Transformer block. We focus on LTB in this work, which simplifies a vanilla Transformer block by removing the non-linearities from the softmax self-attention and the ReLU activation in the MLP layers. This simplification allows us to obtain precise theoretical results for LTB (such as its connection to GD-$\boldsymbol{\beta}$). On the other hand, non-linearities are arguably necessary for Transformers to work well in practice. An important next step is to further consider the theoretical benefits of non-linearities based on our current results.

Roles of MLP layers. The work by [15] (and references thereafter) empirically found that MLP layers operate as key-value memories that store human-interpretable patterns in some pre-trained Transformers. Their work motivated a method for locating and editing information stored in language models by modifying their MLP layers (see, e.g., [9, 23]). Our work proves that the MLP component enables LTB to learn the shared signal in linear regression tasks, which cannot be done by a single LSA component. We leave it as future work to theoretically clarify the information stored in the MLP component.

Acknowledgments

We gratefully acknowledge the support of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639, and of the NSF through grant DMS-2023505.

References

[1] 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.

[2] Kabir Ahuja, Madhur Panwar, and Navin Goyal. In-context learning through the bayesian prism. In *The Twelfth International Conference on Learning Representations*, 2024.

[3] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2022.

[4] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[5] Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. Understanding in-context learning in transformers and LLMs by learning to learn discrete functions. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*, 2021.

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[9] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022.

[10] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[13] Jingwen Fu, Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. How does representation impact in-context learning: A exploration on a synthetic task. *arXiv preprint arXiv:2309.06054*, 2023.

[14] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

[15] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[16] Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do transformers learn in-context beyond simple functions? a case study on learning with representations. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Jonathan N Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *arXiv preprint arXiv:2306.14892*, 2023.

[18] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pages 19565–19594. PMLR, 2023.

[19] Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *The Twelfth International Conference on Learning Representations*, 2024.

[20] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[21] Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2024.

[22] AR Meenakshi and C Rajian. On a product of positive semidefinite matrices. *Linear algebra and its applications*, 295(1-3):3–6, 1999.

[23] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[24] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

[25] OpenAI. Gpt-4 technical report, 2023.

[26] Reese Pathak, Rajat Sen, Weihao Kong, and Abhimanyu Das. Transformers can optimally learn regression mixture models. *arXiv preprint arXiv:2311.08362*, 2023.

[27] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[30] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *arXiv preprint arXiv:2306.15063*, 2023.

[31] George AF Seber. *A matrix handbook for statisticians*. John Wiley & Sons, 2008.

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[33] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1, 2023.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[37] 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? In *The Twelfth International Conference on Learning Representations*, 2024.

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

A Notation and Variable Transformation

In order to simplify the proof, let's first describe a variable transformation scheme for our model and data coming from Assumption 3.1.

Definition A.1 (Variable Transformation)**.** We fix M as the length of the contexts. Recall that $\mathbf{X}, \mathbf{x}$ are, respectively, the features in the context and the query input, while $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ are the true task parameter for the inference prompt and its expectation, respectively. Following the Assumption 3.1, we have $\widetilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}^*, \boldsymbol{\Psi}\right)$. From the definition for multivariate Gaussian distribution, we know there exists a random vector $\widetilde{\boldsymbol{\theta}}$ such that

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}, \tag{A.1}$$

where

$$\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}\left(0, \mathbf{I}_d\right). \tag{A.2}$$

Recall the noise vector is defined as $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and is generated from $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_M\right)$.

Rank deficient case Note that, even when $\boldsymbol{\Psi}$ is rank-deficient, the variable transformation in (A.1) and (A.2) still hold. This can be seen from the definition of the multivariate Gaussian distribution (Def 20.11 and the discussion below in [31]): A vector $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ with mean $\boldsymbol{\beta}^*$ and covariance matrix $\boldsymbol{\Psi}$ has a multivariate normal distribution, if it has the same distribution as $\boldsymbol{A}\widetilde{\boldsymbol{\theta}} + \boldsymbol{\beta}^*$ where $\boldsymbol{A}$ is any $d \times m$ matrix satisfying $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Psi}$ and $\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}\left(\mathbf{0}_m, \mathbf{I}_m\right)$. Here, we can recover the variable transformation above if we take $m = d$ and $\boldsymbol{A} = \boldsymbol{\Psi}^{\frac{1}{2}}$.

Notation Before we delve into the detailed proof, let's repeat the notation part with some additional notations. We use lowercase bold letters to note vectors and uppercase bold letters to denote matrices and tensors. For a vector $\mathbf{x}$ and a positive semi-definite (PSD) matrix $\boldsymbol{A}$, we denote $\|x\|_{\boldsymbol{A}} := \sqrt{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}$. We denote $\langle \cdot, \cdot \rangle$ as the inner product, which is defined as $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^\top \mathbf{y}$ for vectors and $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \operatorname{tr}\left(\boldsymbol{A}\boldsymbol{B}^\top\right)$ for matrices. For a matrix $\boldsymbol{A}$, we denote $\|\boldsymbol{A}\|_{op}, \|\boldsymbol{A}\|_F$ as the operator norm and the Frobenius norm, respectively. We denote $\boldsymbol{A}[i]$ as the i-th row of the matrix $\boldsymbol{A}$, $\boldsymbol{A}_{m,n}$ as the (m, n)-th entry, and $\boldsymbol{A}_{-1,-1}$ as the right-botom entry. We denote $\mathbf{0}_n, \mathbf{0}_{m\times n}, \mathbf{I}_n$ as the zero vector, zero matrix and identity matrix, respectively.

For a positive semi-definite matrix $\boldsymbol{A}$, we denote $\boldsymbol{A}^{\frac{1}{2}}$ for the principle square root of $\boldsymbol{A}$, which is defined as the unique real matrix $\boldsymbol{B}$ such that $\boldsymbol{B}$ is positive semi-definite and $\boldsymbol{B}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{A}$. For positive definite $\boldsymbol{A}$, its principle square root is also positive definite. We denote $\boldsymbol{A}^+$ as the Moore-Penrose pseudo-inverse for any matrix $\boldsymbol{A}$. We also denote $\boldsymbol{A}^{-\frac{1}{2}} = \left(\boldsymbol{A}^{\frac{1}{2}}\right)^+$. We denote $\otimes$ as the Kronecker product. For compatible matrices of proper size $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, $\boldsymbol{B}^\top \otimes \boldsymbol{A}$ is a linear mapping which is defined by $\left(\boldsymbol{B}^\top \otimes \boldsymbol{A}\right) \circ \boldsymbol{C} = \boldsymbol{A}\boldsymbol{C}\boldsymbol{B}$. We denote $\boldsymbol{\Lambda} := \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}$ and $\phi_1 \geq \phi_2 \geq ... \geq \phi_d \geq 0$ are its ordered eigenvalues. We also denote $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d \geq 0$ as the ordered eigenvalues of $\mathbf{H}$, and $\lambda_{-1} > 0$ as its minimal positive eigenvalue. Finally, we denote another important matrix

$$\boldsymbol{\Omega} := \frac{M+1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} + \frac{\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2}{M} \cdot \mathbf{I}_d. \tag{A.3}$$

Note that, under this definition and our assumption on $\mathbf{H}$ and $\boldsymbol{\Psi}$, we have $\boldsymbol{\Omega}$ is invertible.

B Proof of Theorem 4.1

Proof. The fact that $\inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}(f)$ does not depend on the vector $\boldsymbol{\beta}^*$ is subsumed in the Theorem 5.3, so we do not prove it here. We are going to prove the inequality in the theorem. First, from Theorem 5.3 we know that,

$$\inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}(f)=\sigma^2+\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi})-\operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\boldsymbol{\Omega}^{-1}\right), \tag{B.1}$$

where $\boldsymbol{\Omega}$ is defined in (A.3). Then, it suffices to lower bound $\inf_{f\in\mathcal{F}_{\mathsf{LSA}}}\mathcal{R}(f)$. So let's take an arbitrary $f\in\mathcal{F}_{\mathsf{LSA}}$, which is denoted as

$$f(\mathbf{E})=\left[\mathbf{E}+\mathbf{W}_P^\top\mathbf{W}_V\mathbf{E}\mathbf{M}\frac{\mathbf{E}^\top\mathbf{W}_K^\top\mathbf{W}_Q\mathbf{E}}{M}\right]_{-1,-1},$$

where $\mathbf{E}$ is the token matrix defined in (3.1). Since the prediction is the right-bottom entry of the output matrix, we know only the last row of the product $\mathbf{W}_P^\top\mathbf{W}_V$ attends the prediction. Similarly, only the last column of the $\mathbf{E}$ on the far right in the above equation atttends the prediction. Since the last column of $\mathbf{E}$ is $\begin{pmatrix}\mathbf{x}^\top & 0\end{pmatrix}^\top$, we know that only the first d rows of the product $\mathbf{W}_K^\top\mathbf{W}_Q$ enter the calculation (since other parts are multipled by zero). Therefore, we denote

$$\mathbf{W}_P^\top\mathbf{W}_V=\begin{pmatrix}* & *\\ \mathbf{u}_{21}^\top & u_{-1}\end{pmatrix},\quad \mathbf{W}_K^\top\mathbf{W}_Q=\begin{pmatrix}\mathbf{U}_{11} & *\\ \mathbf{u}_{12}^\top & *\end{pmatrix},$$

where $\mathbf{U}_{11}\in\mathbb{R}^{d\times d},\mathbf{u}_{12},\mathbf{u}_{21}\in\mathbb{R}^{d\times 1},u_{-1}\in\mathbb{R}$, and $*$ denotes entries that do not enter the final prediction. The model prediction can be written as

$$\begin{aligned}
&f(\mathbf{E})=\begin{pmatrix}\mathbf{u}_{21}^\top & u_{-1}\end{pmatrix}\cdot\frac{\mathbf{E}\mathbf{M}_M\mathbf{E}^\top}{M}\cdot\begin{pmatrix}\mathbf{U}_{11}\\ \mathbf{u}_{12}^\top\end{pmatrix}\cdot\mathbf{x}\\
&=\left[\mathbf{u}_{21}^\top\cdot\frac{1}{M}\mathbf{X}^\top\mathbf{X}\cdot\mathbf{U}_{11}+\mathbf{u}_{21}^\top\cdot\frac{1}{M}\mathbf{X}^\top\mathbf{y}\cdot\mathbf{u}_{12}^\top+u_{-1}\cdot\frac{1}{M}\mathbf{y}^\top\mathbf{X}\cdot\mathbf{U}_{11}+u_{-1}\cdot\frac{1}{M}\mathbf{y}^\top\mathbf{y}\cdot\mathbf{u}_{12}^\top\right]\cdot\mathbf{x}.
\end{aligned}$$

Step 1: simplify the risk function. We use $\widetilde{\boldsymbol{\beta}}$ to denote the task parameter. From the Assumption 3.1 and Definition A.1, we have

$$\mathbf{y}=\mathbf{X}\widetilde{\boldsymbol{\beta}}+\boldsymbol{\varepsilon},\quad y=\left\langle\widetilde{\boldsymbol{\beta}},\mathbf{x}\right\rangle+\varepsilon,\quad \widetilde{\boldsymbol{\beta}}\sim\mathcal{N}(\boldsymbol{\beta}^*,\boldsymbol{\Psi}),\quad \widetilde{\boldsymbol{\beta}}=\boldsymbol{\beta}^*+\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}};$$

and

$$\mathbf{X}[i],\mathbf{x}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\mathbf{0},\mathbf{H}),\quad \boldsymbol{\varepsilon}[i],\varepsilon\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(0,\sigma^2\right),\quad \widetilde{\boldsymbol{\theta}}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d).$$

Then the model output can be written as

$$\begin{aligned}
&f(\mathbf{E})=\left[\mathbf{u}_{21}^\top\cdot\frac{1}{M}\mathbf{X}^\top\mathbf{X}\cdot\mathbf{U}_{11}+\mathbf{u}_{21}^\top\cdot\frac{1}{M}\mathbf{X}^\top\mathbf{y}\cdot\mathbf{u}_{12}^\top+u_{-1}\cdot\frac{1}{M}\mathbf{y}^\top\mathbf{X}\cdot\mathbf{U}_{11}+u_{-1}\cdot\frac{1}{M}\mathbf{y}^\top\mathbf{y}\cdot\mathbf{u}_{12}^\top\right]\cdot\mathbf{x}\\
&=\left[\left(\mathbf{u}_{21}+u_{-1}\widetilde{\boldsymbol{\beta}}\right)^\top\cdot\frac{1}{M}\mathbf{X}^\top\mathbf{X}\cdot\left(\mathbf{U}_{11}+\widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right)\right]\cdot\mathbf{x}\\
&\quad+\left[\mathbf{u}_{21}^\top\cdot\frac{1}{M}\mathbf{X}^\top\boldsymbol{\varepsilon}\cdot\mathbf{u}_{12}^\top+u_{-1}\cdot\frac{1}{M}\boldsymbol{\varepsilon}^\top\mathbf{X}\cdot\mathbf{U}_{11}+u_{-1}\cdot\frac{2}{M}\boldsymbol{\varepsilon}^\top\mathbf{X}\widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top+\frac{1}{M}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\cdot u_{-1}\mathbf{u}_{12}^\top\right]\cdot\mathbf{x}.
\end{aligned}$$

To simplify the presentation, we denote

$$\begin{aligned}
\boldsymbol{z}_1^\top&=\left(\mathbf{u}_{21}+u_{-1}\widetilde{\boldsymbol{\beta}}\right)^\top\cdot\frac{1}{M}\mathbf{X}^\top\mathbf{X}\cdot\left(\mathbf{U}_{11}+\widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right),\\
\boldsymbol{z}_2^\top&=\mathbf{u}_{21}^\top\cdot\frac{1}{M}\mathbf{X}^\top\boldsymbol{\varepsilon}\cdot\mathbf{u}_{12}^\top+u_{-1}\cdot\frac{1}{M}\boldsymbol{\varepsilon}^\top\mathbf{X}\cdot\mathbf{U}_{11}+u_{-1}\cdot\frac{2}{M}\boldsymbol{\varepsilon}^\top\mathbf{X}\widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\\
\boldsymbol{z}_3^\top&=\frac{1}{M}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\cdot u_{-1}\mathbf{u}_{12}^\top.
\end{aligned}$$

Since $\mathbf{x}, \mathbf{X}, \boldsymbol{\varepsilon}, \widetilde{\boldsymbol{\beta}}$ are independent, we have

$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}\left(f(\mathbf{E}) - \left\langle \widetilde{\boldsymbol{\beta}}, \mathbf{x}\right\rangle - \varepsilon\right)^2 \\ &= \sigma^2 + \mathbb{E}\left(f(\mathbf{E}) - \left\langle \widetilde{\boldsymbol{\beta}}, \mathbf{x}\right\rangle\right)^2 \qquad (\varepsilon \text{ is independent from other variables and zero-mean}) \\ &= \mathbb{E}\left[\left\langle \boldsymbol{z}_1 + \boldsymbol{z}_2 + \boldsymbol{z}_3 - \widetilde{\boldsymbol{\beta}}, \mathbf{x}\right\rangle^2\right] + \sigma^2 \\ &= \left\langle \mathbf{H}, \mathbb{E}\left(\boldsymbol{z}_1 + \boldsymbol{z}_2 + \boldsymbol{z}_3 - \widetilde{\boldsymbol{\beta}}\right)\left(\boldsymbol{z}_1 + \boldsymbol{z}_2 + \boldsymbol{z}_3 - \widetilde{\boldsymbol{\beta}}\right)^\top\right\rangle + \sigma^2.\end{aligned}$$

Note that $\boldsymbol{z}_1$ does not contain $\boldsymbol{\varepsilon}$, $\boldsymbol{z}_2$ is a linear form of $\boldsymbol{\varepsilon}$, and $\boldsymbol{z}_3$ is a quadratic form of $\boldsymbol{\varepsilon}$. Using $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$, we have $\mathbb{E}[(\boldsymbol{z}_1 - \widetilde{\boldsymbol{\beta}}) \cdot \boldsymbol{z}_2^\top] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{z}_2\boldsymbol{z}_3^\top] = \mathbf{0}$. Therefore, we have

$$\begin{aligned}\mathcal{R}(f) - \sigma^2 = &\underbrace{\left\langle \mathbf{H}, \mathbb{E}\left(\boldsymbol{z}_1 - \widetilde{\boldsymbol{\beta}}\right)\left(\boldsymbol{z}_1 - \widetilde{\boldsymbol{\beta}}\right)^\top\right\rangle}_{S_1} + \underbrace{\left\langle \mathbf{H}, \mathbb{E}\boldsymbol{z}_2\boldsymbol{z}_2^\top\right\rangle}_{S_2} + \underbrace{\left\langle \mathbf{H}, \mathbb{E}\boldsymbol{z}_3\boldsymbol{z}_3^\top\right\rangle}_{S_3} \\ &+ \underbrace{2\left\langle \mathbf{H}, \mathbb{E}\left(\boldsymbol{z}_1 - \widetilde{\boldsymbol{\beta}}\right)\boldsymbol{z}_3^\top\right\rangle}_{S_4}. \qquad \text{(B.2)}\end{aligned}$$

Step 2: compute S_1. By Lemma H.4 and that $\mathbf{X}$ is independent of all other random variables, we have

$$\begin{aligned}&\left\langle \mathbf{H}, \mathbb{E}\boldsymbol{z}_1\boldsymbol{z}_1^\top\right\rangle \\ =&\mathbb{E}\mathrm{tr}\left[\left(\mathbf{U}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right)^\top \cdot \frac{1}{M}\mathbf{X}^\top\mathbf{X} \cdot \left(\mathbf{u}_{21} + u_{-1}\widetilde{\boldsymbol{\beta}}\right)\left(\mathbf{u}_{21} + u_{-1}\widetilde{\boldsymbol{\beta}}\right)^\top \cdot \frac{1}{M}\mathbf{X}^\top\mathbf{X} \cdot \left(\mathbf{U}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right)\mathbf{H}\right] \\ =&\frac{M+1}{M}\mathbb{E}_{\widetilde{\boldsymbol{\beta}}}\mathrm{tr}\left[\left(\mathbf{U}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right)^\top \cdot \mathbf{H} \cdot \left(\mathbf{u}_{21} + u_{-1}\widetilde{\boldsymbol{\beta}}\right)\left(\mathbf{u}_{21} + u_{-1}\widetilde{\boldsymbol{\beta}}\right)^\top \cdot \mathbf{H} \cdot \left(\mathbf{U}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right)\mathbf{H}\right] \\ +&\frac{1}{M}\mathbb{E}_{\widetilde{\boldsymbol{\beta}}}\mathrm{tr}\left[\mathrm{tr}\left(\left(\mathbf{u}_{21} + u_{-1}\widetilde{\boldsymbol{\beta}}\right)\left(\mathbf{u}_{21} + u_{-1}\widetilde{\boldsymbol{\beta}}\right)^\top\mathbf{H}\right)\left(\mathbf{U}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right)^\top\mathbf{H}\left(\mathbf{U}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right)\mathbf{H}\right].\end{aligned}$$

We denote

$$\boldsymbol{b} := \mathbf{u}_{21} + u_{-1}\boldsymbol{\beta}^* \in \mathbb{R}^d, \quad \boldsymbol{A} := \mathbf{U}_{11} + \boldsymbol{\beta}^*\mathbf{u}_{12}^\top \in \mathbb{R}^{d\times d}, \qquad \text{(B.3)}$$

then applying $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}$, we get

$$\begin{aligned}&\left\langle \mathbf{H}, \mathbb{E}\boldsymbol{z}_1\boldsymbol{z}_1^\top\right\rangle \\ =&\frac{M+1}{M}\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\mathrm{tr}\left[\left(\boldsymbol{A} + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^\top\right)^\top \cdot \mathbf{H} \cdot \left(\boldsymbol{b} + u_{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)\left(\boldsymbol{b} + u_{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)^\top \cdot \mathbf{H} \cdot \left(\boldsymbol{A} + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^\top\right)\mathbf{H}\right] \\ +&\frac{1}{M}\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\mathrm{tr}\left(\left(\boldsymbol{b} + u_{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)\left(\boldsymbol{b} + u_{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)^\top\mathbf{H}\right)\mathrm{tr}\left[\left(\boldsymbol{A} + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^\top\right)^\top\mathbf{H}\left(\boldsymbol{A} + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^\top\right)\mathbf{H}\right].\end{aligned}$$

