WRAPPED GAUSSIAN MIXTURE ON THE LOW-RANK MATRIX MANIFOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank matrices are widely used in parameter-efficient machine learning models, such as LoRA-based LLM fine-tuning and latent-factor recommender systems. To characterize their generalizability, we commonly use PAC-Bayes bounds, and choose Gaussian priors and posteriors to make the KL divergence practically computable. However, we identify a fundamental issue: when assigning Gaussian distributions on rank-k matrices, no Gaussian can have its support fully cover $\mathcal{M}_k$ – the set of all $m \times n$ matrices of rank-k, also a smooth embedded manifold of dimension $(m+n-k)k$. This partial covering can make the posterior not absolutely continuous with respect to the prior, resulting in infinite KL divergence. To address this issue, we first propose a new statistical tool for characterizing low-rank matrices – a wrapped Gaussian mixture that covers $\mathcal{M}_k$, constructed based on differential-geometric analysis. We then show that this mixture admits a practically computable KL divergence bound in closed-form. Finally, we apply the mixture to establish a PAC-Bayes bound for stochastic LoRA neural networks. Empirical results show that the bound is nonvacuous and captures LoRA generalization behavior, suggesting the utility of the mixture for analyzing low-rank model performance.

1 INTRODUCTION

Low-rank models have been playing a crucial role in modern machine learning areas such as LoRA-based LLM fine-tuning (Hu et al., 2022; Meng et al., 2024; Zhu et al., 2024), latent-factor recommender systems (Koren et al., 2009; Steck, 2020; Nikolakopoulos et al., 2021) and neural network compression (Idelbayev & Carreira-Perpinán, 2020). These models use low-rank weight matrices, typically represented in the form $W = PQ \in \mathbb{R}^{m\times n}$, where $P \in \mathbb{R}^{m\times k}$, $Q \in \mathbb{R}^{k\times n}$, and $k < \min\{m, n\}$. The low-rank structure allows a model embedded in a high-dimensional parameter space to retain expressiveness while using relatively few parameters, thereby improving memory and computational efficiency (Hu et al., 2022).

One typical way to evaluate how well these models perform is through statistical learning theory, where a PAC-Bayes bound is commonly used to characterize their true performance (McAllester, 1998; Alquier, 2024). This approach assigns a prior distribution π and a posterior distribution ρ on W, where π represents the initialization and ρ represents the trained model, and the bound is determined by the Kullback-Leibler(KL) divergence $KL(\rho \,||\, \pi)$ (Kullback & Leibler, 1951). When both ρ and π are chosen to be Gaussian, $KL(\rho \,||\, \pi)$ admits a closed-form expression, thereby making the bound practically computable (Dziugaite & Roy, 2017; Guo et al., 2026).

However, we identify a fundamental challenge in defining and computing KL divergence for low-rank matrices. Formally, we denote $\mathcal{M}_k$ as the set of all $m \times n$ real matrices of rank k, i.e.,

$$\mathcal{M}_k = \{X \in \mathbb{R}^{m\times n} : \operatorname{rank}(X) = k\} \quad \text{for } 0 < k < \min\{m, n\}$$

which is known to be a smooth manifold of dimension $(m+n-k)k$ embedded in $\mathbb{R}^{m\times n}$ (Uschmajew & Vandereycken, 2020). Our analysis shows that it is possible for a Gaussian distribution on $\mathbb{R}^{m\times n}$ to cover a subset of $\mathcal{M}_k$ but impossible for it to cover the entire $\mathcal{M}_k$. (Section 2.2). Since the coverage is not full, requiring both ρ and π to lie on $\mathcal{M}_k$ while forcing them to be Gaussian can cause ρ not absolutely continuous with π, making the KL divergence infinite (Section 2.4).

We address this challenge by constructing a *wrapped Gaussian mixture* that covers the entire $\mathcal{M}_k$: being wrapped allows its support to be a nonlinear manifold that matches the geometric shape of $\mathcal{M}_k$,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

being Gaussian allows us to obtain the KL divergence in closed form to facilitate computation, and being a mixture ensures that the distribution covers the entire $\mathcal{M}_k$, avoiding infinite KL divergence.