Note that the first and third moments of $\widetilde{\boldsymbol{\theta}}$ are zero. So the above equation only involves the zeroth, second, and fourth moments of $\widetilde{\boldsymbol{\theta}}$. Therefore we have

$$\left\langle \mathbf{H}, \mathbb{E}\boldsymbol{z}_1\boldsymbol{z}_1^\top\right\rangle = \underbrace{\frac{M+1}{M}\mathrm{tr}\left(\boldsymbol{A}^\top\mathbf{H}\boldsymbol{b}\boldsymbol{b}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\right) + \frac{1}{M}\mathrm{tr}\left(\boldsymbol{b}\boldsymbol{b}^\top\mathbf{H}\right)\mathrm{tr}\left(\boldsymbol{A}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\right)}_{\text{The leading term}} + T_2 + T_4, \qquad \text{(B.4)}$$

where T_2 and T_4 are the second order and the fourth order term, respectively. More concretely, the second order term is

$$T_2 = \frac{M+1}{M}\mathbb{E}\mathrm{tr}\Big\{u_{-1}\boldsymbol{A}^\top\mathbf{H}\boldsymbol{b}\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^\top\mathbf{H} + u_{-1}\mathbf{u}_{12}\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\boldsymbol{b}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}$$

$$
\begin{aligned}
&+ u_{-1}\mathbf{u}_{12}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{b}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{A}\mathbf{H} + u_{-1}\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\mathbf{H} \\
&+ \mathbf{u}_{12}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{b}\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\mathbf{H} + u_{-1}^{2}\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{A}\mathbf{H}\Big\} \\
&+ \frac{1}{M}\mathbb{E}\Big\{u_{-1}^{2}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\cdot\operatorname{tr}\left(\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{A}\mathbf{H}\right) + \boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{b}\cdot\operatorname{tr}\left(\mathbf{u}_{12}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\mathbf{H}\right) \\
&+ 2u_{-1}\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\cdot\operatorname{tr}\left(\mathbf{u}_{12}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{A}\mathbf{H}\right) + 2u_{-1}\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\cdot\operatorname{tr}\left(\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\mathbf{H}\right)\Big\} \\
&= \frac{M+1}{M}\Big\{2u_{-1}\operatorname{tr}(\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}})\cdot\mathbf{u}_{12}^{\top}\mathbf{H}\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{b} + 2u_{-1}\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12} + \boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{b}\cdot\mathbf{u}_{12}^{\top}\mathbf{H}\mathbf{u}_{12} \\
&+ u_{-1}^{2}\operatorname{tr}\left(\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{A}\mathbf{H}\right)\Big\} + \frac{1}{M}\Big\{u_{-1}^{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\cdot\operatorname{tr}\left(\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{A}\mathbf{H}\right) \\
&+ \operatorname{tr}\left(\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\cdot\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{b}\cdot\mathbf{u}_{12}^{\top}\mathbf{H}\mathbf{u}_{12} + 4u_{-1}\cdot\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12}\Big\} \\
&= \frac{2(M+1)}{M}u_{-1}\boldsymbol{b}^{\top}\left[\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H} + \mathbf{H}\boldsymbol{\Psi}\mathbf{H}\right]\boldsymbol{A}\mathbf{H}\mathbf{u}_{12} + \boldsymbol{b}^{\top}\left(\frac{M+1}{M}\mathbf{H}\boldsymbol{\Psi}\mathbf{H} + \frac{1}{M}\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H}\right)\boldsymbol{b}\cdot\mathbf{u}_{12}^{\top}\mathbf{H}\mathbf{u}_{12} \\
&+ u_{-1}^{2}\operatorname{tr}\left(\boldsymbol{A}^{\top}\left(\frac{M+1}{M}\mathbf{H}\boldsymbol{\Psi}\mathbf{H} + \frac{1}{M}\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H}\right)\boldsymbol{A}\mathbf{H}\right) + \frac{4}{M}u_{-1}\cdot\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12}. \quad \text{(B.5)}
\end{aligned}
$$

The fourth order term is

$$
\begin{aligned}
T_4 = &\frac{M+1}{M}\mathbb{E}\operatorname{tr}\Big\{u_{-1}^{2}\mathbf{u}_{12}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\mathbf{H}\Big\} \\
&+ \frac{1}{M}\mathbb{E}\operatorname{tr}\Big\{u_{-1}^{2}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\cdot\operatorname{tr}\left(\mathbf{u}_{12}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\mathbf{H}\right)\Big\} \\
= &\frac{M+2}{M}u_{-1}^{2}\mathbf{u}_{12}^{\top}\mathbf{H}\mathbf{u}_{12}\cdot\mathbb{E}\left(\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)^{2} \\
= &\frac{M+2}{M}u_{-1}^{2}\cdot\left(2\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}\right) + \operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)^{2}\right)\cdot\mathbf{u}_{12}^{\top}\mathbf{H}\mathbf{u}_{12}. \qquad \text{(B.6)}
\end{aligned}
$$

For the cross term in S_1, we have

$$
\begin{aligned}
&\left\langle\mathbf{H}, \mathbb{E}\boldsymbol{z}_1\widetilde{\boldsymbol{\beta}}^{\top}\right\rangle \\
=&\mathbb{E}\Big\{\left(\mathbf{u}_{21} + u_{-1}\widetilde{\boldsymbol{\beta}}\right)^{\top}\cdot\frac{1}{M}\mathbf{X}^{\top}\mathbf{X}\cdot\left(\mathbf{U}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^{\top}\right)\mathbf{H}\widetilde{\boldsymbol{\beta}}\Big\} \qquad \text{(B.7)} \\
=&\mathbb{E}\Big\{\left(\mathbf{u}_{21} + u_{-1}\widetilde{\boldsymbol{\beta}}\right)^{\top}\mathbf{H}\left(\mathbf{U}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^{\top}\right)\mathbf{H}\widetilde{\boldsymbol{\beta}}\Big\} \qquad \text{(From the distribution of } \mathbf{X}\text{)} \\
=&\mathbb{E}\Big\{\left(\boldsymbol{b} + u_{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)^{\top}\mathbf{H}\left(\boldsymbol{A} + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\right)\mathbf{H}\left(\boldsymbol{\beta}^{*} + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)\Big\} \qquad \text{(By (B.3))} \\
=&\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{\beta}^{*} + \mathbb{E}\Big\{u_{-1}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\mathbf{H}\boldsymbol{\beta}^{*} + u_{-1}\widetilde{\boldsymbol{\theta}}^{\top}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}} + \boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^{\top}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\Big\} \\
=&\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{\beta}^{*} + u_{-1}\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)\cdot\mathbf{u}_{12}^{\top}\mathbf{H}\boldsymbol{\beta}^{*} + u_{-1}\operatorname{tr}\left(\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{\Psi}\right) + \mathbf{u}_{12}^{\top}\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{b}, \qquad \text{(B.8)}
\end{aligned}
$$

where the last row comes from the fact that $\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right)$. Moreover, we have

$$
\left\langle\mathbf{H}, \mathbb{E}\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^{\top}\right\rangle = \boldsymbol{\beta}^{*\top}\mathbf{H}\boldsymbol{\beta}^{*} + \operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right). \qquad \text{(B.9)}
$$

Combining (B.4), (B.5), (B.6), (B.8), (B.9), we have

$$
\begin{aligned}
S_1 = &\frac{M+1}{M}\operatorname{tr}\left(\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{b}\boldsymbol{b}^{\top}\mathbf{H}\boldsymbol{A}\mathbf{H}\right) + \frac{1}{M}\operatorname{tr}\left(\boldsymbol{b}\boldsymbol{b}^{\top}\mathbf{H}\right)\operatorname{tr}\left(\boldsymbol{A}^{\top}\mathbf{H}\boldsymbol{A}\mathbf{H}\right) \\
&+ \frac{M+2}{M}u_{-1}^{2}\cdot\left(2\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}\right) + \operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)^{2}\right)\cdot\mathbf{u}_{12}^{\top}\mathbf{H}\mathbf{u}_{12}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{2(M+1)}{M}u_{-1}\boldsymbol{b}^\top\left[\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H}+\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\right]\boldsymbol{A}\mathbf{H}\mathbf{u}_{12}\\
&+u_{-1}^2\operatorname{tr}\left(\boldsymbol{A}^\top\left(\frac{M+1}{M}\mathbf{H}\boldsymbol{\Psi}\mathbf{H}+\frac{1}{M}\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H}\right)\boldsymbol{A}\mathbf{H}\right)+\frac{4}{M}u_{-1}\cdot\boldsymbol{b}^\top\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12}\\
&-2\bigg[\boldsymbol{b}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{\beta}^*+u_{-1}\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)\cdot\mathbf{u}_{12}^\top\mathbf{H}\boldsymbol{\beta}^*+u_{-1}\operatorname{tr}\left(\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{\Psi}\right)\\
&+\mathbf{u}_{12}^\top\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{b}\bigg]+\boldsymbol{\beta}^{*\top}\mathbf{H}\boldsymbol{\beta}^*+\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)+\boldsymbol{b}^\top\left(\frac{M+1}{M}\mathbf{H}\boldsymbol{\Psi}\mathbf{H}+\frac{1}{M}\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H}\right)\boldsymbol{b}\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}.
\end{aligned}
\tag{B.10}
$$

Step 3: other terms. Let us compute S_2, S_3 and S_4. Using definitions, we rewrite $\boldsymbol{z}_2$ as

$$
\begin{aligned}
\boldsymbol{z}_2^\top&=\mathbf{u}_{21}^\top\cdot\frac{1}{M}\mathbf{X}^\top\boldsymbol{\varepsilon}\cdot\mathbf{u}_{12}^\top+u_{-1}\cdot\frac{1}{M}\boldsymbol{\varepsilon}^\top\mathbf{X}\cdot\mathbf{U}_{11}+u_{-1}\cdot\frac{2}{M}\boldsymbol{\varepsilon}^\top\mathbf{X}\widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\\
&=\boldsymbol{b}^\top\cdot\frac{1}{M}\mathbf{X}^\top\boldsymbol{\varepsilon}\cdot\mathbf{u}_{12}^\top+u_{-1}\cdot\frac{1}{M}\boldsymbol{\varepsilon}^\top\mathbf{X}\cdot\boldsymbol{A}+u_{-1}\cdot\frac{2}{M}\boldsymbol{\varepsilon}^\top\mathbf{X}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^\top.
\end{aligned}
\tag{B.11}
$$

Since $\boldsymbol{\varepsilon}\sim\mathcal{N}\left(\mathbf{0},\sigma^2\mathbf{I}_M\right),\widetilde{\boldsymbol{\theta}}\sim\mathcal{N}\left(\mathbf{0},\mathbf{I}_d\right)$ and they are independent, all terms in S_2 vanish except if the term contains even orders of $\widetilde{\boldsymbol{\theta}}$ or $\boldsymbol{\varepsilon}$. So we have

$$
\begin{aligned}
S_2=\left\langle\mathbf{H},\mathbb{E}\boldsymbol{z}_2\boldsymbol{z}_2^\top\right\rangle=\frac{1}{M^2}\mathbb{E}\bigg\{&\boldsymbol{b}^\top\mathbf{X}^\top\boldsymbol{\varepsilon}\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}\cdot\boldsymbol{\varepsilon}^\top\mathbf{X}\cdot\boldsymbol{b}+u_{-1}^2\boldsymbol{\varepsilon}^\top\mathbf{X}\boldsymbol{A}\mathbf{H}\boldsymbol{A}^\top\mathbf{X}^\top\boldsymbol{\varepsilon}\\
&+2u_{-1}\cdot\boldsymbol{\varepsilon}^\top\mathbf{X}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12}\cdot\boldsymbol{\varepsilon}^\top\mathbf{X}\boldsymbol{b}+4u_{-1}^2\cdot\boldsymbol{\varepsilon}^\top\mathbf{X}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}\cdot\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{X}^\top\boldsymbol{\varepsilon}\bigg\}\\
=\frac{\sigma^2}{M}\bigg\{&\boldsymbol{b}^\top\mathbf{H}\boldsymbol{b}\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}+u_{-1}^2\operatorname{tr}\left(\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{A}^\top\right)+2u_{-1}\mathbf{u}_{12}^\top\mathbf{H}\boldsymbol{A}^\top\mathbf{H}\boldsymbol{b}+4u_{-1}^2\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}\bigg\}
\end{aligned}
\tag{B.12}
$$

For the other two terms, we have

$$
S_3=\frac{1}{M^2}u_{-1}^2\mathbb{E}\{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\}=\frac{\sigma^4(M+2)}{M}u_{-1}^2\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}
\tag{B.13}
$$

and

$$
\begin{aligned}
S_4&=2\left\langle\mathbf{H},\mathbb{E}\left(\boldsymbol{z}_1-\widetilde{\boldsymbol{\beta}}\right)\boldsymbol{z}_3^\top\right\rangle\\
&=2\mathbb{E}\bigg\{\bigg[\left(\mathbf{u}_{21}+u_{-1}\widetilde{\boldsymbol{\beta}}\right)^\top\cdot\frac{1}{M}\mathbf{X}^\top\mathbf{X}\cdot\left(\mathbf{U}_{11}+\widetilde{\boldsymbol{\beta}}\mathbf{u}_{12}^\top\right)-\widetilde{\boldsymbol{\beta}}^\top\bigg]\mathbf{H}\cdot\frac{1}{M}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\cdot u_{-1}\mathbf{u}_{12}^\top\bigg\}\\
&=2\sigma^2u_{-1}\mathbb{E}\bigg\{\bigg[\left(\boldsymbol{b}+u_{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)^\top\cdot\mathbf{H}\cdot\left(\boldsymbol{A}+\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\mathbf{u}_{12}^\top\right)-\widetilde{\boldsymbol{\beta}}^{*\top}-\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{\Psi}^{\frac{1}{2}}\bigg]\mathbf{H}\mathbf{u}_{12}\bigg\}\\
&=2\sigma^2u_{-1}\left[\boldsymbol{b}^\top\mathbf{H}\boldsymbol{A}-\widetilde{\boldsymbol{\beta}}^{*\top}\right]\mathbf{H}\mathbf{u}_{12}+2\sigma^2u_{-1}^2\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}.
\end{aligned}
\tag{B.14}
$$

Step 4: combine all parts. Combining the four parts (B.10), (B.12), (B.13) and (B.14), we have

$$
\mathcal{R}\left(f\right)-\sigma^2=S_1+S_2+S_3+S_4
$$

We remark that in the (B.10), (B.12), (B.13) and (B.14), the parameters are $\boldsymbol{A},\boldsymbol{b},u_{-1}$ and $\mathbf{u}_{12}$, instead of the original parameter $\mathbf{U}_{11},\mathbf{u}_{12},\mathbf{u}_{21},u_{-1}$. From the definitions of $\boldsymbol{A}$ and $\boldsymbol{b}$, we know there is a bijective map between $\left(\mathbf{U}_{11},\mathbf{u}_{12},\mathbf{u}_{21},u_{-1}\right)$ and $\left(\boldsymbol{A},\boldsymbol{b},\mathbf{u}_{12},u_{-1}\right)$, so these two parameterizations are equivalent when computing the minimal risk achieved by a model in a hypothesis class. From the expression of S_1,S_2,s_3,S_4, we can rewrite the risk as

$$
\mathcal{R}\left(f\right)-\sigma^2
$$

$$
\begin{aligned}
&= \boldsymbol{b}^\top \left(\frac{M+1}{M}\mathbf{H\Psi H} + \frac{1}{M}\mathrm{tr}(\mathbf{H\Psi})\mathbf{H}\right)\boldsymbol{b}\cdot \mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12} + \frac{\sigma^2}{M}\cdot\boldsymbol{b}^\top\mathbf{H}\boldsymbol{b}\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}\\
&\quad + u_{-1}^2\mathrm{tr}\left(\boldsymbol{A}^\top\left(\frac{M+1}{M}\mathbf{H\Psi H}+\frac{1}{M}\mathrm{tr}(\mathbf{H\Psi})\mathbf{H}\right)\boldsymbol{A}\mathbf{H}\right) + \frac{\sigma^2}{M}\cdot\mathrm{tr}\left(\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{A}^\top\right)\\
&\quad \underbrace{+2u_{-1}\boldsymbol{b}^\top\left[\frac{1}{M}\mathrm{tr}(\mathbf{H\Psi})\mathbf{H}+\frac{M+1}{M}\mathbf{H\Psi H}+\frac{\sigma^2}{M}\cdot\mathbf{H}\right]\boldsymbol{A}\mathbf{H}\mathbf{u}_{12} - 2u_{-1}\mathrm{tr}\left(\mathbf{H}\boldsymbol{A}\mathbf{H\Psi}\right) - 2\mathbf{u}_{12}^\top\mathbf{H\Psi H}\boldsymbol{b}}_{\mathrm{I}}\\
&+ \underbrace{\frac{1}{M}\left[\boldsymbol{b}^\top\mathbf{H}\boldsymbol{b}\cdot\mathrm{tr}\left(\boldsymbol{A}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\right)+4u_{-1}\cdot\boldsymbol{b}^\top\mathbf{H\Psi H}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12}+4u_{-1}^2\cdot\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right)\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}\right]}_{\mathrm{II}}\\
&+ \underbrace{\boldsymbol{b}^\top\left(\frac{M+1}{M}\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{A}^\top\mathbf{H}\right)\boldsymbol{b} - 2\boldsymbol{b}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{\beta}^* + 2u_{-1}\mathrm{tr}(\mathbf{H\Psi})\cdot\boldsymbol{b}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12} + 2\sigma^2\boldsymbol{b}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\left(u_{-1}\mathbf{u}_{12}\right)}_{\mathrm{III}}\\
&+ \boldsymbol{\beta}^{*\top}\mathbf{H}\boldsymbol{\beta}^* - 2\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)\cdot\boldsymbol{\beta}^{*\top}\mathbf{H}\left(u_{-1}\mathbf{u}_{12}\right)+\mathrm{tr}\left(\mathbf{H\Psi}\right)\\
&\quad \underbrace{+u_{-1}^2\cdot\left[\left(2\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right)+\frac{M+2}{M}\mathrm{tr}\left(\mathbf{H\Psi}\right)^2\right)\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}+\left(2+\frac{4}{M}\right)\sigma^2\mathrm{tr}\left(\mathbf{H\Psi}\right)+\frac{M+2}{M}\sigma^4\right]\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}}_{\mathrm{IV}}\\
&= \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}, \qquad\qquad (\mathrm{B.15})
\end{aligned}
$$

where I, II, III, IV are defined as above.

Step 5: lower bound the risk function. Let's first define a new matrix, which by definition is invertible:

$$
\boldsymbol{\Omega} := \frac{M+1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} + \frac{\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2}{M}\mathbf{I}_d. \qquad (\mathrm{B.16})
$$

So we can write I as

$$
\begin{aligned}
\mathrm{I} &= \boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}}\boldsymbol{b}\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12} + u_{-1}^2\mathrm{tr}\left(\boldsymbol{A}^\top\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}\right) + 2u_{-1}\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12}\\
&\quad - 2u_{-1}\mathrm{tr}\left(\mathbf{H}\boldsymbol{A}\mathbf{H\Psi}\right) - 2\mathbf{u}_{12}^\top\mathbf{H\Psi H}\boldsymbol{b}\\
&\quad = \mathrm{tr}\Bigg[\left(\mathbf{H}^{\frac{1}{2}}\mathbf{u}_{12}\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}} + u_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)\boldsymbol{\Omega}\\
&\qquad\qquad \cdot\left(\mathbf{H}^{\frac{1}{2}}\mathbf{u}_{12}\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}} + u_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)^\top\Bigg] - \mathrm{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)\\
&\quad \geq -\mathrm{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right) \qquad\qquad (\mathrm{B.17})\\
&\quad = -\mathrm{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\boldsymbol{\Omega}^{-1}\right),
\end{aligned}
$$

where the last line comes from the fact that $\boldsymbol{\Omega}$ and $\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}$ commute.

Next, we claim II ≥ 0. To see this, it suffices to notice that

$$
\begin{aligned}
4u_{-1}\cdot\boldsymbol{b}^\top\mathbf{H\Psi H}\boldsymbol{A}\mathbf{H}\mathbf{u}_{12} &\geq -4\left\|\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\right\|_F\cdot\left\|u_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right\|_F\cdot\left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}}\right\|_F\cdot\left\|\mathbf{H}^{\frac{1}{2}}\mathbf{u}_{12}\right\|_F\\
&\geq -\left\|\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\right\|_F^2\left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}}\right\|_F^2 - 4\left\|u_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right\|_F^2\left\|\mathbf{H}^{\frac{1}{2}}\mathbf{u}_{12}\right\|_F^2\\
&\geq -\boldsymbol{b}^\top\mathbf{H}\boldsymbol{b}\cdot\mathrm{tr}\left(\boldsymbol{A}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\right) - 4u_{-1}^2\cdot\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right)\cdot\mathbf{u}_{12}^\top\mathbf{H}\mathbf{u}_{12}, \quad (\mathrm{B.18})
\end{aligned}
$$

where the last line comes from the fact that $\|\boldsymbol{A}\|_F^2 = \mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^\top\right)$ for any matrix $\boldsymbol{A}$.

Then, let's consider III. Notice that actually, III can be viewed as a quadratic function of $\boldsymbol{A}^\top \mathbf{H}\boldsymbol{b}$, which can be easily minimized. More concretely, we have

$$\begin{aligned}
&\mathrm{III} + \frac{M}{M+1}\Big(\boldsymbol{\beta}^* - \left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)u_{-1}\mathbf{u}_{12}\Big)^\top \mathbf{H}\Big(\boldsymbol{\beta}^* - \left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)u_{-1}\mathbf{u}_{12}\Big)\\
&= \left[\boldsymbol{A}^\top\mathbf{H}\boldsymbol{b} - \frac{M}{M+1}\Big(\boldsymbol{\beta}^* - \left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)u_{-1}\mathbf{u}_{12}\Big)\right]^\top \left(\frac{M+1}{M}\mathbf{H}\right)\\
&\qquad\qquad \cdot \left[\boldsymbol{A}^\top\mathbf{H}\boldsymbol{b} - \frac{M}{M+1}\Big(\boldsymbol{\beta}^* - \left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)u_{-1}\mathbf{u}_{12}\Big)\right]\\
&\geq 0. \qquad\qquad \text{(B.19)}
\end{aligned}$$

Therefore, one has

$$\mathrm{III} \geq -\frac{M}{M+1}\Big(\boldsymbol{\beta}^* - \left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)u_{-1}\mathbf{u}_{12}\Big)^\top \mathbf{H}\Big(\boldsymbol{\beta}^* - \left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)u_{-1}\mathbf{u}_{12}\Big).$$

Combining the three parts above, one has

$$\begin{aligned}
&\mathcal{R}\left(f\right) - \sigma^2\\
&\geq \mathrm{IV} - \mathrm{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\mathbf{\Omega}^{-1}\right)\\
&\qquad - \frac{M}{M+1}\Big(\boldsymbol{\beta}^* - \left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)u_{-1}\mathbf{u}_{12}\Big)^\top \mathbf{H}\Big(\boldsymbol{\beta}^* - \left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)u_{-1}\mathbf{u}_{12}\Big)\\
&= \underbrace{\mathrm{tr}\left(\mathbf{H\Psi}\right) - \mathrm{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\mathbf{\Omega}^{-1}\right)}_{\inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}(f)-\sigma^2} + \frac{1}{M+1}\boldsymbol{\beta}^{*\top}\mathbf{H}\boldsymbol{\beta}^* - \frac{2\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)}{M+1}\boldsymbol{\beta}^{*\top}\mathbf{H}\left(u_{-1}\mathbf{u}_{12}\right)\\
&\qquad + \left[2\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right) + \frac{3M+2}{M(M+1)}\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2\right]\left(u_{-1}\mathbf{u}_{12}\right)^\top\mathbf{H}\left(u_{-1}\mathbf{u}_{12}\right).
\end{aligned}$$

Therefore, one has

$$\begin{aligned}
\mathcal{R}\left(f\right) - \inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}(f) \geq{}& \frac{1}{M+1}\boldsymbol{\beta}^{*\top}\mathbf{H}\boldsymbol{\beta}^* - \frac{2\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)}{M+1}\boldsymbol{\beta}^{*\top}\mathbf{H}\left(u_{-1}\mathbf{u}_{12}\right)\\
&+ \left[2\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right) + \frac{3M+2}{M(M+1)}\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2\right]\left(u_{-1}\mathbf{u}_{12}\right)^\top\mathbf{H}\left(u_{-1}\mathbf{u}_{12}\right).
\end{aligned}$$

The right hand side in the above inequality is a quadratic function of $u_{-1}\mathbf{u}_{12}$, so we can take its global minimizer:

$$u_{-1}\mathbf{u}_{12} = \frac{\frac{\left(\mathrm{tr}(\mathbf{H\Psi})+\sigma^2\right)}{M+1}}{2\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right) + \frac{3M+2}{M(M+1)}\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2}\boldsymbol{\beta}^*$$

to lower bound the risk gap as

$$\mathcal{R}\left(f\right) - \inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}(f) \geq \left[\frac{1}{M+1} - \frac{\frac{\left(\mathrm{tr}(\mathbf{H\Psi})+\sigma^2\right)^2}{(M+1)^2}}{2\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right) + \frac{3M+2}{M(M+1)}\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2}\right]\left\|\boldsymbol{\beta}^*\right\|_{\mathbf{H}}^2.$$

Finally, to simplify the results, we notice that on the one hand,

$$\frac{\frac{\left(\mathrm{tr}(\mathbf{H\Psi})+\sigma^2\right)^2}{(M+1)^2}}{2\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right) + \frac{3M+2}{M(M+1)}\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2} \leq \frac{\left(\mathrm{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2}{2(M+1)^2\mathrm{tr}\left(\mathbf{H\Psi H\Psi}\right)}.$$

On the other hand, one has

$$\frac{\frac{\left(\operatorname{tr}(\mathbf{H\Psi})+\sigma^2\right)^2}{(M+1)^2}}{2\operatorname{tr}\left(\mathbf{H\Psi H\Psi}\right)+\frac{3M+2}{M(M+1)}\left(\operatorname{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2}\leq\frac{M}{(M+1)(3M+2)}\leq\frac{1}{3(M+1)}.$$

Therefore, we have

$$\mathcal{R}\left(f\right)-\inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}(f)\geq\max\left\{\frac{2}{3(M+1)},\frac{1}{M+1}-\frac{\left(\operatorname{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2}{2(M+1)^2\operatorname{tr}\left(\mathbf{H\Psi H\Psi}\right)}\right\}\cdot\|\boldsymbol{\beta}^*\|_{\mathbf{H}}^2.$$

When $\frac{1}{M+1}-\frac{\left(\operatorname{tr}(\mathbf{H\Psi})+\sigma^2\right)^2}{2(M+1)^2\operatorname{tr}(\mathbf{H\Psi H\Psi})}\geq\frac{2}{3(M+1)}$, we have $\frac{1}{M+1}\geq\frac{3\left(\operatorname{tr}(\mathbf{H\Psi})+\sigma^2\right)^2}{2(M+1)^2\operatorname{tr}(\mathbf{H\Psi H\Psi})}$, which implies

$$\mathcal{R}\left(f\right)-\inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}(f)\geq\left[\frac{1}{M+1}-\frac{\left(\operatorname{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2}{2(M+1)^2\operatorname{tr}\left(\mathbf{H\Psi H\Psi}\right)}\right]\cdot\|\boldsymbol{\beta}^*\|_{\mathbf{H}}^2\geq\frac{\left(\operatorname{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2}{(M+1)^2\operatorname{tr}\left(\mathbf{H\Psi H\Psi}\right)}\cdot\|\boldsymbol{\beta}^*\|_{\mathbf{H}}^2.$$

Therefore, we finally have

$$\mathcal{R}\left(f\right)-\inf_{f\in\mathcal{F}_{\mathsf{LTB}}}\mathcal{R}(f)\geq\max\left\{\frac{2}{3(M+1)},\frac{\left(\operatorname{tr}\left(\mathbf{H\Psi}\right)+\sigma^2\right)^2}{(M+1)^2\operatorname{tr}\left(\mathbf{H\Psi H\Psi}\right)}\right\}\cdot\|\boldsymbol{\beta}^*\|_{\mathbf{H}}^2.$$

Note that this holds for an arbitrary $f\in\mathcal{F}_{\mathsf{LSA}}$, so the proof finishes by taking infimum on the left hand side. □

C Proof of Theorem 5.2

Proof. Recall that from Assumption 3.1,

$$\mathbf{x} \sim \mathcal{N}(0, \mathbf{H}), \quad y = \mathbf{x}^\top \widetilde{\boldsymbol{\beta}} + \varepsilon, \quad \mathbf{X}[i] \sim \mathcal{N}(0, \mathbf{H}), \quad \mathbf{y} = \mathbf{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon},$$

where

$$\widetilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}^*, \boldsymbol{\Psi}\right), \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_M\right).$$

Step 1: compute the risk function. From the independence of $\varepsilon, \boldsymbol{\varepsilon}$ with other random variables, we have

$$\begin{aligned}
\mathcal{R}_M(\boldsymbol{\beta}, \boldsymbol{\Gamma}) &= \mathbb{E}\left(\mathbf{x}^\top\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \varepsilon + \mathbf{x}^\top \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \mathbf{X}}{M}\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right) - \mathbf{x}^\top \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \boldsymbol{\varepsilon}}{M}\right)^2 \\
&= \mathbb{E}\left(\mathbf{x}^\top\left(\mathbf{I}_d - \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \mathbf{X}}{M}\right) \cdot \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2 + \frac{1}{M^2}\mathbb{E}\left(\mathbf{x}^\top \boldsymbol{\Gamma} \mathbf{X}^\top \boldsymbol{\varepsilon}\right)^2 + \sigma^2 \\
&= \left\langle \mathbf{H}, \mathbb{E}\left(\mathbf{I}_d - \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \mathbf{X}}{M}\right)^{\otimes 2} \circ \mathbb{E}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\otimes 2}\right\rangle + \frac{1}{M^2}\mathbb{E}\left(\mathbf{x}^\top \boldsymbol{\Gamma} \mathbf{X}^\top \boldsymbol{\varepsilon}\right)^2 + \sigma^2,
\end{aligned} \tag{C.1}$$

where the last line comes from the independence between $\widetilde{\boldsymbol{\beta}}$ and $\mathbf{X}, \mathbf{x}$, as well as the fact that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{H})$ and the property of tensor product $\circ$. Note that, $\boldsymbol{\beta}$ ad $\boldsymbol{\Gamma}$ are learnable parameters here, and we should set them apart from the task vector $\widetilde{\boldsymbol{\beta}}$ or the prior mean vector $\widetilde{\boldsymbol{\beta}}^*$. Recall that $\widetilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}^*, \boldsymbol{\Psi}\right)$, we have $\mathbb{E}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\otimes 2} = \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^{\otimes 2} + \boldsymbol{\Psi}$. Moreover, using the following

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{H}), \quad \mathbf{X}[i] \sim \mathcal{N}(\mathbf{0}, \mathbf{H}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_M\right),$$

we have that

$$\mathbb{E}\left(\mathbf{x}^\top \boldsymbol{\Gamma} \mathbf{X}^\top \boldsymbol{\varepsilon}\right)^2 = \sigma^2 \mathbb{E}\mathrm{tr}\left(\mathbf{X}\boldsymbol{\Gamma}^\top \mathbf{x}\mathbf{x}^\top \boldsymbol{\Gamma} \mathbf{X}^\top\right) = M\sigma^2 \cdot \left\langle \mathbf{H}\boldsymbol{\Gamma}\mathbf{H}, \boldsymbol{\Gamma}^\top\right\rangle.$$

Bridging the two terms above into (C.1), we get

$$\begin{aligned}
\mathcal{R}_M(\boldsymbol{\beta}, \boldsymbol{\Gamma}) &= \left\langle \mathbf{H}, \mathbb{E}\left(\mathbf{I}_d - \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \mathbf{X}}{M}\right)^{\otimes 2} \circ \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^{\otimes 2}\right\rangle \\
&\quad + \left\langle \mathbf{H}, \mathbb{E}\left(\mathbf{I}_d - \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \mathbf{X}}{M}\right)^{\otimes 2} \circ \boldsymbol{\Psi} + \frac{\sigma^2}{M}\boldsymbol{\Gamma}\mathbf{H}\boldsymbol{\Gamma}^\top\right\rangle + \sigma^2 \\
&= \underbrace{\left\langle \mathbb{E}\left(\mathbf{I}_d - \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \mathbf{X}}{M}\right)^{\otimes 2} \circ \mathbf{H}, \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^{\otimes 2}\right\rangle}_{V_1} \\
&\quad + \underbrace{\left\langle \mathbf{H}, \mathbb{E}\left(\left(\mathbf{I}_d - \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \mathbf{X}}{M}\right)^\top\right)^{\otimes 2} \circ \boldsymbol{\Psi} + \frac{\sigma^2}{M}\boldsymbol{\Gamma}\mathbf{H}\boldsymbol{\Gamma}^\top\right\rangle}_{V_2} + \sigma^2
\end{aligned} \tag{C.2}$$

First, let's compute V_1. From the definition above, we have

$$V_1 = \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H}_{\boldsymbol{\Gamma}} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right),$$

where

$$\mathbf{H}_{\boldsymbol{\Gamma}} := \mathbb{E}\left(\left(\mathbf{I}_d - \boldsymbol{\Gamma} \cdot \frac{\mathbf{X}^\top \mathbf{X}}{M}\right)^\top\right)^{\otimes 2} \circ \mathbf{H} \tag{C.3}$$

$$= \mathbb{E}\left(\mathbf{I}_d - \mathbf{\Gamma}\cdot\frac{\mathbf{X}^\top\mathbf{X}}{M}\right)^\top \mathbf{H}\left(\mathbf{I}_d - \mathbf{\Gamma}\cdot\frac{\mathbf{X}^\top\mathbf{X}}{M}\right) \qquad \text{(the definition of the tensor product)}$$

$$= \mathbf{H} - \mathbf{H}\left(\mathbf{\Gamma}+\mathbf{\Gamma}^\top\right)\mathbf{H} + \frac{\operatorname{tr}\left(\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\right)}{M}\mathbf{H} + \frac{M+1}{M}\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\mathbf{H}$$

$$(\mathbf{X}[i] \sim \mathcal{N}(\mathbf{0},\mathbf{H}) \text{ and Lemma H.4})$$

$$= \left(\mathbf{I}_d - \mathbf{\Gamma}\mathbf{H}\right)^\top \mathbf{H}\left(\mathbf{I}_d - \mathbf{\Gamma}\mathbf{H}\right) + \frac{\operatorname{tr}\left(\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\right)}{M}\mathbf{H} + \frac{1}{M}\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\mathbf{H} \succeq \mathbf{0}. \qquad \text{(C.4)}$$

Then, let's compute V_2. We have

$$\mathbb{E}\left(\mathbf{I}_d - \mathbf{\Gamma}\cdot\frac{\mathbf{X}^\top\mathbf{X}}{M}\right)^{\otimes 2}\circ\mathbf{\Psi} + \frac{\sigma^2}{M}\mathbf{\Gamma}\mathbf{H}\mathbf{\Gamma}^\top = \mathbb{E}\left(\mathbf{I}_d - \mathbf{\Gamma}\cdot\frac{\mathbf{X}^\top\mathbf{X}}{M}\right)\mathbf{\Psi}\left(\mathbf{I}_d - \mathbf{\Gamma}\cdot\frac{\mathbf{X}^\top\mathbf{X}}{M}\right)^\top + \frac{\sigma^2}{M}\mathbf{\Gamma}\mathbf{H}\mathbf{\Gamma}^\top$$

$$= \mathbf{\Psi} - \mathbf{\Gamma}\mathbf{H}\mathbf{\Psi} - \mathbf{\Psi}\mathbf{H}\mathbf{\Gamma}^\top + \mathbf{\Gamma}\left(\frac{\operatorname{tr}(\mathbf{H}\mathbf{\Psi})+\sigma^2}{M}\cdot\mathbf{H} + \frac{M+1}{M}\mathbf{H}\mathbf{\Psi}\mathbf{H}\right)\mathbf{\Gamma}^\top.$$

$$(\mathbf{X}[i] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0},\mathbf{H}) \text{ and the Lemma H.4})$$

From the definition of $\mathbf{\Omega}$ in (A.3), we know that it is invertible. Moreover, it holds that

$$\mathbb{E}\left(\mathbf{I}_d - \mathbf{\Gamma}\cdot\frac{\mathbf{X}^\top\mathbf{X}}{M}\right)^{\otimes 2}\circ\mathbf{\Psi} + \frac{\sigma^2}{M}\mathbf{\Gamma}\mathbf{H}\mathbf{\Gamma}^\top = \mathbf{\Psi} - \mathbf{\Gamma}\mathbf{H}\mathbf{\Psi} - \mathbf{\Psi}\mathbf{H}\mathbf{\Gamma}^\top + \mathbf{\Gamma}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}^\top$$

$$= \left(\mathbf{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\right)\mathbf{\Omega}\left(\mathbf{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\right)^\top + \mathbf{\Psi} - \mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}.$$

Therefore, we have

$$V_2 = \operatorname{tr}\left[\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\right)\mathbf{\Omega}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\right)^\top\right]$$
$$+ \operatorname{tr}(\mathbf{H}\mathbf{\Psi}) - \operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right)$$

Step 2: solve the global minimum. Now we have

$$\mathcal{R}(\boldsymbol{\beta},\mathbf{\Gamma}) - \sigma^2$$
$$= (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\left[\left(\mathbf{I}_d - \mathbf{\Gamma}\mathbf{H}\right)^\top \mathbf{H}\left(\mathbf{I}_d - \mathbf{\Gamma}\mathbf{H}\right) + \frac{\operatorname{tr}\left(\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\right)}{M}\mathbf{H} + \frac{1}{M}\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\mathbf{H}\right](\boldsymbol{\beta}-\boldsymbol{\beta}^*)$$
$$+ \operatorname{tr}\left[\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\right)\mathbf{\Omega}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\right)^\top\right]$$
$$+ \operatorname{tr}(\mathbf{H}\mathbf{\Psi}) - \operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) \qquad \text{(C.5)}$$
$$\geq \operatorname{tr}(\mathbf{H}\mathbf{\Psi}) - \operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right)$$
$$= \operatorname{tr}(\mathbf{H}\mathbf{\Psi}) - \operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\mathbf{\Omega}^{-1}\right),$$

where the last line comes from the fact that $\mathbf{\Omega}$ and $\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}$ commute. Taking infimum over all $\boldsymbol{\beta},\mathbf{\Gamma}$ in the left hand side, we have

$$\inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}}\mathcal{R}(f) - \sigma^2 = \inf_{\boldsymbol{\beta},\mathbf{\Gamma}}\mathcal{R}(\boldsymbol{\beta},\mathbf{\Gamma}) - \sigma^2 \geq \operatorname{tr}(\mathbf{H}\mathbf{\Psi}) - \operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\mathbf{\Omega}^{-1}\right).$$

On the other hand, we take

$$\boldsymbol{\beta} = \boldsymbol{\beta}^*, \quad \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^* := \boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}, \tag{C.6}$$

where $\mathbf{H}^{\frac{1}{2}}$ is the principle square root of $\mathbf{H}$ and $\mathbf{H}^{-\frac{1}{2}}$ is its Moore-Penrose pseudo inverse (see notation part in Section A). Then, we have

$$\begin{aligned}
&\mathcal{R}\left(\boldsymbol{\beta}^*, \boldsymbol{\Gamma}^*\right) - \sigma^2 \\
&= \operatorname{tr}\left[\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)\boldsymbol{\Omega}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)^\top\right] \\
&+ \operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) - \operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right).
\end{aligned}$$

Since

$$\begin{aligned}
\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1} &= \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1} \\
&= \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} - \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} \\
&\qquad (\boldsymbol{\Omega} \text{ and } \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} \text{ commute}) \\
&= \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} - \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} = \mathbf{0}_{d\times d}.. \\
&\qquad (\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}})
\end{aligned}$$

Therefore, we have

$$\mathcal{R}\left(\boldsymbol{\beta}^*, \boldsymbol{\Gamma}^*\right) - \sigma^2 = \operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) - \operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\boldsymbol{\Omega}^{-1}\right) \geq \inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) - \sigma^2 = \inf_{\boldsymbol{\beta},\boldsymbol{\Gamma}} \mathcal{R}\left(\boldsymbol{\beta}, \boldsymbol{\Gamma}\right) - \sigma^2.$$

Combining both directions, we conclude that

$$\inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) - \sigma^2 = \inf_{\boldsymbol{\beta},\boldsymbol{\Gamma}} \mathcal{R}\left(\boldsymbol{\beta}, \boldsymbol{\Gamma}\right) - \sigma^2 = \operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) - \operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\boldsymbol{\Omega}^{-1}\right). \tag{C.7}$$

Step 3: identify all global minimizers Above we show that $\mathcal{R}\left(\boldsymbol{\beta}^*, \boldsymbol{\Gamma}^*\right)$ achieves the global minimum of the ICL risk over GD-$\boldsymbol{\beta}$ class. Now we will figure out the sufficient and necessary condition to achieve the global minimal risk. From the proof above, we know that for any $\boldsymbol{\beta} \in \mathbb{R}^d, \boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}$, it holds that

$$\mathcal{R}\left(\boldsymbol{\beta}, \boldsymbol{\Gamma}\right) = \inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) \iff \begin{cases} V_1 = \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H}_{\boldsymbol{\Gamma}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right) = 0 \\ \operatorname{tr}\left[\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)\boldsymbol{\Omega}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)^\top\right] = 0 \end{cases}$$

Since $\boldsymbol{\Omega}$ is invertible, the second equation is equivalent to

$$\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1} = \mathbf{0}_{d\times d},$$

which is a linear system of $\boldsymbol{\Gamma}$. Since $\boldsymbol{\Gamma}^*$ is proved to be one solution, all solutions of this equation is

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^* + \left\{\boldsymbol{Z} \in \mathbb{R}^{d\times d} : \mathbf{H}^{\frac{1}{2}}\boldsymbol{Z}\mathbf{H}^{\frac{1}{2}} = \mathbf{0}_{d\times d}\right\} = \boldsymbol{\Gamma}^* + \mathsf{Im}\left(\mathbf{H}^{\otimes 2}\right),$$

where the last equation comes from Lemma H.6. Here, $+$ denotes Minkowski sum. This is defined as $A + B = \{a + b, a \in A, b \in B\}$ for two sets A, B and $a + B = \{a\} + B$ for an entry a and a set B. Under this condition, we know that

$$\mathbf{H}_{\boldsymbol{\Gamma}} = \mathbf{H} - \mathbf{H}\left(\boldsymbol{\Gamma}^* + \boldsymbol{\Gamma}_M^{*\top}\right)\mathbf{H} + \frac{\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Gamma}_M^{*\top}\mathbf{H}\boldsymbol{\Gamma}^*\right)}{M}\mathbf{H} + \frac{M+1}{M}\mathbf{H}\boldsymbol{\Gamma}_M^{*\top}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}.$$

Consider the following inequality:

$$\mathbf{I}_d - \mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\Gamma}^* + \boldsymbol{\Gamma}^{*\top}\right)\mathbf{H}^{\frac{1}{2}} + \frac{\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Gamma}_M^{*\top}\mathbf{H}\boldsymbol{\Gamma}^*\right)}{M}\mathbf{I}_d + \frac{M+1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}_M^{*\top}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}}$$

$$
\begin{aligned}
&\succeq \mathbf{I}_d - \mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\Gamma}^* + \boldsymbol{\Gamma}^{*\top}\right)\mathbf{H}^{\frac{1}{2}} + \frac{M+1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^{*\top}_{M}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}} \\
&= \left(\sqrt{\frac{M}{M+1}}\mathbf{I}_d - \sqrt{\frac{M+1}{M}}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right)\cdot\left(\sqrt{\frac{M}{M+1}}\mathbf{I}_d - \sqrt{\frac{M+1}{M}}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right)^\top + \frac{M}{M+1}\mathbf{I}_d \\
&\succ \mathbf{0}_{d\times d}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \mathbf{H}_{\boldsymbol{\Gamma}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right) = 0 \\
\iff &\left[\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}^{\frac{1}{2}}\right]\left(\mathbf{I}_d - \mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\Gamma}^* + \boldsymbol{\Gamma}^{*\top}\right)\mathbf{H}^{\frac{1}{2}} + \frac{\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Gamma}^{*\top}_{M}\mathbf{H}\boldsymbol{\Gamma}^*\right)}{M}\mathbf{I}_d\right. \\
&\left. + \frac{M+1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^{*\top}_{M}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right)\left[\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\right] = 0 \\
\iff &\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right) = \mathbf{0}_d \iff \boldsymbol{\beta} = \boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}^{\frac{1}{2}}\right) \iff \boldsymbol{\beta} = \boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}\right),
\end{aligned}
$$

where $\mathsf{null}\left(\cdot\right)$ denotes the null space of a matrix and $+$ denotes the Minkowski sum. The last equivalence comes from Lemma H.6. Therefore, we conclude that the $f_{\boldsymbol{\beta},\boldsymbol{\Gamma}}$ achieves the minimal ICL risk in GD-$\boldsymbol{\beta}$ class if and only if

$$
\boldsymbol{\beta} = \boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}\right), \quad \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^* + \mathsf{Im}\left(\mathbf{H}^{\otimes 2}\right).
$$

Specially, if $\mathbf{H}$ is positive definite, there is unique global minimizer in GD-$\boldsymbol{\beta}$ class with parameters

$$
\boldsymbol{\beta} = \boldsymbol{\beta}^*, \quad \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^*.
$$