The details of our contributions are as follows:

- Based on the differential-geometric analysis of (Uschmajew & Vandereycken, 2020) (Proposition 2.1), we construct finitely many charts that cover $\mathcal{M}_k$. In each chart, we define a Gaussian distribution in Euclidean space and push it forward to $\mathcal{M}_k$ through a diffeomorphism, yielding a wrapped Gaussian covering a subset of $\mathcal{M}_k$. All these wrapped Gaussians form a mixture that covers the entire $\mathcal{M}_k$, denoted as X [1]. (Section 3)
- To estimate the KL divergence of two mixtures $KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}})$, we utilize the upper bound $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) \geq KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}})$, where XC is a joint distribution of X with a discrete distribution over charts C. We show that $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ is finite and has a closed-form expression, and that it can be computed practically by reducing the total number of parameters through parameter sharing across charts. (Section 4)
- Using the mixture X, a classic LoRA model $W = W_0 + \Delta W$ can be naturally converted into a stochastic LoRA model $\mathsf{W} = W_0 + \mathsf{X}$, for which we establish a PAC-Bayes bound by defining a prior g_{X} and a posterior f_{X}, and compute the KL divergence using $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$. Empirical results show that the bound is nonvacuous and captures the generalization behavior of LoRA, suggesting the utility of the mixture for analyzing low-rank model performance. (Section 5)

2 Low-Rank Matrix Manifolds and Challenges in Distribution Covering

We first review the differential-geometric structure of $\mathcal{M}_k$ in Section 2.1, then show that earlier approaches, such as Gaussian distributions (Section 2.2) and single-chart wrapped Gaussians (Section 2.3), cannot cover $\mathcal{M}_k$. In Section 2.4, we demonstrate a critical issue: this partial covering can make the posterior not absolutely continuous with respect to the prior, which poses a fundamental obstacle for computing KL divergence on low-rank matrix manifolds.

2.1 The manifold of $\mathcal{M}_k$

Since $\mathcal{M}_k$ is a subset of the Euclidean space $\mathbb{R}^{m\times n}$, we let $\Omega_R = \{U \subset \mathbb{R}^{m\times n} : U \text{ is open}\}$ be the topology (i.e., the collection of all open subsets) of $\mathbb{R}^{m\times n}$, $\Omega = \{\mathcal{M}_k \cap U : U \in \Omega_R\}$ be the subspace topology of $\mathcal{M}_k$, and denote $(\mathcal{M}_k, \Omega)$ as a topological subspace of $(\mathbb{R}^{m\times n}, \Omega_R)$.

Let $\mathcal{I}_k = \{(i_1, i_2, ..., i_k) : 1 \leq i_1 \leq i_2 ... \leq m\}$ be the set of the combination of k rows from m, and $\mathcal{J}_k = \{(j_1, j_2, ..., j_k) : 1 \leq j_1 \leq j_2 ... \leq n\}$ be the set of the combination of k columns from n. We have $|\mathcal{I}_k| = \binom{m}{k}$ and $|\mathcal{J}_k| = \binom{n}{k}$. Let $X \in \mathbb{R}^{m\times n}$, given $I \in \mathcal{I}_k$ and $J \in \mathcal{J}_k$, we denote X_{IJ} as the $k \times k$ submatrix of X with row indices in I and column indices in J. For any I, J, we define the set $\mathcal{U}_{IJ} = \{X \in \mathbb{R}^{m\times n} : X_{IJ} \text{ invertible}\}$. Then $\mathcal{U}_{IJ}$ is an open subset of $\mathbb{R}^{m\times n}$ (Proposition B.1).

Since for any rank-k matrix, there must be some $k \times k$ submatrices invertible (0.4.4, (Horn & Johnson, 2012)), we can write $\mathcal{M}_k = \bigcup_{I\in\mathcal{I}_k, J\in\mathcal{J}_k} (\mathcal{M}_k \cap \mathcal{U}_{IJ})$. Since $\mathcal{M}_k \cap \mathcal{U}_{IJ} \in \Omega$, each $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ is an open set on $\mathcal{M}_k$ with respect to Ω. This implies that the collection of sets $\{\mathcal{M}_k \cap \mathcal{U}_{IJ}\}_{I\in\mathcal{I}_k, J\in\mathcal{J}_k}$ is an open cover of $\mathcal{M}_k$.