Step 4: equivalence in the hypothesis class Finally, we show when $\mathbf{H}$ is rank-deficient, any global minimizer actually corresponds to one single function in $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}$ almost surely. For an arbitrary global minimizer, we assume $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^* + \boldsymbol{Z}$, where $\mathbf{h}\in\mathsf{null}\left(\mathbf{H}\right), \boldsymbol{Z}\in\mathsf{Im}\left(\mathbf{H}^{\otimes 2}\right)$. Suppose we have a prompt $\mathbf{X},\mathbf{y},\mathbf{x},y$ which follows the Assumption 3.1 and the token matrix is formed by (3.1), we have

$$
\begin{aligned}
f_{\boldsymbol{\beta},\boldsymbol{\Gamma}}\left(\mathbf{E}\right) &= \left\langle \boldsymbol{\beta} - \frac{\boldsymbol{\Gamma}}{M}\mathbf{X}^\top\left(\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\right),\mathbf{x}\right\rangle \\
&= \left\langle \boldsymbol{\beta}^* - \frac{\boldsymbol{\Gamma}^*}{M}\mathbf{X}^\top\left(\mathbf{X}\boldsymbol{\beta}^*-\mathbf{y}\right),\mathbf{x}\right\rangle \\
&\qquad + \mathbf{h}^\top\mathbf{x} - \frac{1}{M}\mathbf{x}^\top\boldsymbol{Z}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}^* - \frac{1}{M}\mathbf{x}^\top\boldsymbol{Z}\mathbf{X}^\top\mathbf{X}\mathbf{h} - \frac{1}{M}\mathbf{x}^\top\boldsymbol{\Gamma}^*\mathbf{X}^\top\mathbf{X}\mathbf{h} + \mathbf{x}^\top\boldsymbol{Z}\mathbf{X}^\top\mathbf{y}.
\end{aligned}
$$

Notice that $\mathbf{X}[i],\mathbf{x}\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(\mathbf{0}_d,\mathbf{H}\right)$, we have

$$
\mathbb{E}\left(\mathbf{h}^\top\mathbf{x}\right)^2 = \mathbf{h}^\top\mathbf{H}\mathbf{h} = 0
$$

since $\mathbf{h}\in\mathsf{null}\left(\mathbf{H}\right)$. Therefore, we know $\mathbf{h}^\top\mathbf{x} = 0$ almost surely, and similarly, $\mathbf{X}\mathbf{h} = \mathbf{0}_d$ almost surely. Finally, we have

$$
\mathbb{E}\left[\left(\frac{1}{M}\mathbf{x}^\top\boldsymbol{Z}\mathbf{X}^\top\right)\cdot\left(\frac{1}{M}\mathbf{x}^\top\boldsymbol{Z}\mathbf{X}^\top\right)^\top\right] = \frac{1}{M}\mathbb{E}\operatorname{tr}\left[\mathbf{H}\boldsymbol{Z}\mathbf{H}\right] = 0,
$$

which imples $\mathbf{x}^\top\boldsymbol{Z}\mathbf{X}^\top = \mathbf{0}_d$ almost surely. Therefore, we conclude

$$
f_{\boldsymbol{\beta},\boldsymbol{\Gamma}}\left(\mathbf{E}\right) = f_{\boldsymbol{\beta}^*,\boldsymbol{\Gamma}^*}\left(\mathbf{E}\right) \quad \text{almost surely.}
$$

□

D Proof of Theorem 5.3

Proof. Recall the definition of Linear Transformer Block class:

$$
\begin{aligned}
f_{\mathsf{LTB}} &: \mathbb{R}^{(d+1)\times(M+1)} \to \mathbb{R} \\
\mathbf{E} &\mapsto \left[\mathbf{W}_2^\top \mathbf{W}_1 \left(\mathbf{E} + \mathbf{W}_P^\top \mathbf{W}_V \mathbf{E}\mathbf{M}\frac{\mathbf{E}^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{E}}{M}\right)\right]_{-1,-1},
\end{aligned}
$$

where $\mathbf{E}$ is the token matrix defined by (3.1). Let's take an arbitrary function f in the LTB class with trainable matrices

$$
\mathbf{W}_K,\ \mathbf{W}_Q \in \mathbb{R}^{d_k\times(d+1)}, \quad \mathbf{W}_P,\ \mathbf{W}_V \in \mathbb{R}^{d_v\times(d+1)}, \quad \mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_f\times(d+1)}.
$$

Similar to the proof of the Theorem 4.1, we know that only the last row of $\mathbf{W}_2^\top \mathbf{W}_1$ and the first d-columns of $\mathbf{W}_K^\top \mathbf{W}_Q$ attend the prediction. Therefore, we denote

$$
\mathbf{W}_2^\top \mathbf{W}_1 = \begin{pmatrix} * & * \\ \boldsymbol{\gamma}^\top & * \end{pmatrix}, \quad \mathbf{W}_2^\top \mathbf{W}_1\, \mathbf{W}_P^\top \mathbf{W}_V = \begin{pmatrix} * & * \\ \mathbf{v}_{21}^\top & v_{-1} \end{pmatrix}, \quad \mathbf{W}_K^\top \mathbf{W}_Q = \begin{pmatrix} \mathbf{V}_{11} & * \\ \mathbf{v}_{12}^\top & * \end{pmatrix}, \tag{D.1}
$$

where $\boldsymbol{\gamma}, \mathbf{v}_{12}, \mathbf{v}_{21}, \in \mathbb{R}^d, v_{-1} \in \mathbb{R}, \mathbf{V}_{11} \in \mathbb{R}^{d\times d}$, and $*$ denotes entries that do not enter the prediction. Then, the prediction of LTB function can be written as

$$
\begin{aligned}
f(\mathbf{E}) &= \boldsymbol{\gamma}^\top \mathbf{x} + \begin{pmatrix} \mathbf{v}_{21}^\top & v_{-1} \end{pmatrix} \cdot \frac{\mathbf{E}\mathbf{M}_M \mathbf{E}^\top}{M} \cdot \begin{pmatrix} \mathbf{V}_{11} \\ \mathbf{v}_{12}^\top \end{pmatrix} \cdot \mathbf{x} \\
&= \left[\boldsymbol{\gamma}^\top + \mathbf{v}_{21}^\top \cdot \frac{1}{M}\mathbf{X}^\top \mathbf{X} \cdot \mathbf{V}_{11} + \mathbf{v}_{21}^\top \cdot \frac{1}{M}\mathbf{X}^\top \mathbf{y} \cdot \mathbf{v}_{12}^\top + v_{-1} \cdot \frac{1}{M}\mathbf{y}^\top \mathbf{X} \cdot \mathbf{V}_{11} + v_{-1} \cdot \frac{1}{M}\mathbf{y}^\top \mathbf{y} \cdot \mathbf{v}_{12}^\top\right] \cdot \mathbf{x}.
\end{aligned}
$$

Step 1: simplify the risk function. We use $\widetilde{\boldsymbol{\beta}}$ to denote the task parameter. From the Assumption 3.1 and Definition A.1, we have

$$
\mathbf{y} = \mathbf{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, \quad y = \left\langle \widetilde{\boldsymbol{\beta}}, \mathbf{x} \right\rangle + \varepsilon, \quad \widetilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}^*, \boldsymbol{\Psi}\right), \quad \widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}};
$$

and

$$
\mathbf{X}[i], \mathbf{x} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{H}\right), \quad \boldsymbol{\varepsilon}[i], \varepsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \quad \widetilde{\boldsymbol{\theta}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right).
$$

Then the model output can be written as

$$
\begin{aligned}
f(\mathbf{E}) &= \left[\boldsymbol{\gamma}^\top + \mathbf{v}_{21}^\top \cdot \frac{1}{M}\mathbf{X}^\top \mathbf{X} \cdot \mathbf{V}_{11} + \mathbf{v}_{21}^\top \cdot \frac{1}{M}\mathbf{X}^\top \mathbf{y} \cdot \mathbf{v}_{12}^\top + v_{-1} \cdot \frac{1}{M}\mathbf{y}^\top \mathbf{X} \cdot \mathbf{V}_{11} + v_{-1} \cdot \frac{1}{M}\mathbf{y}^\top \mathbf{y} \cdot \mathbf{v}_{12}^\top\right] \cdot \mathbf{x} \\
&= \left[\boldsymbol{\gamma}^\top + \left(\mathbf{v}_{21} + v_{-1}\widetilde{\boldsymbol{\beta}}\right)^\top \cdot \frac{1}{M}\mathbf{X}^\top \mathbf{X} \cdot \left(\mathbf{V}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{v}_{12}^\top\right)\right] \cdot \mathbf{x} \\
&\quad + \left[\mathbf{v}_{21}^\top \cdot \frac{1}{M}\mathbf{X}^\top \boldsymbol{\varepsilon} \cdot \mathbf{v}_{12}^\top + v_{-1} \cdot \frac{1}{M}\boldsymbol{\varepsilon}^\top \mathbf{X} \cdot \mathbf{V}_{11} + v_{-1} \cdot \frac{2}{M}\boldsymbol{\varepsilon}^\top \mathbf{X}\widetilde{\boldsymbol{\beta}}\mathbf{v}_{12}^\top + \frac{1}{M}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \cdot v_{-1}\mathbf{v}_{12}^\top\right] \cdot \mathbf{x}.
\end{aligned}
$$

To simplify the presentation, we denote

$$
\begin{aligned}
\boldsymbol{z}_1^\top &= \left(\mathbf{v}_{21} + v_{-1}\widetilde{\boldsymbol{\beta}}\right)^\top \cdot \frac{1}{M}\mathbf{X}^\top \mathbf{X} \cdot \left(\mathbf{V}_{11} + \widetilde{\boldsymbol{\beta}}\mathbf{v}_{12}^\top\right), \\
\boldsymbol{z}_2^\top &= \mathbf{v}_{21}^\top \cdot \frac{1}{M}\mathbf{X}^\top \boldsymbol{\varepsilon} \cdot \mathbf{v}_{12}^\top + v_{-1} \cdot \frac{1}{M}\boldsymbol{\varepsilon}^\top \mathbf{X} \cdot \mathbf{V}_{11} + v_{-1} \cdot \frac{2}{M}\boldsymbol{\varepsilon}^\top \mathbf{X}\widetilde{\boldsymbol{\beta}}\mathbf{v}_{12}^\top \\
\boldsymbol{z}_3^\top &= \frac{1}{M}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \cdot v_{-1}\mathbf{v}_{12}^\top.
\end{aligned}
$$

Since $\mathbf{x}, \mathbf{X}, \boldsymbol{\varepsilon}, \widetilde{\boldsymbol{\beta}}$ are independent, we have

$$
\mathcal{R}\left(f\right) = \mathbb{E}\left(f(\mathbf{E}) - \left\langle \widetilde{\boldsymbol{\beta}}, \mathbf{x} \right\rangle - \varepsilon\right)^2
$$

$$
\begin{aligned}
&= \sigma^2 + \mathbb{E}\left(f(\mathbf{E}) - \left\langle \widetilde{\boldsymbol{\beta}}, \mathbf{x} \right\rangle\right)^2 \qquad (\varepsilon \text{ is independent from other variables and zero-mean}) \\
&= \mathbb{E}\left[\left\langle \boldsymbol{z}_1 + \boldsymbol{z}_2 + \boldsymbol{z}_3 + \boldsymbol{\gamma} - \widetilde{\boldsymbol{\beta}}, \mathbf{x} \right\rangle^2\right] + \sigma^2 \\
&= \left\langle \mathbf{H}, \mathbb{E}\left(\boldsymbol{z}_1 + \boldsymbol{z}_2 + \boldsymbol{z}_3 + \boldsymbol{\gamma} - \widetilde{\boldsymbol{\beta}}\right)\left(\boldsymbol{z}_1 + \boldsymbol{z}_2 + \boldsymbol{z}_3 + \boldsymbol{\gamma} - \widetilde{\boldsymbol{\beta}}\right)^\top \right\rangle + \sigma^2.
\end{aligned}
$$

Note that $\boldsymbol{z}_1$ does not contain $\boldsymbol{\varepsilon}$, $\boldsymbol{z}_2$ is a linear form of $\boldsymbol{\varepsilon}$, and $\boldsymbol{z}_3$ is a quadratic form of $\boldsymbol{\varepsilon}$. Using $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$, we have $\mathbb{E}[(\boldsymbol{z}_1 + \boldsymbol{\gamma} - \widetilde{\boldsymbol{\beta}}) \cdot \boldsymbol{z}_2^\top] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{z}_2 \boldsymbol{z}_3^\top] = \mathbf{0}$. Therefore, we have

$$
\begin{aligned}
\mathcal{R}(f) - \sigma^2 = {} & \underbrace{\left\langle \mathbf{H}, \mathbb{E}\left(\boldsymbol{z}_1 + \boldsymbol{\gamma} - \widetilde{\boldsymbol{\beta}}\right)\left(\boldsymbol{z}_1 + \boldsymbol{\gamma} - \widetilde{\boldsymbol{\beta}}\right)^\top \right\rangle}_{S_1'} + \underbrace{\left\langle \mathbf{H}, \mathbb{E}\boldsymbol{z}_2\boldsymbol{z}_2^\top \right\rangle}_{S_2'} + \underbrace{\left\langle \mathbf{H}, \mathbb{E}\boldsymbol{z}_3\boldsymbol{z}_3^\top \right\rangle}_{S_3'} \\
& + 2\underbrace{\left\langle \mathbf{H}, \mathbb{E}\left(\boldsymbol{z}_1 + \boldsymbol{\gamma} - \widetilde{\boldsymbol{\beta}}\right)\boldsymbol{z}_3^\top \right\rangle}_{S_4'}. \qquad \text{(D.2)}
\end{aligned}
$$

Step 2: compute the risk function. Let's compute the risk function. Note that, the risk function is in the same form as the risk function of LSA layer, which is in the proof of Theorem 4.1 (see Section B and equation (B.2)). There are two differences: one is we replace $\mathbf{U}_{11}, \mathbf{u}_{12}, \mathbf{u}_{21}, u_{-1}$ here with $\mathbf{V}_{11}, \mathbf{v}_{12}, \mathbf{v}_{21}, v_{-1}$. The second difference is that we replace $\widetilde{\boldsymbol{\beta}}$ in (B.2) with $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\gamma}$. This is equivalent to replacing $\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^* - \boldsymbol{\gamma}$, since $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\gamma} \sim \mathcal{N}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma}, \boldsymbol{\Psi}\right)$. Therefore, similar to (B.15), the risk function in (D.2) can be written as

$$
\begin{aligned}
& \mathcal{R}(f) - \sigma^2 \\
= {} & \boldsymbol{b}^\top \left(\frac{M+1}{M}\mathbf{H}\boldsymbol{\Psi}\mathbf{H} + \frac{1}{M}\mathrm{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H}\right)\boldsymbol{b} \cdot \mathbf{v}_{12}^\top \mathbf{H}\mathbf{v}_{12} + \frac{\sigma^2}{M} \cdot \boldsymbol{b}^\top \mathbf{H}\boldsymbol{b} \cdot \mathbf{v}_{12}^\top \mathbf{H}\mathbf{v}_{12} \\
& \underbrace{\begin{aligned} & + v_{-1}^2 \mathrm{tr}\left(\boldsymbol{A}^\top \left(\frac{M+1}{M}\mathbf{H}\boldsymbol{\Psi}\mathbf{H} + \frac{1}{M}\mathrm{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H}\right)\boldsymbol{A}\mathbf{H}\right) + \frac{\sigma^2}{M} \cdot \mathrm{tr}\left(\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{A}^\top\right) \\ & + 2v_{-1}\boldsymbol{b}^\top \left[\frac{1}{M}\mathrm{tr}(\mathbf{H}\boldsymbol{\Psi})\mathbf{H} + \frac{M+1}{M}\mathbf{H}\boldsymbol{\Psi}\mathbf{H} + \frac{\sigma^2}{M} \cdot \mathbf{H}\right]\boldsymbol{A}\mathbf{H}\mathbf{v}_{12} - 2v_{-1}\mathrm{tr}\left(\mathbf{H}\boldsymbol{A}\mathbf{H}\boldsymbol{\Psi}\right) - 2\mathbf{v}_{12}^\top \mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{b} \end{aligned}}_{\text{V}} \\
& + \underbrace{\frac{1}{M}\left[\boldsymbol{b}^\top \mathbf{H}\boldsymbol{b} \cdot \mathrm{tr}\left(\boldsymbol{A}^\top \mathbf{H}\boldsymbol{A}\mathbf{H}\right) + 4v_{-1} \cdot \boldsymbol{b}^\top \mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{v}_{12} + 4v_{-1}^2 \cdot \mathrm{tr}\left(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}\right) \cdot \mathbf{v}_{12}^\top \mathbf{H}\mathbf{v}_{12}\right]}_{\text{VI}} \\
& + \underbrace{\boldsymbol{b}^\top \left(\frac{M+1}{M}\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{A}^\top \mathbf{H}\right)\boldsymbol{b} - 2\boldsymbol{b}^\top \mathbf{H}\boldsymbol{A}\mathbf{H}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma}\right) + 2v_{-1}\mathrm{tr}(\mathbf{H}\boldsymbol{\Psi}) \cdot \boldsymbol{b}^\top \mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{v}_{12} + 2\sigma^2 \boldsymbol{b}^\top \mathbf{H}\boldsymbol{A}\mathbf{H}\left(v_{-1}\mathbf{v}_{12}\right)}_{\text{VII}} \\
& + \underbrace{\begin{aligned} & \left(\boldsymbol{\beta}^* - \boldsymbol{\gamma}\right)^\top \mathbf{H}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma}\right) - 2\left(\mathrm{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right) \cdot \left(\boldsymbol{\beta}^* - \boldsymbol{\gamma}\right)^\top \mathbf{H}\left(v_{-1}\mathbf{v}_{12}\right) + \mathrm{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) \\ & + v_{-1}^2 \cdot \left[\left(2\mathrm{tr}\left(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}\right) + \frac{M+2}{M}\mathrm{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right)^2\right) \cdot \mathbf{v}_{12}^\top \mathbf{H}\mathbf{v}_{12} + \left(2 + \frac{4}{M}\right)\sigma^2 \mathrm{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \frac{M+2}{M}\sigma^4\right] \cdot \mathbf{v}_{12}^\top \mathbf{H}\mathbf{v}_{12} \end{aligned}}_{\text{VIII}} \\
= {} & \text{V} + \text{VI} + \text{VII} + \text{VIII}, \qquad \text{(D.3)}
\end{aligned}
$$

where V, VI, VII, VIII are defined as above and $\boldsymbol{\Omega} = \frac{M+1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} + \frac{\mathrm{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2}{M} \cdot \mathbf{I}_d$, and $\boldsymbol{b} := \mathbf{v}_{21} + v_{-1}\boldsymbol{\beta}^* \in \mathbb{R}^d$, $\boldsymbol{A} := \mathbf{V}_{11} + \boldsymbol{\beta}^* \mathbf{v}_{12}^\top \in \mathbb{R}^{d \times d}$,

Step 3: solve the global minimum of the risk function. This is very similar to step 5 in Appendix B. The V term in (D.3) is actually equal to the I term in (B.15), so we have

$$\begin{aligned}
\mathrm{V} = \operatorname{tr}&\left[\left(\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}} + v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)\boldsymbol{\Omega}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}} + v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)^\top\right] \\
&\quad - \operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right) \\
\geq& -\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right) \\
=& -\operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\boldsymbol{\Omega}^{-1}\right),
\end{aligned}$$

where the last line comes from the fact that $\boldsymbol{\Omega}$ and $\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}$ commute. For the same reason, we know that the VI term above is equal to the II term in (B.15), which implies VI ≥ 0, since

$$\begin{aligned}
4v_{-1}\cdot\boldsymbol{b}^\top\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{v}_{12} &\geq -4\left\|\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\right\|_F\cdot\left\|v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right\|_F\cdot\left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}}\right\|_F\cdot\left\|\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\right\|_F \\
&\geq -\left\|\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\right\|_F^2\left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}}\right\|_F^2 - 4\left\|v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right\|_F^2\left\|\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\right\|_F^2 \\
&= -\boldsymbol{b}^\top\mathbf{H}\boldsymbol{b}\cdot\operatorname{tr}\left(\boldsymbol{A}^\top\mathbf{H}\boldsymbol{A}\mathbf{H}\right) - 4v_{-1}^2\cdot\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}\right)\cdot\mathbf{v}_{12}^\top\mathbf{H}\mathbf{v}_{12}, \quad \text{(D.4)}
\end{aligned}$$

where the last line comes from the fact that $\|\boldsymbol{A}\|_F^2 = \operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^\top\right)$ for any matrix $\boldsymbol{A}$. The term VII above is equal to III term in (B.15), except that we replace $\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^* - \boldsymbol{\gamma}$. Therefore, we have

$$\begin{aligned}
&\mathrm{VII} + \frac{M}{M+1}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)v_{-1}\mathbf{v}_{12}\right)^\top\mathbf{H}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)v_{-1}\mathbf{v}_{12}\right) \\
=& \left[\boldsymbol{A}^\top\mathbf{H}\boldsymbol{b} - \frac{M}{M+1}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)v_{-1}\mathbf{v}_{12}\right)\right]^\top\left(\frac{M+1}{M}\mathbf{H}\right) \\
&\quad\cdot\left[\boldsymbol{A}^\top\mathbf{H}\boldsymbol{b} - \frac{M}{M+1}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)v_{-1}\mathbf{v}_{12}\right)\right] \\
\geq& 0, \qquad \text{(D.5)}
\end{aligned}$$

which implies

$$\mathrm{VII} \geq -\frac{M}{M+1}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)v_{-1}\mathbf{v}_{12}\right)^\top\mathbf{H}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)v_{-1}\mathbf{v}_{12}\right).$$

Combining the three parts above, one has

$$\begin{aligned}
&\mathcal{R}\left(f\right) - \sigma^2 \\
&\geq \mathrm{VIII} - \operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\boldsymbol{\Omega}^{-1}\right) \\
&\quad -\frac{M}{M+1}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)v_{-1}\mathbf{v}_{12}\right)^\top\mathbf{H}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)v_{-1}\mathbf{v}_{12}\right) \\
&= \underbrace{\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) - \operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2\boldsymbol{\Omega}^{-1}\right)}_{\inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}}\mathcal{R}(f)-\sigma^2} + \frac{1}{M+1}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma}\right)^\top\mathbf{H}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma}\right) \\
&\qquad - \frac{2\left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)}{M+1}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma}\right)^\top\mathbf{H}\left(v_{-1}\mathbf{v}_{12}\right) \\
&\qquad + \left[2\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}\right) + \frac{3M+2}{M(M+1)}\left(\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Psi}\right) + \sigma^2\right)^2\right]\left(v_{-1}\mathbf{v}_{12}\right)^\top\mathbf{H}\left(v_{-1}\mathbf{v}_{12}\right).
\end{aligned}$$

Here, the global minimum of GD-$\boldsymbol{\beta}$ class is taken from Theorem 5.2, whose proof does not reply on the proof here. Therefore, one has

$$\mathcal{R}(f) - \inf_{f \in \mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) \geq \frac{1}{M+1} (\boldsymbol{\beta}^* - \boldsymbol{\gamma})^\top \mathbf{H} (\boldsymbol{\beta}^* - \boldsymbol{\gamma}) - \frac{2\left(\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2\right)}{M+1} (\boldsymbol{\beta}^* - \boldsymbol{\gamma})^\top \mathbf{H} (v_{-1}\mathbf{v}_{12}) + \left[2\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}) + \frac{3M+2}{M(M+1)} \left(\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2\right)^2\right] (v_{-1}\mathbf{v}_{12})^\top \mathbf{H} (v_{-1}\mathbf{v}_{12}).$$

The right hand side in the above inequality is a quadratic function of $v_{-1}\mathbf{v}_{12}$, so we can take its global minimizer:

$$v_{-1}\mathbf{v}_{12} = \frac{\frac{\left(\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2\right)}{M+1}}{2\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}) + \frac{3M+2}{M(M+1)} (\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2)^2} (\boldsymbol{\beta}^* - \boldsymbol{\gamma})$$

to lower bound the risk gap as

$$\mathcal{R}(f) - \inf_{f \in \mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) \geq \left[\frac{1}{M+1} - \frac{\frac{\left(\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2\right)^2}{(M+1)^2}}{2\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}) + \frac{3M+2}{M(M+1)} (\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2)^2}\right] \|\boldsymbol{\beta}^* - \boldsymbol{\gamma}\|_{\mathbf{H}}^2 \geq 0.$$

Note that, taking infimum on the left hand side, we get

$$\inf_{f \in \mathcal{F}_{\mathsf{LTB}}} \mathcal{R}(f) - \inf_{f \in \mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) \geq 0.$$

On the other hand, in the main text we have showed that $\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}} \subset \mathcal{F}_{\mathsf{LTB}}$, which implies

$$\inf_{f \in \mathcal{F}_{\mathsf{LTB}}} \mathcal{R}(f) - \inf_{f \in \mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) \leq 0.$$

Therefore, we conclude

$$\inf_{f \in \mathcal{F}_{\mathsf{LTB}}} \mathcal{R}(f) = \inf_{f \in \mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) = \operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) - \operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2 \boldsymbol{\Omega}^{-1}\right),$$

where the last equation is from Theorem 5.2.

Step 4: sufficient and necessary conditions for global minimizers. Let's now verify the conditions for the global minimizers. For a function in LTB class, the sufficient and necessary condition for it to be a global minimizer is that inequalities in the above step all hold, which are

$$\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}} + v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}, \tag{D.6}$$

$$\left\|\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\right\|_F^2 \left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}}\right\|_F^2 = 4\left\|v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right\|_F^2 \left\|\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\right\|_F^2, \tag{D.7}$$

$$v_{-1} \cdot \boldsymbol{b}^\top\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{A}\mathbf{H}\mathbf{v}_{12} = \left\|\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\right\|_F \cdot \left\|v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right\|_F \cdot \left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}}\right\|_F \cdot \left\|\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\right\|_F, \tag{D.8}$$

$$\mathbf{H}^{\frac{1}{2}}\left[\boldsymbol{A}^\top\mathbf{H}\boldsymbol{b} - \frac{M}{M+1}\left(\boldsymbol{\beta}^* - \boldsymbol{\gamma} - \left(\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2\right) v_{-1}\mathbf{v}_{12}\right)\right] = \mathbf{0}_{d\times d}, \tag{D.9}$$

$$v_{-1}\mathbf{v}_{12} = \frac{\frac{\left(\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2\right)}{M+1}}{2\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}\mathbf{H}\boldsymbol{\Psi}) + \frac{3M+2}{M(M+1)} (\operatorname{tr}(\mathbf{H}\boldsymbol{\Psi}) + \sigma^2)^2} (\boldsymbol{\beta}^* - \boldsymbol{\gamma}) \tag{D.10}$$

$$\|\boldsymbol{\beta}^* - \boldsymbol{\gamma}\|_{\mathbf{H}} = 0. \tag{D.11}$$

Let's first verify the necessary conditions of the system defined above. From (D.11) and Lemma H.5, we know $\boldsymbol{\gamma} \in \boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}^{\frac{1}{2}}\right) = \boldsymbol{\beta}^* + \mathsf{null}(\mathbf{H})$. Then, we have $v_{-1}\mathbf{v}_{12} \in \mathsf{null}(\mathbf{H})$. Then, in (D.7), we know

$$\left\|\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}\right\|_F^2 \left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}}\right\|_F^2 = 4\left\|v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right\|_F^2 \left\|\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\right\|_F^2 = 4\left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right\|_F^2 \left\|v_{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\right\|_F^2 = 0,$$

which implies either $\mathbf{H}^{\frac{1}{2}}\boldsymbol{b} = \mathbf{0}_d$ or $\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}} = \mathbf{0}_{d\times d}$. If we assume $\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}\mathbf{H}^{\frac{1}{2}} = \mathbf{0}_{d\times d}$, then (D.6) implies $\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}$. From our assumption, we know $\mathsf{rank}(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}^{\frac{1}{2}}) \geq 2$, which implies $\mathsf{rank}(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}) \geq 2$, since for any matrix $\boldsymbol{Z}$, it holds that $\mathsf{rank}(\boldsymbol{Z}) = \mathsf{rank}(\boldsymbol{Z}\boldsymbol{Z}^\top)$. Since multiplication by an invertible matrix does not change the rank, we know $\mathsf{rank}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right) \geq 2$, while $\mathbf{H}^{\frac{1}{2}}\mathbf{v}_{12}\boldsymbol{b}^\top\mathbf{H}^{\frac{1}{2}}$ is a matrix of rank at most one, which contradicts with (D.6). Therefore, we have

$$\mathbf{H}^{\frac{1}{2}}\boldsymbol{b} = \mathbf{0}, \tag{D.12}$$

$$v_{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}. \tag{D.13}$$

Recall in Theorem 5.2, we have defined

$$\boldsymbol{\Gamma}^* := \boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}.$$

Simple calculation shows

$$\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} = \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} = \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1},$$

where the second and the last equalities come from the fact that $\boldsymbol{\Omega}$ and $\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}$ commute, and the third equality comes from the property of Moor Penrose pseudo-inverse. Therefore, one solution of (D.13) is $v_{-1}\boldsymbol{A} = \boldsymbol{\Gamma}^{*\top}$. Since (D.13) is a linear equation to $v_{-1}\boldsymbol{A}$, we know its full solution is

$$v_{-1}\boldsymbol{A} \in \boldsymbol{\Gamma}^{*\top} + \left\{\boldsymbol{Z} : \mathbf{H}^{\frac{1}{2}}\boldsymbol{Z}\mathbf{H}^{\frac{1}{2}} = \mathbf{0}_{d\times d}\right\} = \boldsymbol{\Gamma}^{*\top} + \mathsf{Im}\left(\mathbf{H}^{\otimes 2}\right).$$

The solution to (D.12) is

$$\mathbf{v}_{21} = -v_{-1}\boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}^{\frac{1}{2}}\right) = -v_{-1}\boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}\right).$$

Therefore, we know the necessary conditions for (D.6) to (D.11) are

$$\begin{cases} v_{-1} \neq 0, \\ v_{-1}\mathbf{v}_{12} \in \mathsf{null}\left(\mathbf{H}\right), \\ \mathbf{v}_{21} = -v_{-1}\boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}\right), \\ v_{-1}\mathbf{V}_{11} \in \boldsymbol{\Gamma}^{*\top} - v_{-1}\boldsymbol{\beta}^*\mathbf{v}_{12}^\top + \mathsf{Im}\left(\mathbf{H}^{\otimes 2}\right), \\ \boldsymbol{\gamma} = \boldsymbol{\beta}^* + \mathsf{null}\left(\mathbf{H}\right) \end{cases} \tag{D.14}$$

It is easy to verify these equations above are also sufficient conditions of (D.6) to (D.11) by directly replacing each variable with its value and validating equaions from (D.6) to (D.11).

Specially, if $\mathbf{H}$ is positive definite and hence, invertible, the global minimizer is unique up to a scaling to v_{-1} :

$$v_{-1} \neq 0, \quad \mathbf{v}_{12} = \mathbf{0}_d, \quad \mathbf{v}_{21} = -v_{-1}\boldsymbol{\beta}^*, \quad \mathbf{V}_{11} = \frac{1}{v_{-1}} \cdot \boldsymbol{\Gamma}^{*\top}, \quad \boldsymbol{\gamma} = \boldsymbol{\beta}^*. \tag{D.15}$$

Step 5: equivalence in the hypothesis class. Finally, we will prove that any global minimizer of $\mathcal{F}_{\mathsf{LTB}}$ is actually equivalent to one single function in $\mathcal{F}_{\mathsf{LTB}}$ almost surely. More concretely, let's take a function $f \in \mathcal{F}_{\mathsf{LTB}}$ with parameters $\mathbf{W}_K,\ \mathbf{W}_Q,\ \mathbf{W}_P,\ \mathbf{W}_V, \mathbf{W}_1,\ \mathbf{W}_2$ and recall the parameter transformation in (D.1). We assume equations in (D.14) and Assumption 3.1 hold. Then, for any vector $\mathbf{a} \in \mathsf{null}\left(\mathbf{H}\right)$, one has $\mathbf{x}^\top\mathbf{a} = \mathbf{x}_i^\top\mathbf{a} = 0$ since $\mathbf{x}, \mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}_d, \mathbf{H}\right)$. We denote

$$\mathbf{v}_{21} = -v_{-1}\boldsymbol{\beta}^* + \mathbf{a}_1, \quad v_{-1}\mathbf{V}_{11} = \boldsymbol{\Gamma}^{*\top} - v_{-1}\boldsymbol{\beta}^*\mathbf{v}_{12}^\top + \boldsymbol{Z}, \quad \boldsymbol{\gamma} = \boldsymbol{\beta}^* + \mathbf{a}_2,$$

where $\mathbf{a}_1, \mathbf{a}_2 \in \mathbf{H}, \mathbf{H}^{\frac{1}{2}}\boldsymbol{Z}\mathbf{H}^{\frac{1}{2}} = \mathbf{0}_{d\times d}$. Then, we have

$$f(\mathbf{E}) = \left[\mathbf{W}_2^\top\mathbf{W}_1\left(\mathbf{E} + \mathbf{W}_P^\top\,\mathbf{W}_V\mathbf{E}\mathbf{M}\frac{\mathbf{E}^\top\mathbf{W}_K^\top\,\mathbf{W}_Q\mathbf{E}}{M}\right)\right]_{-1,-1}$$

$$
\begin{aligned}
&= \left(\boldsymbol{\beta}^* + \mathbf{a}_2\right)^\top \mathbf{x} + \begin{pmatrix} -v_{-1}\boldsymbol{\beta}^{*\top} + \mathbf{a}_1^\top & v_{-1} \end{pmatrix} \cdot \frac{1}{M} \begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{y} \\ \mathbf{y}^\top\mathbf{X} & \mathbf{y}^\top\mathbf{y} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{V}_{11} \\ \mathbf{v}_{12}^\top \end{pmatrix} \cdot \mathbf{x} \\
&= \boldsymbol{\beta}^{*\top}\mathbf{x} + \begin{pmatrix} -\boldsymbol{\beta}^{*\top} & 1 \end{pmatrix} \cdot \frac{1}{M} \begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{y} \\ \mathbf{y}^\top\mathbf{X} & \mathbf{y}^\top\mathbf{y} \end{pmatrix} \cdot \begin{pmatrix} v_{-1}\mathbf{V}_{11}\mathbf{x} \\ v_{-1}\mathbf{v}_{12}^\top\mathbf{x} \end{pmatrix} \qquad (\mathbf{X}\mathbf{a}_1 = \mathbf{0}, \mathbf{a}_2^\top\mathbf{x} = 0) \\
&= \boldsymbol{\beta}^{*\top}\mathbf{x} + \begin{pmatrix} -\boldsymbol{\beta}^{*\top} & 1 \end{pmatrix} \cdot \frac{1}{M} \begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{y} \\ \mathbf{y}^\top\mathbf{X} & \mathbf{y}^\top\mathbf{y} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\Gamma}^{*\top}\mathbf{x} \\ \mathbf{0}_d^\top \end{pmatrix} \qquad (\mathbf{v}_{12}^\top\mathbf{x} = 0, \mathbf{X}\mathbf{V}_{11}\mathbf{x} = \mathbf{0}) \\
&= \left\langle \boldsymbol{\beta}^* - \frac{\boldsymbol{\Gamma}^*}{M}\mathbf{X}^\top\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\right), \mathbf{x} \right\rangle = f_{\boldsymbol{\beta}^*,\boldsymbol{\Gamma}^*}\left(\mathbf{E}\right),
\end{aligned}
$$

where $f_{\boldsymbol{\beta}^*,\boldsymbol{\Gamma}^*}\left(\cdot\right)$ is the GD-$\boldsymbol{\beta}$ function defined in (5.1). □

E Proof of Corollary 6.2

Proof. We denote $\phi_1 \geq \phi_2 \geq ... \geq \phi_d \geq 0$ are ordered eigenvalues of $\mathbf{\Psi}^{\frac{1}{2}}\mathbf{H}\mathbf{\Psi}^{\frac{1}{2}}$. From Theorem 5.2, we have

$$\inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) - \sigma^2 = \operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) - \operatorname{tr}\left(\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right)^2 \mathbf{\Omega}^{-1}\right),$$

where

$$\mathbf{\Omega} := \frac{M+1}{M}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}} + \frac{\operatorname{tr}\left(\mathbf{H}\mathbf{\Psi}\right) + \sigma^2}{M} \cdot \mathbf{I}_d.$$

Therefore, we have

$$\begin{aligned}\inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) - \sigma^2 &= \operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}} \cdot \left(\mathbf{\Omega} - \mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right)\mathbf{\Omega}^{-1}\right) \\ &= \frac{1}{M}\operatorname{tr}\left(\mathbf{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}} \cdot \left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}} + \left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2\right) \cdot \mathbf{I}_d\right)\right).\end{aligned}$$

Since

$$\left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2\right)\cdot\mathbf{I}_d \preceq \mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}} + \left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2\right)\cdot\mathbf{I}_d \preceq 2\left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2\right)\cdot\mathbf{I}_d,$$

we have

$$\begin{aligned}\inf_{f\in\mathcal{F}_{\mathsf{GD}\text{-}\boldsymbol{\beta}}} \mathcal{R}(f) - \sigma^2 &\simeq \frac{\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2}{M}\operatorname{tr}\left(\mathbf{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) \\ &= \bar{\phi} \cdot \sum_{i=1}^{d} \frac{\phi_i}{\frac{M+1}{M}\phi_i + \bar{\phi}} \\ &\simeq \sum_{i=1}^{d} \min\left\{\phi_i, \bar{\phi}\right\}.\end{aligned}$$

Therefore, we conclude. □

F Proof of Theorem 6.1

F.1 Proof of first equation

Proof. From the classical bias-variance decomposition, we know

$$\begin{aligned}\mathcal{L}\left(g; \mathbf{X}\right) &:= \mathbb{E}\left[\left(g(\mathbf{X}, \mathbf{y}, \mathbf{x}) - y\right)^2 \mid \mathbf{X}\right] \\ &= \mathbb{E}\left[\left(g(\mathbf{X}, \mathbf{y}, \mathbf{x}) - \mathbb{E}\left[y \mid \mathbf{X}, \mathbf{y}, \mathbf{x}\right]\right)^2 \mid \mathbf{X}\right] + \mathbb{E}\left[\left(\mathbb{E}\left[y \mid \mathbf{X}, \mathbf{y}, \mathbf{x}\right] - y\right)^2 \mid \mathbf{X}\right]\end{aligned}$$

since the cross term vanishes. The second term does not depend on g and hence, the Bayesian optimal estimator is given by the posterior mean, i.e.,

$$\widehat{y}_{\mathsf{Bayes}} = \mathbb{E}\left[y \mid \mathbf{X}, \mathbf{y}, \mathbf{x}\right] = \left\langle \mathbb{E}\left[\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}, \mathbf{y}, \mathbf{x}\right], \mathbf{x} \right\rangle.$$

Since $\widetilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}^*, \boldsymbol{\Psi}\right)$, we have there exists a random vector $\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}\left(\mathbf{0}_d, \mathbf{I}_d\right)$ such that $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}$ almost surely. Therefore, one has

$$\widehat{y}_{\mathsf{Bayes}} = \left\langle \boldsymbol{\beta}^*, \mathbf{x} \right\rangle + \left\langle \mathbb{E}\left[\widetilde{\boldsymbol{\theta}} \mid \mathbf{X}, \mathbf{y}, \mathbf{x}\right], \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{x} \right\rangle.$$

To compute $\mathbb{E}\left[\widetilde{\boldsymbol{\theta}} \mid \mathbf{X}, \mathbf{y}, \mathbf{x}\right]$, it suffices to solve the posterior distribution of $\boldsymbol{\beta}$ given $\mathbf{X}, \mathbf{y}$. From $\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}\left(\mathbf{0}_d, \mathbf{I}_d\right)$, we have

$$\begin{aligned}\mathbb{P}\left(\widetilde{\boldsymbol{\theta}} \mid \mathbf{X}, \mathbf{y}, \mathbf{x}\right) \propto \mathbb{P}\left(\widetilde{\boldsymbol{\theta}}\right)\mathbb{P}\left(\mathbf{y} \mid \mathbf{X}, \widetilde{\boldsymbol{\theta}}\right) &\propto \exp\left(-\frac{\left\|\mathbf{y} - \mathbf{X}\left(\boldsymbol{\beta}^* + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)\right\|_2^2}{2\sigma^2} - \frac{1}{2}\widetilde{\boldsymbol{\theta}}^\top \cdot \widetilde{\boldsymbol{\theta}}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\left[\widetilde{\boldsymbol{\theta}}^\top\left(\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Psi}^{\frac{1}{2}} + \sigma^2\mathbf{I}_d\right)\widetilde{\boldsymbol{\theta}} - 2\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\right)^\top\mathbf{X}\boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right]\right).\end{aligned}$$

Note that, the function above matches the probability density function of a multivariate Gaussian distrbution. Therefore, the posterior mean is given by

$$\mathbb{E}\left[\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, \mathbf{x}\right] = \left(\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Psi}^{\frac{1}{2}} + \sigma^2\mathbf{I}_d\right)^{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{X}^\top\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\right).$$

Therefore, we conclude

$$\widehat{y}_{\mathsf{Bayes}} = \mathbf{x}^\top\boldsymbol{\Psi}^{\frac{1}{2}}\left(\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Psi}^{\frac{1}{2}} + \sigma^2\mathbf{I}_d\right)^{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{X}^\top\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\right) + \mathbf{x}^\top\boldsymbol{\beta}^*. \tag{F.1}$$

□

F.2 Proof of second equation

Proof. Let's first do a variable transformation. We denote

$$\widetilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Psi}^{\frac{1}{2}}, \quad \widetilde{\mathbf{x}} = \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{x}, \quad \widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*, \quad \boldsymbol{\Lambda} = \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}. \tag{F.2}$$

Then, we know $\widetilde{\mathbf{X}}[i], \widetilde{\mathbf{x}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}_d, \boldsymbol{\Lambda}\right)$. We can write the Bayesian optimal estimator in (F.1) as

$$\widehat{y}_{\mathsf{Bayes}} = \widetilde{\mathbf{x}}^\top\left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \sigma^2\mathbf{I}_d\right)^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}} + \mathbf{x}^\top\boldsymbol{\beta}^*.$$

The true label is

$$y = \mathbf{x}^\top\widetilde{\boldsymbol{\beta}} + \varepsilon = \mathbf{x}^\top\left(\boldsymbol{\beta}^* + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right).$$

Therefore, we have

$$\mathcal{L}\left(\widehat{y}_{\mathsf{Bayes}}; \mathbf{X}\right) - \sigma^2 = \mathbb{E}\left(\widetilde{\mathbf{x}}^\top\left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \sigma^2\mathbf{I}_d\right)^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}} + \mathbf{x}^\top\boldsymbol{\beta}^* - \mathbf{x}^\top\left(\boldsymbol{\beta}^* + \boldsymbol{\Psi}^{\frac{1}{2}}\widetilde{\boldsymbol{\theta}}\right)\right)^2$$

$$= \mathbb{E}\left(\widetilde{\mathbf{x}}^{\top}\left[\left(\widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}}+\sigma^2\mathbf{I}_d\right)^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}}-\widetilde{\boldsymbol{\theta}}\right]\right)^2 = \mathbb{E}\left\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\right\|_{\boldsymbol{\Lambda}}^2,$$

where

$$\widehat{\boldsymbol{\theta}} := \left(\widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}}+\sigma^2\mathbf{I}_d\right)^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}}.$$

Note that, this is equivalent to the risk of estimating the ground true linear weight $\widetilde{\boldsymbol{\theta}}$ using $\widehat{\boldsymbol{\theta}}$ under the Gaussian prior $\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right)$. The estimator $\widehat{\boldsymbol{\theta}}$ is a function of transformed input-output pairs in the context $\left(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}}\right)$ and the transformed query input $\mathbf{x}$, and takes the form of standard ridge estimator with regularization coefficient being σ^2. We then revoke the standard results for the risk of a ridge estimator. Applying the Theorem 1 and 2 in [33], we know for a fixed weight vector $\widetilde{\boldsymbol{\theta}}$, with probability at least $1-\exp\left(-\Omega(M)\right)$ we have

$$\left\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\right\|_{\mathbf{H}}^2 \simeq \left(\frac{\sigma^2+\sum_{i>k^*}\phi_i}{M}\right)^2\left\|\widetilde{\boldsymbol{\theta}}\right\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \left\|\widetilde{\boldsymbol{\theta}}\right\|_{\mathbf{H}_{k^*:\infty}}^2 + \sigma^2\left(\frac{k^*}{M}+\frac{M\sum_{i>k^*}\phi_i^2}{\sigma^2+\sum_{i>k^*}\phi_i}\right),$$

where $\phi_1 \geq \phi_2 \geq ... \geq \phi_d \geq 0$ are ordered eigenvalues of $\boldsymbol{\Lambda}$.