By considering $\mathcal{M}_k$ to be a smooth manifold of dimension d, we mean that for any $p \in \mathcal{M}_k$, we can find an open subset $N_p \in \Omega$ with $p \in N_p$ (called a neighborhood of p), an open subset $\widehat{N}_p \subset \mathbb{R}^d$, and a diffeomorphism (a bijective infinitely differentiable map with infinitely differentiable inverse) $\psi : N_p \to \widehat{N}_p$. For any open subset $U \subset \mathcal{M}_k$ with a diffeomorphism ψ such that $\psi(U) \subset \mathbb{R}^d$, we say (U, ψ) is a chart. A known fundamental result is that

Proposition 2.1. *((Uschmajew & Vandereycken, 2020; Vandereycken, 2013)) $\mathcal{M}_k$ is a smooth submanifold of dimension $(m + n - k)k$ embedded in $\mathbb{R}^{m\times n}$.*

Proof: Since any $X \in \mathcal{M}_k$ contains a $k \times k$ invertible submatrix X_{IJ} for some I, J, it must be contained in $\mathcal{U}_{IJ}$ for some I, J. Since $\mathcal{U}_{IJ}$ is an open subset of $\mathbb{R}^{m\times n}$, it is a neighborhood of X in

[1] In this paper, all symbols in mathsf font such as A, B represent random variables.

$\mathbb{R}^{m\times n}$. By Proposition 5.16 of (Lee, 2012), if for every I, J, there exists a smooth submersion $\phi : \mathcal{U}_{IJ} \to \mathbb{R}^{(m-k)\times(n-k)}$ such that $\mathcal{M}_k \cap \mathcal{U}_{IJ} = \phi^{-1}(C)$ for some constant matrix $C \in \mathbb{R}^{(m-k)\times(n-k)}$, then $\mathcal{M}_k$ is an embedded smooth manifold of dimension $mn-(m-k)(n-k) = (m+n-k)k$. To find such a ϕ, given any I, J, for any $X \in \mathcal{U}_{IJ}$, we first permute its rows and columns to obtain X', so that X_{IJ} appears in the upper-left corner: $X' = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, where $A = X_{IJ} \in \mathbb{R}^{k\times k}$ is invertible. By Schur complement, we have $\operatorname{rank}(X) = \operatorname{rank}(A) + \operatorname{rank}(D - CA^{-1}B)$. Therefore,

$$X \in \mathcal{M}_k \cap \mathcal{U}_{IJ} \subset \mathcal{U}_{IJ} \iff \operatorname{rank}(X) = \operatorname{rank}(A) = k \iff D - CA^{-1}B = 0$$

Hence, we can define $\phi : \mathcal{U}_{IJ} \mapsto \mathbb{R}^{(m-k)\times(n-k)}$ by $\phi(X) = D - CA^{-1}B$, where ϕ is a submersion since its partial derivative with respect to D is identity, and we have $\mathcal{M}_k \cap \mathcal{U}_{IJ} = \phi^{-1}(0)$. □

2.2 Gaussian Distribution cannot Cover $\mathcal{M}_k$

The simplest distribution that could be assigned on $\mathcal{M}_k$ may be a Gaussian distribution. We show that any $m \times n$ Gaussian random matrix, with dependent or independent entries, cannot cover $\mathcal{M}_k$.

Let $\mathcal{P}$ be a probability measure (distribution) on $\mathbb{R}^{m\times n}$, $\mathsf{W} \sim \mathcal{P}$ be a real random matrix, f_{W} be the probability density of W, and $\operatorname{supp}(\mathcal{P}) = \{W : f_{\mathsf{W}}(W) > 0\}$ be the support of $\mathcal{P}$ such that $\mathcal{P}(\operatorname{supp}(\mathcal{P})) = 1$.