$$k^* := \min\left\{k : \phi_k \geq c\cdot\frac{\sigma^2+\sum_{i>k^*}\phi_i}{M}\right\}$$

and $c > 1$ is an absolute constant. Here, $\mathbf{H}_{0:k^*}$ is SVD approximation with respect to the largest k^* singular values and $\mathbf{H}_{k^*:\infty}$ is the SVD approximation in the remaining singular values. Namely, if we have the eigen-decomposition of $\mathbf{H} = \mathbf{Q}\cdot\operatorname{diag}\left(\phi_1,\phi_2,...,\phi_d\right)\cdot\mathbf{Q}^{\top}$, where $\mathbf{Q}$ is an orthogonal matrix, then $\mathbf{H}_{0:k^*}$ and $\mathbf{H}_{k^*:\infty}$ are given by

$$\mathbf{H}_{0:k^*} = \mathbf{Q}\cdot\operatorname{diag}\left(\phi_1,\phi_2,...,\phi_{k^*},0,0,...,0\right)\cdot\mathbf{Q}^{\top}, \quad \mathbf{H}_{k^*:\infty} = \mathbf{H}-\mathbf{H}_{0:k^*}.$$

Taking expectation over $\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}\left(0, \mathbf{I}_d\right)$, we have

$$\begin{aligned}\mathcal{L}\left(\widehat{y}_{\mathsf{Bayes}};\mathbf{X}\right)-\sigma^2 &= \mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim\mathcal{N}(0,\mathbf{I}_d)}\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{Bayes}}-\widetilde{\boldsymbol{\theta}}\right\|_{\mathbf{H}}^2\\ &= \left(\frac{\sigma^2+\sum_{i>k^*}\phi_i}{M}\right)^2\cdot\sum_{i\leq k^*}\frac{1}{\phi_i}+\sum_{i>k^*}\phi_i+\sigma^2\left(\frac{k^*}{M}+\frac{M\sum_{i>k^*}\phi_i^2}{\sigma^2+\sum_{i>k^*}\phi_i}\right)\end{aligned}$$

Now we simplify this expression. First, we define

$$\bar{\phi} := c\cdot\frac{\sigma^2+\sum_{i>k^*}\phi_i}{M}.$$

From the definition, we see $\bar{\phi} \geq \frac{c\sigma^2}{M}$. On the other hand, from the assumption that $\operatorname{tr}\left(\mathbf{H}\right) = \operatorname{tr}\left(\boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Psi}^{\frac{1}{2}}\right) = \sum_{i=1}^{d}\phi_i \lesssim \sigma^2$, we know that $\bar{\phi} \lesssim \frac{c\sigma^2}{M}$. Combining two parts, we get

$$\bar{\phi} \simeq \frac{\sigma^2}{M}. \tag{F.3}$$

Therefore, we have

$$\begin{aligned}\mathcal{L}\left(\widehat{y}_{\mathsf{Bayes}};\mathbf{X}\right)-\sigma^2 &\simeq \bar{\phi}^2\cdot\sum_{i\leq k^*}\frac{1}{\phi_i}+\sum_{i>k^*}\phi_i+\frac{\sigma^2}{M}\cdot\left(k^*+\frac{\sum_{i>k^*}\phi_i^2}{\bar{\phi}^2}\right)\\ &\simeq \sum_i\min\left\{\frac{\bar{\phi}^2}{\phi_i},\phi_i\right\}+\bar{\phi}\cdot\sum_i\min\left\{1,\frac{\phi_i^2}{\bar{\phi}^2}\right\} && \text{(from (F.3))}\\ &\simeq \sum_i\left(\min\left\{\frac{\bar{\phi}^2}{\phi_i},\phi_i\right\}+\min\left\{\bar{\phi},\frac{\phi_i^2}{\bar{\phi}}\right\}\right)\\ &\simeq \sum_i\min\left\{\phi_i,\bar{\phi}\right\}.\end{aligned}$$

This finishes the proof. □

G Proof of Theorem 6.3

G.1 Training dynamics

Now we consider doing gradient flow on the risk function, or equivalently, on the excess risk function which differs by only a constant:

$$\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t} = -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\beta}}\left[\mathcal{R}(\boldsymbol{\beta}, \boldsymbol{\Gamma}) - \min \mathcal{R}(\cdot,\cdot)\right]; \tag{G.1}$$

$$\frac{\mathrm{d}\boldsymbol{\Gamma}}{\mathrm{d}t} = -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Gamma}}\left[\mathcal{R}(\boldsymbol{\beta}, \boldsymbol{\Gamma}) - \min \mathcal{R}(\cdot,\cdot)\right]. \tag{G.2}$$

First, we have the following corollary of Theorem 5.2, which computes the excess risk $\mathcal{R}(\boldsymbol{\beta}, \boldsymbol{\Gamma}) - \min \mathcal{R}(\cdot,\cdot)$.

Corollary G.1 (Excess Risk)**.** *We fix M as the context length. Consider the ICL risk in* (3.5)*, assume the data is generated following Assumption 3.1. Then, we have*

$$\begin{aligned}
&\mathcal{R}(\boldsymbol{\beta}, \boldsymbol{\Gamma}) - \min \mathcal{R}(\cdot,\cdot) \\
&= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \left[(\mathbf{I}_d - \boldsymbol{\Gamma}\mathbf{H})^\top \mathbf{H} (\mathbf{I}_d - \boldsymbol{\Gamma}\mathbf{H}) + \frac{\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Gamma}^\top \mathbf{H}\boldsymbol{\Gamma}\right)}{M}\mathbf{H} + \frac{1}{M}\mathbf{H}\boldsymbol{\Gamma}^\top\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}\right](\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\
&+ \operatorname{tr}\left[\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)\boldsymbol{\Omega}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)^\top\right].
\end{aligned} \tag{G.3}$$

Proof. This is obtained directly from equation (C.5) and (C.7). □

Now, we write out the differential equations explicitly in the following lemma.

Lemma G.2 (Dynamical system)**.** *The dynamical system of gradient flow described in* (G.1) *and* (G.2) *is*

$$\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t} = -\left[(\mathbf{I}_d - \boldsymbol{\Gamma}\mathbf{H})^\top \mathbf{H} (\mathbf{I}_d - \boldsymbol{\Gamma}\mathbf{H}) + \frac{\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Gamma}^\top \mathbf{H}\boldsymbol{\Gamma}\right)}{M}\mathbf{H} + \frac{1}{M}\mathbf{H}\boldsymbol{\Gamma}^\top\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}\right](\boldsymbol{\beta} - \boldsymbol{\beta}^*) \tag{G.4}$$

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{\Gamma}}{\mathrm{d}t} &= -\left(\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}\boldsymbol{\Psi}\mathbf{H}\right) - \frac{M+1}{M}\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top\mathbf{H} \\
&\quad - \frac{1}{M}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top\mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\cdot\mathbf{H}\boldsymbol{\Gamma}\mathbf{H} + \mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top\mathbf{H}.
\end{aligned} \tag{G.5}$$

Proof. This can be obtained by directly calculating the derivatives over the excess risk (G.3). To write out the dynamics of $\boldsymbol{\beta}$, it suffices to notice that $\boldsymbol{\beta}$ only attends the first term of the RHS of (G.3), which is a standard quadratic function. For the derivatives of $\boldsymbol{\Gamma}$, we first have

$$\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Gamma}}\operatorname{tr}\left[\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)\boldsymbol{\Omega}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)^\top\right] \\
&= \mathbf{H}^{\frac{1}{2}}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\right)\boldsymbol{\Omega}^{\frac{1}{2}}\cdot\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} \qquad \text{(the sixth equation in Lemma H.3)} \\
&= \mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}} - \mathbf{H}\boldsymbol{\Psi}\mathbf{H}.
\end{aligned}$$

Now it suffices to compute

$$\frac{\partial}{\partial\boldsymbol{\Gamma}}\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top\left[(\mathbf{I}_d - \boldsymbol{\Gamma}\mathbf{H})^\top \mathbf{H} (\mathbf{I}_d - \boldsymbol{\Gamma}\mathbf{H}) + \frac{\operatorname{tr}\left(\mathbf{H}\boldsymbol{\Gamma}^\top \mathbf{H}\boldsymbol{\Gamma}\right)}{M}\mathbf{H} + \frac{1}{M}\mathbf{H}\boldsymbol{\Gamma}^\top\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}\right](\boldsymbol{\beta} - \boldsymbol{\beta}^*).$$

Let's compute the partial derivativess separately. From Lemma H.3, we have

$$\frac{\partial}{\partial \mathbf{\Gamma}} \frac{1}{2} \left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \left(-\mathbf{H}\mathbf{\Gamma}^\top \mathbf{H}\right) \left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right) = \frac{\partial}{\partial \mathbf{\Gamma}} \frac{1}{2} \left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \left(-\mathbf{H}\mathbf{\Gamma}\mathbf{H}\right) \left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)$$
$$= -\frac{1}{2}\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \mathbf{H};$$
$$\frac{\partial}{\partial \mathbf{\Gamma}} \frac{1}{2} \left(\frac{M+1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\right) = \frac{M+1}{M}\mathbf{H}\mathbf{\Gamma}\cdot\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \mathbf{H};$$
$$\frac{\partial}{\partial \mathbf{\Gamma}} \frac{1}{2} \left(\frac{\operatorname{tr}\left\{\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\right\}}{M}\cdot\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\right) = \frac{1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\cdot\mathbf{H}\mathbf{\Gamma}\mathbf{H}.$$

Summing over the three equations above and applying the definition of gradient flow, we conclude the dynamics of $\mathbf{\Gamma}$ in (G.5). □

G.2 Proof of the global convergence

Let's now prove Theorem 6.3. Since the first part of Theorem 6.3 is directly implied by the second part, we only deal with the case with general $\mathbf{H}$. In this section, we denote $\lambda_{-1} > 0$ as the minimal non-zero eigenvalue of $\mathbf{H}$. As in the main text, we define

$$\mathcal{H} := \mathsf{Im}\left(\mathbf{H}\right)$$

and $\mathcal{H}^\perp$ as its orthogonal complement. First, let's prove the convergence of $\boldsymbol{\beta}$.

Lemma G.3 (Convergence of $\boldsymbol{\beta}$)**.** *Under the dynamical system* (G.4) *and* (G.5)*, one has*

$$\left\|\mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}(t)\right)-\mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}^*\right)\right\|_2^2 \le \exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}(0)\right)-\mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}^*\right)\right\|_2^2, \tag{G.6}$$

which implies

$$\left\|\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*\right)\right\|_2^2 \le \lambda_1 \exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\boldsymbol{\beta}(0)-\boldsymbol{\beta}^*\right\|_2^2, \tag{G.7}$$

and

$$\mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}(t)\right) \to \mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}^*\right)$$

when $t \to \infty$, from arbitrary initialization $\boldsymbol{\beta}(0)$ and $\mathbf{\Gamma}(0)$. Moreover, one has for any $t > 0$, it holds that

$$\mathcal{P}_{\mathcal{H}^\perp}\left(\boldsymbol{\beta}(t)\right) = \mathcal{P}_{\mathcal{H}^\perp}\left(\boldsymbol{\beta}(0)\right) \tag{G.8}$$

Proof. We first consider the orthogonal projection operator $\mathcal{P}$. From Lemma H.5 and equation (G.4), we know

$$\frac{\mathrm{d}\mathcal{P}_\mathcal{H}(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*)}{\mathrm{d}t} = -\mathbf{H}\mathbf{H}^+\mathbf{H}_\mathbf{\Gamma}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right),$$

where

$$\mathbf{H}_\mathbf{\Gamma} := \left(\mathbf{I}_d-\mathbf{\Gamma}\mathbf{H}\right)^\top\mathbf{H}\left(\mathbf{I}_d-\mathbf{\Gamma}\mathbf{H}\right)+\frac{\operatorname{tr}\left(\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\right)}{M}\mathbf{H}+\frac{1}{M}\mathbf{H}\mathbf{\Gamma}^\top\mathbf{H}\mathbf{\Gamma}\mathbf{H}.$$

From the property of pseudo-inverse, we know

$$\frac{\mathrm{d}\mathcal{P}_\mathcal{H}(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*)}{\mathrm{d}t} = -\mathbf{H}_\mathbf{\Gamma}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right) = -\mathbf{H}_\mathbf{\Gamma}\mathbf{H}^+\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right) = -\mathbf{H}_\mathbf{\Gamma}\mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right).$$

Therefore, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2}\left\|\mathcal{P}_\mathcal{H}(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*)\right\|_2^2\right] = -\mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\cdot\mathbf{H}_\mathbf{\Gamma}\cdot\mathcal{P}_\mathcal{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right).$$

Notice that

$$\mathbf{H}_\mathbf{\Gamma} \succeq \left(\sqrt{\frac{M}{M+1}}\mathbf{I}-\sqrt{\frac{M+1}{M}}\mathbf{\Gamma}\mathbf{H}\right)^\top\mathbf{H}\left(\sqrt{\frac{M}{M+1}}\mathbf{I}-\sqrt{\frac{M+1}{M}}\mathbf{\Gamma}\mathbf{H}\right)+\frac{1}{M+1}\mathbf{H} \succeq \frac{1}{M+1}\mathbf{H}.$$

This implies

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2}\left\|\mathcal{P}_{\mathcal{H}}(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*)\right\|_2^2\right] &\leq -\frac{1}{M+1}\mathcal{P}_{\mathcal{H}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\cdot\mathbf{H}\cdot\mathcal{P}_{\mathcal{H}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\\ &\leq -\frac{\lambda_{-1}}{M+1}\left\|\mathcal{P}_{\mathcal{H}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\right\|_2^2,\end{aligned}$$

where the last line comes from Lemma H.5 and λ_{-1} is the minimal non-zero eigenvalue of $\mathbf{H}$. Via standard integration method in ODE, we know this suggests

$$\left\|\mathcal{P}_{\mathcal{H}}\left(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*\right)\right\|_2^2\leq\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\mathcal{P}_{\mathcal{H}}\left(\boldsymbol{\beta}(0)-\boldsymbol{\beta}^*\right)\right\|_2^2\to 0\quad\text{when } t\to\infty.$$

This indicates that $\mathcal{P}_{\mathcal{H}}\left(\boldsymbol{\beta}(t)\right)\to\mathcal{P}_{\mathcal{H}}\left(\boldsymbol{\beta}^*\right)$ when $t\to\infty$. Finally, we have

$$\frac{\mathrm{d}\mathcal{P}_{\mathcal{H}^\perp}(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*)}{\mathrm{d}t}=-\left(\mathbf{I}_d-\mathbf{H}\mathbf{H}^+\right)\mathbf{H}_{\boldsymbol{\Gamma}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)=\mathbf{0}_d,$$

which implies $\mathcal{P}_{\mathcal{H}^\perp}(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*)=\mathcal{P}_{\mathcal{H}^\perp}(\boldsymbol{\beta}(0)-\boldsymbol{\beta}^*)$ and hence, $\mathcal{P}_{\mathcal{H}^\perp}(\boldsymbol{\beta}(t))=\mathcal{P}_{\mathcal{H}^\perp}(\boldsymbol{\beta}(0))$ for any $t>0$. □

Next, let's consider the convergence of $\boldsymbol{\Gamma}$. As defined in the main text, we have

$$\mathcal{Z}:=\mathsf{Im}\left(\mathbf{H}\otimes\mathbf{H}\right)$$

and $\mathcal{Z}^\perp$ is its orthogonal complement. We have the following lemma.

Lemma G.4. *Under the dynamical system* (G.4) *and* (G.5)*, one has*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{P}_{\mathcal{Z}^\perp}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)=\mathbf{0}_{d\times d},\tag{G.9}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)=&-\mathbf{H}\cdot\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\cdot\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}}-\frac{M+1}{M}\mathbf{H}\cdot\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\cdot\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\\ &-\frac{1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\cdot\mathbf{H}\cdot\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\cdot\mathbf{H}\\ &-\frac{1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\cdot\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}-\frac{1}{M}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\\ &+\frac{1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}+\left(\mathrm{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)+\sigma^2\right)\mathbf{I}_d\right)\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}.\end{aligned}\tag{G.10}$$

Proof. The first ODE is trivial since the RHS of (G.5) always lies in $\mathcal{Z}:=\mathsf{Im}\left(\mathbf{H}\otimes\mathbf{H}\right)$, so its projection on $\mathcal{Z}^\perp$ always vanishes. To prove the second equation, we can rewrite (G.5) as

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{\Gamma}}{\mathrm{d}t}=&-\left(\mathbf{H}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}}\underbrace{-\mathbf{H}\boldsymbol{\Psi}\mathbf{H}+\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}}}_{A}\right)\\ &-\frac{M+1}{M}\mathbf{H}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\\ &-\frac{1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\cdot\mathbf{H}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\mathbf{H}+\underbrace{\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}}_{B}\\ &-\frac{1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\cdot\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}\underbrace{-\frac{M+1}{M}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}}_{C}.\end{aligned}$$

For term A, recalling $\boldsymbol{\Gamma}^*=\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}$, we have

$$\begin{aligned}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}}&=\mathbf{H}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}}\\ &=\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}} &&(\boldsymbol{\Omega}\text{ and }\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\text{ commute.})\end{aligned}$$

$$\begin{aligned}
&= \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}} && (\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}})\\
&= \mathbf{H}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}} && (\boldsymbol{\Omega} \text{ and } \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} \text{ commute.})\\
&= \mathbf{H}\boldsymbol{\Psi}\mathbf{H}.
\end{aligned}$$

This suggests $A = 0$. For $B + C$, we have

$$B + C = -\frac{1}{M}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H} + \left(\mathbf{I}_d - \mathbf{H}\boldsymbol{\Gamma}^*\right)\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}.$$

We can compute $\left(\mathbf{I}_d - \mathbf{H}\boldsymbol{\Gamma}^*\right)\mathbf{H}$ as

$$\begin{aligned}
\left(\mathbf{I}_d - \mathbf{H}\boldsymbol{\Gamma}^*\right)\mathbf{H} &= \mathbf{H}^{\frac{1}{2}}\left(\mathbf{H}^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}\right)\\
&= \mathbf{H}^{\frac{1}{2}}\left(\mathbf{H}^{\frac{1}{2}} - \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}\right) && (\boldsymbol{\Omega} \text{ and } \mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} \text{ commute.})\\
&= \mathbf{H}^{\frac{1}{2}}\left(\mathbf{H}^{\frac{1}{2}} - \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}\right) && (\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}})\\
&= \mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}} - \boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}\right)\\
&= \frac{1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} + \left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2\right)\mathbf{I}_d\right)\mathbf{H}^{\frac{1}{2}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
B + C = &-\frac{1}{M}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\\
&+ \frac{1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} + \left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2\right)\mathbf{I}_d\right)\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}.
\end{aligned}$$

Bridging $A = 0$ and the result of $B + C$ into the ODE, we have

$$\begin{aligned}
&\frac{\mathrm{d}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)}{\mathrm{d}t}\\
=& -\mathbf{H}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{H}^{\frac{1}{2}} - \frac{M+1}{M}\mathbf{H}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\\
&- \frac{1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\cdot\mathbf{H}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\mathbf{H}\\
&- \frac{1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\cdot\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H} - \frac{1}{M}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\\
&+ \frac{1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} + \left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2\right)\mathbf{I}_d\right)\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}. \qquad \text{(G.11)}
\end{aligned}$$

Note that, all terms in the RHS above lie in $\mathcal{Z} = \mathsf{Im}\left(\mathbf{H}^{\otimes 2}\right)$, so this summation also equals $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)$. Moreover, since

$$\begin{aligned}
\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\mathbf{H}^{\frac{1}{2}} &= \mathbf{H}^{\frac{1}{2}}\left[\mathcal{P}_{\mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}}\otimes\mathbf{H}^{\frac{1}{2}}\right)}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right) + \mathcal{P}_{\mathsf{null}\left(\mathbf{H}^{\frac{1}{2}}\otimes\mathbf{H}^{\frac{1}{2}}\right)}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right]\mathbf{H}^{\frac{1}{2}}\\
&= \mathbf{H}^{\frac{1}{2}}\cdot\mathcal{P}_{\mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}}\otimes\mathbf{H}^{\frac{1}{2}}\right)}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\mathbf{H}^{\frac{1}{2}}. \qquad \text{(G.12)}
\end{aligned}$$

Bridging (G.12) into (G.11), we conclude. □

Then, let's upper bound the dynamics of $\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F^2$ in the following lemma.

Lemma G.5 (Dynamics of Frobenius norm)**.** *Under the dynamical system* (G.4) *and* (G.5)*, one has*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2}\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F^2\right] \leq -A_1\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F^2 + A_2\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F, \qquad \text{(G.13)}$$

where

$$A_1 = \lambda_{-1}^2 \lambda_{\min}\left(\mathbf{\Omega}\right), \qquad A_2 = \left(2 + \frac{2}{M}\right) \lambda_1^2 \left\|\boldsymbol{\beta}(0) - \boldsymbol{\beta}^*\right\|_2^2 . \tag{G.14}$$

are two positive constant. Here, $\lambda_{-1} > 0$ *is the minimal positive eigenvalue of* $\mathbf{H}$, λ_1 *is the maximal eigenvalue of* $\mathbf{H}$, $\lambda_{\min}\left(\mathbf{\Omega}\right)$ *is the minimal eigenvalue of* $\mathbf{\Omega}$ *(defined in* (A.3)*) and is strictly positive.*

Proof. From the dynamics of $\mathcal{P}_{\mathcal{Z}}\mathbf{\Gamma}$ in (G.10), we know

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2}\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)\right\|_F^2\right] &= \operatorname{tr}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right) \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top\right) \\ &= G_1 + G_2 + G_3, \end{aligned} \tag{G.15}$$

where

$$\begin{aligned} G_1 &= -\operatorname{tr}\left[\mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right) \cdot \mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}\mathbf{H}^{\frac{1}{2}} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top\right], \\ G_2 &= -\operatorname{tr}\left[\frac{M+1}{M}\mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right) \cdot \mathbf{H}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top\right] \\ &\qquad - \operatorname{tr}\left[\frac{1}{M}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right) \cdot \mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right) \cdot \mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top\right], \\ G_3 &= -\operatorname{tr}\left[\frac{1}{M}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right) \cdot \mathbf{H}\mathbf{\Gamma}^*\mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top\right] \\ &\quad + \frac{1}{M}\operatorname{tr}\left[\mathbf{H}\mathbf{\Gamma}^*\mathbf{H}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top\right] \\ &\quad + \operatorname{tr}\left[\frac{1}{M}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}} + \left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\right) + \sigma^2\right)\mathbf{I}_d\right)\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top\right]. \end{aligned} \tag{G.16}$$

From Lemma H.2, we know that $G_2 \le 0$. Then, we consider G_1 and G_3. For G_1, we have

$$\begin{aligned} G_1 &= -\operatorname{tr}\left[\left(\mathbf{H}^{\frac{1}{2}}\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top \mathbf{H}^{\frac{1}{2}}\right) \cdot \left(\mathbf{H}^{\frac{1}{2}}\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)\mathbf{H}^{\frac{1}{2}}\right) \cdot \mathbf{\Omega}\right] \\ &\le -\lambda_{-1}^2 \operatorname{tr}\left[\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right) \cdot \mathbf{\Omega}\right] \quad \text{((3) in Lemma H.6 and (2) in Lemma H.2)} \\ &\le -\lambda_{-1}^2 \lambda_{\min}\left(\mathbf{\Omega}\right)\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)\right\|_F^2, \end{aligned}$$

where $\lambda_{\min}(\cdot)$ denote the minimal eigenvalue. Since $\mathbf{\Omega}$ is positive definite, we know $\lambda_{\min}\left(\mathbf{\Omega}\right) > 0$. Finally, let's upper bound G_3. First, we have

$$\begin{aligned} &- \operatorname{tr}\left[\frac{1}{M}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right) \cdot \mathbf{H}\mathbf{\Gamma}^*\mathbf{H} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top\right] \\ \le& \frac{\sqrt{d}}{M}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^\top \mathbf{H}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right) \cdot \left\|\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right\|_2 \left\|\mathbf{H}^{\frac{1}{2}} \cdot \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)^\top \mathbf{H}^{\frac{1}{2}}\right\|_F \qquad \text{(Lemma H.1)} \\ \le& \frac{\sqrt{d}\lambda_1}{M}\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\boldsymbol{\beta}(0) - \boldsymbol{\beta}^*\right\|_2^2 \lambda_{\max}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right) \cdot \left\|\mathbf{H}\right\|_F \left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)\right\|_F \\ & \qquad \text{(from equation (G.7))} \\ \le& \frac{\sqrt{d}\lambda_1^2}{M}\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\boldsymbol{\beta}(0) - \boldsymbol{\beta}^*\right\|_2^2 \lambda_{\max}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right) \cdot \left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma} - \mathbf{\Gamma}^*\right)\right\|_F . \end{aligned}$$

For $\lambda_{\max}\left(\mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right)$, we recall $\mathbf{\Gamma}^* = \mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}$ and obtain

$$\begin{aligned} \mathbf{H}^{\frac{1}{2}}\mathbf{\Gamma}^*\mathbf{H}^{\frac{1}{2}} &= \mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{\Omega}^{-1}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} = \mathbf{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} \qquad (\mathbf{\Omega} \text{ and } \mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}} \text{ commute}) \\ &= \mathbf{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{H}^{\frac{1}{2}}. \qquad (\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}}) \end{aligned}$$

Since $\boldsymbol{\Omega}^{-1}$ and $\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}}$ commute, they are simultaneously diagonalizable. The eigenvalues of $\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}$ are

$$\frac{\phi_i}{\frac{M+1}{M}\phi_i + \frac{\sum_i \phi_i + \sigma^2}{M}}, \quad i = 1, 2, ..., d.$$

Note that, every eigenvalue is upper bounded by 1, so we simply get

$$\lambda_{\max}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right) \le 1. \tag{G.17}$$

Therefore, we have

$$\begin{aligned}
&-\operatorname{tr}\left[\frac{1}{M}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\cdot \mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}\cdot \mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)^\top\right] \\
\le &\frac{\sqrt{d}\lambda_1^2}{M}\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\boldsymbol{\beta}(0)-\boldsymbol{\beta}^*\right\|_2^2 \cdot \left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F.
\end{aligned} \tag{G.18}$$

Similarly, we have

$$\begin{aligned}
&-\operatorname{tr}\left[\frac{1}{M}\mathbf{H}\boldsymbol{\Gamma}^*\mathbf{H}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top \mathbf{H}\cdot \mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)^\top\right] \\
\le &\frac{\sqrt{d}}{M}\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)^\top\right\|_F \left\|\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}^{\frac{1}{2}}\right\|_F \left\|\mathbf{H}\right\|_F \left\|\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Gamma}^*\mathbf{H}^{\frac{1}{2}}\right\|_2 \quad \text{(Lemma H.1)} \\
\le &\frac{\sqrt{d}\lambda_1^2}{M}\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\boldsymbol{\beta}(0)-\boldsymbol{\beta}^*\right\|_2^2 \cdot \left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F,
\end{aligned} \tag{G.19}$$

and

$$\begin{aligned}
&\operatorname{tr}\left[\frac{1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}+\left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)+\sigma^2\right)\mathbf{I}_d\right)\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\cdot\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)^\top\right] \\
\le &\frac{\sqrt{d}}{M}\left\|\boldsymbol{\Omega}^{-1}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}+\left(\operatorname{tr}\left(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\right)+\sigma^2\right)\mathbf{I}_d\right)\right\|_2\left\|\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}\cdot\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)^\top\mathbf{H}^{\frac{1}{2}}\right\|_F \\
&\text{(Lemma H.1)} \\
\le &\frac{\sqrt{d}}{M}\cdot M\cdot\left\|\mathbf{H}^{\frac{1}{2}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)^\top\mathbf{H}^{\frac{1}{2}}\right\|_F\cdot\left\|\mathbf{H}\right\|_F\cdot\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F \\
\le &\sqrt{d}\lambda_1^2\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\boldsymbol{\beta}(0)-\boldsymbol{\beta}^*\right\|_2^2\cdot\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F
\end{aligned} \tag{G.20}$$

Bridging (G.18), (G.19) and (G.20) into (G.16), we get

$$G_3 \le \left(2+\frac{2}{M}\right)\lambda_1^2\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right)\left\|\boldsymbol{\beta}(0)-\boldsymbol{\beta}^*\right\|_2^2\cdot\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F \tag{G.21}$$

□

Finally, we finish this section by proving the convergence of $\boldsymbol{\Gamma}$.

Lemma G.6 (Convergence of $\boldsymbol{\Gamma}$)**.** *Under the dynamical system* (G.4) *and* (G.5)*, one has*

$$\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}(t)\right)\to\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}^*\right), \quad \mathcal{P}_{\mathcal{Z}^\perp}\left(\boldsymbol{\Gamma}(t)\right)=\mathcal{P}_{\mathcal{Z}^\perp}\left(\boldsymbol{\Gamma}(0)\right)$$

when $t\to\infty$, from arbitrary initialization $\boldsymbol{\beta}(0)$ and $\boldsymbol{\Gamma}(0)$.

Proof. We observe that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2}\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F^2\right]=\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F\cdot\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F.$$

Combining it with Lemma G.5, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F\le -A_1\left\|\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\right)\right\|_F+A_2\exp\left(\frac{-2\lambda_{-1}t}{M+1}\right),$$

where $A_1 > 0, A_2 > 0$ are defined in (G.14). Simple calculation shows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\exp\left(A_1 t\right)\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}-\mathbf{\Gamma}^*\right)\right\|_F\right] \leq A_2 \exp\left[\left(A_1 - \frac{2\lambda_{-1}}{M+1}\right)t\right]. \tag{G.22}$$

When $A_1 \neq \frac{2\lambda_{-1}}{M+1}$, we integrate both sides from $t = 0$ to $t = T$, then divide them by $\exp\left(A_1 T\right)$. This gives

$$\begin{aligned}\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}(T)-\mathbf{\Gamma}^*\right)\right\|_F &\leq \frac{\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}(0)-\mathbf{\Gamma}^*\right)\right\|_F}{\exp\left(A_1 T\right)} + \frac{A_2}{A_1 - \frac{2\lambda_{-1}}{M+1}} \cdot \frac{\exp\left[\left(A_1 - \frac{2\lambda_{-1}}{M+1}\right)T\right] - 1}{\exp\left(A_1 T\right)} \\ &= \frac{\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}(0)-\mathbf{\Gamma}^*\right)\right\|_F}{\exp\left(A_1 T\right)} + \frac{A_2}{A_1 - \frac{2\lambda_{-1}}{M+1}} \cdot \left[\exp\left(-\frac{2\lambda_{-1}T}{M+1}\right) - \exp\left(-A_1 T\right)\right] \to 0\end{aligned}$$

when $T \to \infty$. Otherwise, if $A_1 = \frac{2\lambda_{-1}}{M+1}$, the right hand side of (G.22) reduces to a constant, which implies $\exp\left(A_1 t\right)\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}(t)-\mathbf{\Gamma}^*\right)\right\|_F$ grows at most at a linear rate, which implies $\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}(t)-\mathbf{\Gamma}^*\right)\right\|_F \to 0$ when $t \to \infty$. Together, in all cases, we have $\left\|\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}(t)-\mathbf{\Gamma}^*\right)\right\|_F \to 0$. From (G.9), we soon get $\mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}(t)\right) = \mathcal{P}_{\mathcal{Z}}\left(\mathbf{\Gamma}(0)\right)$ foo all $t > 0$. Therefore, we conclude. □

The Theorem 6.3 is proved by simplly combining Lemma G.3 and Lemma G.6.

H Technical lemmas

Lemma H.1 (Von-Neumann's Trace Inequality)**.** *Let $U, V \in \mathbb{R}^{d\times n}$ with $d \le n$. We have*

$$\operatorname{tr}\left(U^\top V\right) \le \sum_{i=1}^{d} \sigma_i(U)\sigma_i(V) \le \|U\|_{\mathrm{op}} \times \sum_{i=1}^{d} \sigma_i(V) \le \sqrt{d}\cdot \|U\|_{\mathrm{op}}\|V\|_F$$

where $\sigma_1(X) \ge \sigma_2(X) \ge \cdots \ge \sigma_d(X)$ are the ordered singular values of $X \in \mathbb{R}^{d\times n}$.

Lemma H.2 ([22])**.** *For any two positive semi-definite matrices $A, B \in \mathbb{R}^{d\times d}$, we have*

- $\operatorname{tr}[AB] \ge 0$.
- $AB \succeq 0$ *if and only if A and B commute.*

Lemma H.3 (Derivatives, [27])**.** *We denote $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{X}$ as matrices and $\mathbf{a}, \mathbf{b}, \mathbf{x}$ as vectors. Then, we have*

- $\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}\left[\mathbf{X}^\top \mathbf{B}\mathbf{X}\mathbf{C}\right] = \mathbf{B}\mathbf{X}\mathbf{C} + \mathbf{B}^\top \mathbf{X}\mathbf{C}^\top$.
- $\frac{\partial \mathbf{x}^\top \mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = \left(\mathbf{B} + \mathbf{B}^\top\right)\mathbf{x}$.
- $\frac{\partial \mathbf{a}^\top \mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^\top$.
- $\frac{\partial \mathbf{a}^\top \mathbf{X}^\top \mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^\top$.
- $\frac{\partial \mathbf{b}^\top \mathbf{X}^\top \mathbf{D}\mathbf{X}\mathbf{c}}{\partial \mathbf{X}} = \mathbf{D}^\top \mathbf{X}\mathbf{b}\mathbf{c}^\top + \mathbf{D}\mathbf{X}\mathbf{c}\mathbf{b}^\top$.
- $\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}\left[(\mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{C})(\mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{C})^\top\right] = 2\mathbf{A}^\top(\mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{C})\mathbf{B}^\top$

Lemma H.4 (Lemma D.2 in [38], Lemma 4.2 in [37])**.** *If $\mathbf{x}$ is Gaussian random vector of d dimension, mean zero and covariance matrix $\mathbf{H}$, and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is a fixed matrix. Then*

$$\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top \boldsymbol{A}\mathbf{x}\mathbf{x}^\top\right] = \mathbf{H}\left(\boldsymbol{A} + \boldsymbol{A}^\top\right)\mathbf{H} + \operatorname{tr}(\boldsymbol{A}\mathbf{H})\mathbf{H}.$$

If $\boldsymbol{A}$ is symmetric and the rows in $\mathbf{X} \in \mathbb{R}^{M\times d}$ are generated independently from

$$\mathbf{X}[i] \sim \mathcal{N}(0, \mathbf{H}), \quad i = 1, \ldots, M.$$

Then, it holds that

$$\mathbb{E}\left[\mathbf{X}^\top \mathbf{X}\boldsymbol{A}\mathbf{X}^\top \mathbf{X}\right] = M \cdot \operatorname{tr}(\mathbf{H}\boldsymbol{A}) \cdot \mathbf{H} + M(M+1) \cdot \mathbf{H}\boldsymbol{A}\mathbf{H}.$$

Lemma H.5 (Linear Algebra)**.** *Suppose $\mathbf{H}$ is a (non-zero) positive semi-definite matrix in $\mathbb{R}^{d\times d}$ and $\mathbf{H}^{\frac{1}{2}}$ is its principle square root. We denote λ_{-1} as the minimal non-zero eigenvector of $\mathbf{H}$. Then, we have*

- *1.* $\mathsf{null}(\mathbf{H}) = \mathsf{null}\left(\mathbf{H}^{\frac{1}{2}}\right), \mathsf{Im}(\mathbf{H}) = \mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}}\right)$.
- *2.* $\mathbf{H}\mathbf{H}^+ = \mathbf{H}^+\mathbf{H}$, *where $(\cdot)^+$ denotes the Moore-Penrose pseudo-inverse.*
- *3. For any vector $\boldsymbol{\alpha} \in \mathbb{R}^d$, the orthogonal projection operator on $\mathsf{Im}(\mathbf{H})$ and $\mathsf{null}(\mathbf{H})$ are respectively*

  $$\mathcal{P}_{\mathsf{Im}(\mathbf{H})}(\boldsymbol{\alpha}) = \mathbf{H}\mathbf{H}^+\boldsymbol{\alpha} = \mathbf{H}^+\mathbf{H}\boldsymbol{\alpha}, \quad \mathcal{P}_{\mathsf{null}(\mathbf{H})}(\boldsymbol{\alpha}) = \left(\mathbf{I}_d - \mathbf{H}\mathbf{H}^+\right)\boldsymbol{\alpha} = \left(\mathbf{I}_d - \mathbf{H}^+\mathbf{H}\right)\boldsymbol{\alpha}.$$

- *4. For any vector $\boldsymbol{\alpha} \in \mathbb{R}^d$, we have*

  $$\mathcal{P}_{\mathsf{Im}(\mathbf{H})}(\boldsymbol{\alpha})^\top \mathbf{H}\mathcal{P}_{\mathsf{Im}(\mathbf{H})}(\boldsymbol{\alpha}) \ge \lambda_{-1}\left\|\mathcal{P}_{\mathsf{Im}(\mathbf{H})}(\boldsymbol{\alpha})\right\|_2^2$$

Proof. We consider the eigen-decomposition of matrix $\mathbf{H}$, which is

$$\mathbf{H} = \mathbf{Q}\boldsymbol{D}\mathbf{Q}^\top, \tag{H.1}$$

where $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \lambda_2, ..., \lambda_d)$ is a diagonal matrix with diagonal entries being the eigenvalues of $\mathbf{H}$, and $\mathbf{Q}$ is an orthogonal matrix. Then, from the definition of principle square root, we know

$$\mathbf{H}^{\frac{1}{2}} = \mathbf{Q}\boldsymbol{D}^{\frac{1}{2}}\mathbf{Q}^\top, \tag{H.2}$$

where $\boldsymbol{D}^{\frac{1}{2}} = \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_d}\right)$. We denote columns of $\mathbf{Q}$ as $\mathbf{q}_1, ...\mathbf{q}_d$. We know they are the eigenvectors of $\mathbf{H}$. From the eigen-decomposition of $\mathbf{H}$ and $\mathbf{H}^{\frac{1}{2}}$, we know they share the same set of eigenvectors. Without loss of generality, we assume $\mathsf{rank}(\mathbf{H}) = r$ and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0, \lambda_i = 0, i = r+1, ..., d$. Then, we denote $\mathcal{H}$ as the linear vector space spanned by $\mathbf{q}_1, ..., \mathbf{q}_r$ and $\mathcal{H}^\perp$ as its orthogonal complement (which is the subspace spanned by $\mathbf{q}_{r+1}, ..., \mathbf{q}_d$). For any vector $\boldsymbol{\alpha} \in \mathbb{R}^d$, we can write its orthogonal decomposition as $\boldsymbol{\alpha} = \sum_{i=1}^d a_i \mathbf{q}_i$, so we have

$$\mathbf{H}\boldsymbol{\alpha} = \sum_{i=1}^d a_i \mathbf{H}\mathbf{q}_i = \sum_{i=1}^r a_i \lambda_i \mathbf{q}_i \in \mathcal{H},$$

which implies $\mathsf{Im}(\mathbf{H}) \subset \mathcal{H}$. On the other hand, we know for any $\boldsymbol{\alpha} \in \mathcal{H}$, we can write it as $\boldsymbol{\alpha} = \sum_{i=1}^r a_i \mathbf{q}_i$ for some $a_1, ..., a_r$, then we have

$$\boldsymbol{\alpha} = \sum_{i=1}^r \frac{a_i}{\lambda_i} \cdot \lambda_i \mathbf{q}_i = \mathbf{H}\left(\sum_{i=1}^r \frac{a_i}{\lambda_i}\mathbf{q}_i\right) \in \mathsf{Im}(\mathbf{H}),$$

which implies $\mathcal{H} \subset \mathsf{Im}(\mathbf{H})$. This shows

$$\mathsf{Im}(\mathbf{H}) = \mathcal{H} = \mathsf{span}\{\mathbf{q}_1, ...\mathbf{q}_r\},$$

and

$$\mathsf{null}(\mathbf{H}) = \mathcal{H} = \mathsf{span}\{\mathbf{q}_{r+1}, ...\mathbf{q}_d\},$$

since $\mathsf{null}(\mathbf{H})$ is the orthogonal complement of $\mathsf{Im}(\mathbf{H})$. Then, (1) is proved by noticing that $\mathbf{H}$ and $\mathbf{H}^{\frac{1}{2}}$ share the same set of eigenvectors. To prove (2), it suffices to notice that

$$\mathbf{H}\mathbf{H}^+ = \mathbf{Q} \cdot \operatorname{diag}(\underbrace{1, 1, ..., 1}_{\text{r}}, 0, 0, .., 0) \cdot \mathbf{Q}^\top = \mathbf{H}^+\mathbf{H}.$$

To prove (3), it suffices to notice that actually $\mathbf{H}\mathbf{H}^+\boldsymbol{\alpha} \in \mathsf{Im}(\mathbf{H})$ and

$$\left\langle \mathbf{H}\mathbf{H}^+\boldsymbol{\alpha}, \left(\mathbf{I}_d - \mathbf{H}\mathbf{H}^+\right)\boldsymbol{\alpha}\right\rangle = \left\langle \mathbf{H}^+\mathbf{H}\boldsymbol{\alpha}, \left(\mathbf{I}_d - \mathbf{H}\mathbf{H}^+\right)\boldsymbol{\alpha}\right\rangle = \boldsymbol{\alpha}^\top\left(\mathbf{H}\mathbf{H}^+ - \mathbf{H}\mathbf{H}^+\mathbf{H}\mathbf{H}^+\right)\boldsymbol{\alpha} = 0.$$

Finally, to prove (4), we can write $\boldsymbol{\alpha} = \sum_{i=1}^d a_i \mathbf{q}_i$ for some $a_1, ..., a_d$. Then, by orthogonality we have

$$\begin{aligned}\mathcal{P}_{\mathsf{Im}(\mathbf{H})}(\boldsymbol{\alpha})^\top \mathbf{H}\mathcal{P}_{\mathsf{Im}(\mathbf{H})}(\boldsymbol{\alpha}) &= \left(\sum_{i=1}^d a_i \mathbf{q}_i\right)^\top \mathbf{H}\left(\sum_{i=1}^d a_i \mathbf{q}_i\right) = \sum_{i=1}^r a_i^2 \lambda_i \mathbf{q}_i^\top \mathbf{q}_i \geq \lambda_{-1} \cdot \sum_{i=1}^r a_i^2 \mathbf{q}_i^\top \mathbf{q}_i \\ &= \lambda_{-1}\left\|\mathcal{P}_{\mathsf{Im}(\mathbf{H})}(\boldsymbol{\alpha})\right\|_2^2.\end{aligned}$$

Therefore, we conclude. □

Lemma H.6 (Tensor product). *Suppose* $\mathbf{H}$ *is a (non-zero) positive semi-definite matrix in* $\mathbb{R}^{d\times d}$ *and* $\mathbf{H}^{\frac{1}{2}}$ *is its principle square root. We denote* $\otimes$ *as Kronecker product, which is defined as*

$$(\boldsymbol{A} \otimes \boldsymbol{B}) \circ \boldsymbol{C} = \boldsymbol{B}\boldsymbol{C}\boldsymbol{A}^\top.$$

We define

$$\mathsf{Im}(\mathbf{H} \otimes \mathbf{H}) := \left\{(\mathbf{H} \otimes \mathbf{H}) \circ \boldsymbol{Z} : \boldsymbol{Z} \in \mathbb{R}^{d\times d}\right\}, \quad \mathsf{null}(\mathbf{H} \otimes \mathbf{H}) := \left\{\boldsymbol{Z} \in \mathbb{R}^{d\times d} : (\mathbf{H} \otimes \mathbf{H}) \circ \boldsymbol{Z} = \mathbf{0}\right\},$$

and define $\mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right)$ *and* $\mathsf{null}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right)$ *similarly. Then, we have*

- *1.* $\mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) = \mathsf{Im}\left(\mathbf{H} \otimes \mathbf{H}\right), \mathsf{null}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) = \mathsf{null}\left(\mathbf{H} \otimes \mathbf{H}\right).$

- *2. We denote* $\mathcal{Z} = \mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) = \mathsf{Im}\left(\mathbf{H} \otimes \mathbf{H}\right)$ *and* $\mathcal{Z}^{\perp} = \mathsf{null}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) = \mathsf{null}\left(\mathbf{H} \otimes \mathbf{H}\right)$. *For any matrix* $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$, *we have*

$$\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{Z}\right) = \left(\mathbf{H}\mathbf{H}^{+}\right)^{\otimes 2} \circ \boldsymbol{Z} = \mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H} = \mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\boldsymbol{Z}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} \tag{H.3}$$

$$\mathcal{P}_{\mathcal{Z}^{\perp}}\left(\boldsymbol{Z}\right) = \left[\mathbf{I}_{d}^{\otimes 2} - \left(\mathbf{H}\mathbf{H}^{+}\right)^{\otimes 2}\right] \circ \boldsymbol{Z} = \boldsymbol{Z} - \mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H} = \boldsymbol{Z} - \mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\boldsymbol{Z}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}}. \tag{H.4}$$

- *3. For any matrix* $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$, *it holds that*

$$\left(\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) \circ \mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{Z}\right)\right) \cdot \left(\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) \circ \mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{Z}\right)\right)^{\top} \succeq \lambda_{-1}^{2}\mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{Z}\right) \cdot \mathcal{P}_{\mathcal{Z}}\left(\boldsymbol{Z}\right)^{\top}, \tag{H.5}$$

where λ_{-1} *is the minimal positive eigenvector of* $\mathbf{H}$.

Proof. Let's first prove the second part. From the definition of tensor product and the fact that $\mathbf{H}\mathbf{H}^{+} = \mathbf{H}^{+}\mathbf{H}$ (see Lemma H.5), we know the second equation in H.3 holds. To prove the first equation, it suffices to notice that for any matrix $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$, the matrix $\mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H}$ is in $\mathsf{Im}\left(\mathbf{H}^{\otimes 2}\right)$, together with the fact that

$$\begin{aligned}\left\langle \mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H}, \boldsymbol{Z} - \mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H}\right\rangle &= \operatorname{tr}\left(\mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H}\boldsymbol{Z}^{\top}\right) - \operatorname{tr}\left(\mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H}\mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}^{\top}\mathbf{H}^{+}\mathbf{H}\right)\\ &= \operatorname{tr}\left(\mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H}\boldsymbol{Z}^{\top}\right) - \operatorname{tr}\left(\mathbf{H}\mathbf{H}^{+}\mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H}\mathbf{H}^{+}\mathbf{H}\boldsymbol{Z}^{\top}\right)\\ &\qquad\qquad (\mathbf{H}\mathbf{H}^{+} = \mathbf{H}^{+}\mathbf{H})\\ &= 0.\end{aligned}$$

The proof of (H.4) is similar. Then, from Lemma H.5, we know $\mathsf{Im}\left(\mathbf{H}\right) = \mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}}\right)$, which implies the projection operator onto those two subspace are identical, indicating $\mathbf{H}\mathbf{H}^{+} = \mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}$. Therefore, we have

$$\mathcal{P}_{\mathsf{Im}(\mathbf{H}\otimes\mathbf{H})}\left(\boldsymbol{Z}\right) = \left(\mathbf{H}\mathbf{H}^{+}\right)^{\otimes 2} \circ \boldsymbol{Z} = \mathbf{H}\mathbf{H}^{+}\boldsymbol{Z}\mathbf{H}^{+}\mathbf{H} = \mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\boldsymbol{Z}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} = \mathcal{P}_{\mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}}\otimes\mathbf{H}^{\frac{1}{2}}\right)}\left(\boldsymbol{Z}\right).$$

Since this holds for any matrix $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$, this suggests $\mathsf{Im}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) = \mathsf{Im}\left(\mathbf{H} \otimes \mathbf{H}\right)$. Similarly, we also have $\mathsf{null}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) = \mathsf{null}\left(\mathbf{H} \otimes \mathbf{H}\right)$. Finally, to prove (3), we use the eigen-decomposition of $\mathbf{H}$ and $\mathbf{H}^{\frac{1}{2}}$:

$$\mathbf{H} = \mathbf{Q}\boldsymbol{D}\mathbf{Q}^{\top}, \quad \mathbf{H}^{\frac{1}{2}} = \mathbf{Q}\boldsymbol{D}^{\frac{1}{2}}\mathbf{Q}^{\top},$$

where $\mathbf{Q} = \left(\mathbf{q}_1\ \mathbf{q}_2\ ...\ \mathbf{q}_d\right)$ is an orthogonal matrix and $\boldsymbol{D} = \operatorname{diag}\left(\lambda_1, ..., \lambda_d\right)$, where $\lambda_1 \geq \lambda_2 \geq ....\lambda_d \geq 0$ are ordered eigenvalues. We assume $\mathsf{rank}(\mathbf{H}) = r$ and $\lambda_r > 0 = \lambda_{r+1}$. Then, from the property of Kronecker product (eg. see [27]), the eigenvalues of $\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}$ are $\left\{\sqrt{\lambda_i\lambda_j} : 1 \leq i, j \leq d\right\}$. The eigenvectors are $\left\{\mathbf{q}_i \otimes \mathbf{q}_j : 1 \leq i, j \leq d\right\}$, which forms an orthogonal unit basis of $\mathbb{R}^{d\times d}$. We define

$$\mathcal{S} := \mathsf{span}\left\{\mathbf{q}_i \otimes \mathbf{q}_j : 1 \leq i, j \leq r\right\}$$

as the eigenspace spanned by all eigenvectors corresponding to positive eigenvalues. For any matrix $\boldsymbol{Z}$, we can decompose it as

$$\boldsymbol{Z} = \sum_{i,j} a_{i,j} \cdot \mathbf{q}_i \otimes \mathbf{q}_j,$$

so

$$\left(\mathbf{H} \otimes \mathbf{H}\right) \circ \boldsymbol{Z} = \sum_{i,j=1}^{d} a_{i,j}\left(\mathbf{H} \otimes \mathbf{H}\right) \circ \left(\mathbf{q}_i \otimes \mathbf{q}_j\right) = \sum_{i,j}\left(\mathbf{H}\mathbf{q}_i\right) \otimes \left(\mathbf{H}\mathbf{q}_j\right)$$

$$= \sum_{i,j=1}^{d} a_{i,j}\lambda_i\lambda_j\mathbf{q}_i \otimes \mathbf{q}_j = \sum_{i,j=1}^{r} a_{i,j}\lambda_i\lambda_j\mathbf{q}_i \otimes \mathbf{q}_j \in \mathcal{S},$$

which implies $\mathcal{Z} \subset \mathcal{S}$. On the other hand, for any $\boldsymbol{Z} \in \mathcal{S}$, we can write it as $\boldsymbol{Z} = \sum_{i,j=1}^{r} b_{i,j} \cdot \mathbf{q}_i \otimes \mathbf{q}_j$, so we have

$$\boldsymbol{Z} = \sum_{i,j=1}^{r} \frac{b_{i,j}}{\lambda_i\lambda_j} \cdot (\lambda_i\mathbf{q}_i) \otimes (\lambda_j\mathbf{q}_j) = \sum_{i,j=1}^{r} \frac{b_{i,j}}{\lambda_i\lambda_j} \cdot (\mathbf{H}\mathbf{q}_i) \otimes (\mathbf{H}\mathbf{q}_j) = (\mathbf{H} \otimes \mathbf{H}) \circ \left[\sum_{i,j=1}^{r} \frac{b_{i,j}}{\lambda_i\lambda_j} (\mathbf{q}_i \otimes \mathbf{q}_j)\right] \in \mathcal{Z},$$

which implies $\mathcal{S} \subset \mathcal{Z}$. Combining two directions, we have $\mathcal{S} = \mathcal{Z}$. Therefore, for any matrix $\boldsymbol{Z} \in \mathbb{R}^{d \times d}$, we can write it as $\boldsymbol{Z} = \sum_{i,j=1}^{d} c_{i,j} \cdot \mathbf{q}_i \otimes \mathbf{q}_j$ for some $c_{i,j}$. Then, we have

$$\begin{aligned}\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) \circ \mathcal{P}_{\mathcal{Z}}(\boldsymbol{Z}) = \left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) \circ \mathcal{P}_{\mathcal{S}}(\boldsymbol{Z}) &= \left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) \circ \sum_{i,j=1}^{r} c_{i,j} \cdot \mathbf{q}_i \otimes \mathbf{q}_j \\ &= \sum_{i,j=1}^{r} c_{i,j} \left(\mathbf{H}^{\frac{1}{2}}\mathbf{q}_i\right) \otimes \left(\mathbf{H}^{\frac{1}{2}}\mathbf{q}_j\right) = \sum_{i,j=1}^{r} c_{i,j}\sqrt{\lambda_i\lambda_j} \cdot \mathbf{q}_i \otimes \mathbf{q}_j.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\left(\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) \circ \mathcal{P}_{\mathcal{Z}}(\boldsymbol{Z})\right)\left(\left(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{H}^{\frac{1}{2}}\right) \circ \mathcal{P}_{\mathcal{Z}}(\boldsymbol{Z})\right)^{\top} \\ &= \left(\sum_{i,j=1}^{r} c_{i,j}\sqrt{\lambda_i\lambda_j} \cdot \mathbf{q}_i \otimes \mathbf{q}_j\right)\left(\sum_{k,l=1}^{r} c_{k,l}\sqrt{\lambda_k\lambda_l} \cdot \mathbf{q}_k \otimes \mathbf{q}_l\right)^{\top} \\ &= \sum_{i,j=1}^{r} c_{i,j}^2\lambda_i\lambda_j (\mathbf{q}_i \otimes \mathbf{q}_j)(\mathbf{q}_i \otimes \mathbf{q}_j)^{\top} && \text{(By orthogonality)} \\ &\succeq \lambda_{-1}^2 \sum_{i,j=1}^{r} c_{i,j}^2 (\mathbf{q}_i \otimes \mathbf{q}_j)(\mathbf{q}_i \otimes \mathbf{q}_j)^{\top} \\ &= \lambda_{-1}^2 \mathcal{P}_{\mathcal{Z}}(\boldsymbol{Z}) \cdot \mathcal{P}_{\mathcal{Z}}(\boldsymbol{Z})^{\top}.\end{aligned}$$

Therefore, we conclude. □

I Experiment Details

Our experiments on GPT2 mostly follow the setting in [14], except that we use the token matrix defined in (3.1). In our experiments, we use $d = 20$, $M = 40$, $\boldsymbol{\Psi} = \mathbf{H} = \mathbf{I}_d$ and $\sigma = 0$. We train the GPT2 with and without MLP layers on two settings: $\boldsymbol{\beta}^* = (0, 0, ..., 0)^{\top}$ and $\boldsymbol{\beta}^* = (10, 10, ..., 10)^{\top}$. For GPT2 with and without MLP layers, we initialize the $\mathbf{W}_V$ and $\mathbf{W}_2$ by normal distribution with standard deviation $0.02/\sqrt{\text{number of residual connections}}$, where the number of residual connections equals the number of layers for GPT2 without MLP layers. For GPT2 with MLP layers, this is 2 times the number of layers. Initialization for other matrices follow the default setting in [36]. We use Adam with learning rate 0.0001. We train the model for 200000 steps and we sample a batch of 256 new tasks for each step.

J Does scratchpad help?

In this section, we show the limitations of adding a scratchpad to the token matrix. We will show that by using a single LSA layer and the token matrix with scratchpad, one cannot recover the GD-$\boldsymbol{\beta}$ estimator defined in Section 5. We leave it as future work to see whether the token matrix with scratchpad could implement other types of estimators that more effectively address the linear regression tasks defined in Assumption 3.1, as well as whether additional structures could help alleviate this approximation error. We follow the notations in previous sections. The token matrix with scratchpad is defined as

$$\mathbf{E} := \begin{pmatrix} \mathbf{X}^\top & \mathbf{x} \\ \mathbf{1}_M^\top & 1 \\ \mathbf{y}^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d+2)\times(M+1)}. \tag{J.1}$$

where $\mathbf{1}_M \in \mathbb{R}^M$ refers to the vector filled with ones.

A LSA model f, is defined by

$$f(\mathbf{E}) := \left[\mathbf{E} + \mathbf{W}_P^\top \mathbf{W}_V \mathbf{E}\mathbf{M} \frac{\mathbf{E}^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{E}}{M}\right]_{-1,-1} = \left[\mathbf{W}_P^\top \mathbf{W}_V \mathbf{E}\mathbf{M} \frac{\mathbf{E}^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{E}}{M}\right]_{-1,-1},$$

where the second equality is because the bottom right entry of $\mathbf{E}$ is zero (see (3.1)). Note that the prediction is the bottom right entry of the output matrix. So only the last row of $\mathbf{W}_P^\top \mathbf{W}_V$ and the last column of $\mathbf{E}$ attend the prediction. Denote

$$\mathbf{W}_P^\top \mathbf{W}_V = \begin{pmatrix} * & * & * \\ \boldsymbol{w}^\top & \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{pmatrix}, \quad \mathbf{W}_K^\top \mathbf{W}_Q = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{b} & * \\ \boldsymbol{q}_1^\top & \boldsymbol{b}_1 & * \\ \boldsymbol{q}_2^\top & \boldsymbol{b}_2 & * \end{pmatrix},$$

where

$$\boldsymbol{w} \in \mathbb{R}^d,\ \boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R},\ \boldsymbol{Q} \in \mathbb{R}^{d\times d},\ \boldsymbol{b}, \boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathbb{R}^d,\ \boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R},$$

and $*$ denotes entries that do not enter the final prediction. Then we have

$$\begin{aligned} f(\mathbf{E}) &= \begin{pmatrix} \boldsymbol{w}^\top & \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{pmatrix} \frac{\mathbf{E}\mathbf{M}_M \mathbf{E}^\top}{M} \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{b} & * \\ \boldsymbol{q}_1^\top & \boldsymbol{b}_1 & * \\ \boldsymbol{q}_2^\top & \boldsymbol{b}_2 & * \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{w}^\top & \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{pmatrix} \frac{\mathbf{E}\mathbf{M}_M \mathbf{E}^\top}{M} \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{b} \\ \boldsymbol{q}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{q}_2^\top & \boldsymbol{b}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{w}^\top & \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{pmatrix} \frac{1}{M} \begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top \mathbf{1}_M & \mathbf{X}^\top \mathbf{y} \\ \mathbf{1}_M^\top \mathbf{X} & M & \mathbf{1}_M^\top \mathbf{y} \\ \mathbf{y}^\top \mathbf{X} & \mathbf{y}^\top \mathbf{1}_M & \mathbf{y}^\top \mathbf{y} \end{pmatrix} \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{b} \\ \boldsymbol{q}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{q}_2^\top & \boldsymbol{b}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}. \end{aligned}$$

Following the Assumption 3.1, we use $\widetilde{\boldsymbol{\beta}}$ to refer to the task parameter, then

$$\widetilde{\boldsymbol{\beta}} := \boldsymbol{\beta}^* + \boldsymbol{\Psi}^{\frac{1}{2}} \widetilde{\boldsymbol{\theta}}, \quad \text{where} \quad \widetilde{\boldsymbol{\theta}} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d).$$

We then decompose $f(\mathbf{E})$ as

$$f(\mathbf{E}) = \underbrace{\begin{pmatrix} \boldsymbol{w}^\top & \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{pmatrix} \frac{1}{M} \begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top \mathbf{1}_M & \mathbf{X}^\top \mathbf{y} \\ \mathbf{1}_M^\top \mathbf{X} & M & \mathbf{1}_M^\top \mathbf{y} \\ \mathbf{y}^\top \mathbf{X} & \mathbf{y}^\top \mathbf{1}_M & \mathbf{y}^\top \mathbf{y} \end{pmatrix} \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{q}_1^\top \\ \boldsymbol{q}_2^\top \end{pmatrix}}_{\mathrm{I}} \cdot \mathbf{x} + \mathrm{II},$$

where terms I and II are independent of $\mathbf{x}$. Therefore, we have

$$\begin{aligned} \mathcal{R}(f) &:= \mathbb{E}\left(f(\mathbf{E}) - y\right)^2 \\ &= \mathbb{E}\left(f(\mathbf{E}) - \langle \widetilde{\boldsymbol{\beta}}, \mathbf{x} \rangle\right)^2 + \sigma^2 && \text{since } y | \widetilde{\boldsymbol{\beta}}, \mathbf{x} \sim \mathcal{N}(0, \sigma^2) \\ &= \mathbb{E}\left(\mathrm{I} \cdot \mathbf{x} + \mathrm{II} - \langle \widetilde{\boldsymbol{\beta}}, \mathbf{x} \rangle\right)^2 + \sigma^2 && \text{since } f(\mathbf{E}) = \mathrm{I} \cdot \mathbf{x} + \mathrm{II} \\ &= \mathbb{E}\| \mathrm{I}^\top - \widetilde{\boldsymbol{\beta}} \|_{\mathbf{H}}^2 + \mathbb{E}\mathrm{II}^2 + \sigma^2 && \text{since } \mathbf{x} \sim \mathcal{N}(0, \mathbf{H}) \end{aligned}$$

$$\geq \mathbb{E}\|\mathrm{I}^\top - \widetilde{\boldsymbol{\beta}}\|_{\mathbf{H}}^2 + \sigma^2.$$

Note that the above equation holds if $\mathrm{II} = 0$, which can be easily achieved by setting $\boldsymbol{b} = \mathbf{0}_d, \boldsymbol{b}_1 = \boldsymbol{b}_2 = 0$. Therefore, without loss of generality, we can consider the LSA function which takes the following form:

$$f(\mathbf{E}) = \mathrm{I} \cdot \mathbf{x} = \begin{pmatrix} \boldsymbol{w}^\top & \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{pmatrix} \frac{1}{M} \begin{pmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{X}^\top \mathbf{1}_M & \mathbf{X}^\top \mathbf{y} \\ \mathbf{1}_M^\top \mathbf{X} & M & \mathbf{1}_M^\top \mathbf{y} \\ \mathbf{y}^\top \mathbf{X} & \mathbf{y}^\top \mathbf{1}_M & \mathbf{y}^\top \mathbf{y} \end{pmatrix} \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{q}_1^\top \\ \boldsymbol{q}_2^\top \end{pmatrix} \cdot \mathbf{x}. \tag{J.2}$$

Let's then try to determine whether such a function can effectively represent a GD-$\boldsymbol{\beta}$ function, which takes the form of

$$f_{\mathsf{GD}\text{-}\boldsymbol{\beta}}(\mathbf{E}) = \langle \boldsymbol{\beta}^*, \mathbf{x} \rangle - \left\langle \frac{\boldsymbol{\Gamma}^* \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y})}{M}, \mathbf{x} \right\rangle, \tag{J.3}$$

where $\boldsymbol{\beta}^*$ is the prior mean in Assumption 3.1 in our submission, and $\boldsymbol{\Gamma}^* = \boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Omega}^{-1}\mathbf{H}^{\frac{1}{2}}$ and $\boldsymbol{\Omega} = \frac{M+1}{M}\mathbf{H}^{\frac{1}{2}}\boldsymbol{\Psi}\mathbf{H}^{\frac{1}{2}} + \frac{\sigma^2 + \mathrm{tr}(\mathbf{H}\boldsymbol{\Psi})}{M}\mathbf{I}_d$ is defined in (5.3) in our submission. In order to achieve this, one simple way is to let

$$\boldsymbol{w} = -\boldsymbol{\beta}^*, \quad \boldsymbol{a}_1 = \boldsymbol{a}_2 = 1, \quad \boldsymbol{Q} = (\boldsymbol{\Gamma}^*)^\top, \quad \boldsymbol{q}_1 = \boldsymbol{\beta}^*, \quad \boldsymbol{q}_2 = \mathbf{0}_d. \tag{J.4}$$

However, this will incur some additive terms and will potentially enlarge the ICL risk. Inserting the parameters above, we have

$$f(\mathbf{E}) = f_{\mathsf{GD}\text{-}\boldsymbol{\beta}}(\mathbf{E}) + \frac{\mathbf{1}_M^\top}{M}\left(\mathbf{X}(\boldsymbol{\Gamma}^*)^\top - \mathbf{x}\boldsymbol{\beta}^*(\boldsymbol{\beta}^*)^\top + \mathbf{y}(\boldsymbol{\beta}^*)^\top\right) \cdot \mathbf{x}$$

This shows that the extended token with a single LSA cannot easily implement the GD-$\boldsymbol{\beta}$ function class and may incur an additive ICL risk depending on $\boldsymbol{\beta}^*$ (as shown in Theorem 4.1).

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: We believe our introduction and abstract are factually accurate in describing the contributions of the paper and of its result.

   Guidelines:

   - The answer NA 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 NA 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: We discuss the limitation in Section 4 and the future work section.

   Guidelines:

   - The answer NA 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 provide the assumption in Assumption 3.1.

Guidelines:

- The answer NA 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: We provide all detailed experiment setup in Appendix I.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The data we use are simulated. We plan to release our code upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include all details and hyperparameters in our experiments in Appendix I.

Guidelines:

- The answer NA 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: [No]

Justification: The cost for producing the experiments precludes us from reporting error bars. Moreover, the experiment section is not the main contribution of this paper.

Guidelines:

- The answer NA 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: [NA]

   Justification: There is no strict requirement for the computer resources for our experiments.

   Guidelines:

   - The answer NA 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: None to report.

   Guidelines:

   - The answer NA 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: [NA]

    Justification: This work focuses on the theory of in-context learning so it is not expected to have a direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The data used in this paper are fully simulated, so there is no data or model that have a high risk.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: We do not relase any asset.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: We do not relase any asset.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: We do not involve humans in our research.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: Our paper does not involve crowdsourcing.

    Guidelines:

    - The answer NA 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.