Definition 2.2. We say $\mathcal{P}$ *covers* $\mathcal{M}_k$ if $\mathcal{M}_k \subset \operatorname{supp}(\mathcal{P})$ and $\mathcal{P}(\mathcal{M}_k) = 1$.

This definition includes two cases: $\operatorname{supp}(\mathcal{P}) = \mathcal{M}_k$, or $\operatorname{supp}(\mathcal{P}) \setminus \mathcal{M}_k$ is nonempty but has measure zero. It is possible to construct a $\mathcal{P}$ covering $\mathcal{M}_k$: define two random matrices $\mathsf{P} \in \mathbb{R}^{m\times k}$ and $\mathsf{Q} \in \mathbb{R}^{k\times n}$ with independent Gaussian entries, and let $\mathsf{W} = \mathsf{PQ}$. But this $\mathcal{P}$ is not Gaussian – it is a dependent non-central chi-squared distribution that is irregular and difficult to analyze. To simplify the analysis, we focus on the case where $\mathcal{P}$ is Gaussian and denote it as $\mathcal{Q}$. The following example shows a possible construction of $\mathcal{Q}$ that $\mathsf{W} \sim \mathcal{Q}$ is of rank-k with probability 1.

Example 2.3. Given a fixed full-rank matrix $P \in \mathbb{R}^{m\times k}$ and a Gaussian matrix $\mathsf{Q} \in \mathbb{R}^{k\times n}$ whose entries are independent, let $\mathsf{W} = P\mathsf{Q}$. Then W is a Gaussian matrix, with each column $\mathsf{W}_{*i} = P\mathsf{Q}_{*i}$ for $i = 1, ..., n$ being a linear transformation of the Gaussian vector Q_{*i}. In this case,

(a) Since all entries in Q are independent, Q is of full rank with probability 1 (Proposition B.5). Since any realization $\mathsf{Q} = Q$ gives $\operatorname{rank}(PQ) = k$, W is of rank-k with probability 1.

(b) The support of W, denoted as $\operatorname{supp}(\mathcal{Q})$, is a kn dimensional submanifold: since $\operatorname{vec}(\mathsf{W}) = (I_n \otimes P)\operatorname{vec}(\mathsf{Q})$, the support of $\operatorname{vec}(\mathsf{W})$ is a linear subspace spanned by the columns of $I_n \otimes P$, where $\operatorname{rank}(I_n \otimes P) = kn$.

(c) $\mathcal{Q}$ covers only a subset of $\mathcal{M}_k$, i.e., $\mathcal{Q}(\operatorname{supp}(\mathcal{Q}) \cap \mathcal{M}_k) = 1$, but $\mathcal{M}_k \not\subset \operatorname{supp}(\mathcal{Q})$. As $\operatorname{supp}(\mathcal{Q}) = \operatorname{range}(I_n \otimes P)$, choosing a different P will make $\mathcal{Q}$ cover a different subset of $\mathcal{M}_k$.

(d) The Gaussian entries in each W_{*i} are dependent. In fact, for any $\mathcal{Q}$ with $\operatorname{supp}(\mathcal{Q})$ being a lower dimensional subspace of $\mathbb{R}^{m\times n}$, the entries of W must be dependent (Proposition B.5).

Example 2.3 shows the existence of a Gaussian $\mathcal{Q}$ on $\mathbb{R}^{m\times n}$ that covers a subset of $\mathcal{M}_k$. Since $\mathcal{M}_k$ is a smooth submanifold with $\dim(\mathcal{M}_k) = (m+n-k)k$ (Proposition 2.1), while $\operatorname{supp}(\mathcal{Q})$ is an affine subspace (linear submanifold) whose dimension can vary in $\{0, 1, ..., mn\}$ (Proposition B.6), a nature question arises: Can we construct a Gaussian $\mathcal{Q}$ on $\mathbb{R}^{m\times n}$ that covers the entire $\mathcal{M}_k$, i.e., $\mathcal{M}_k \subset \operatorname{supp}(\mathcal{Q})$ and $\mathcal{Q}(\mathcal{M}_k) = 1$?

We show that this is **impossible**, based on the following fact:

Fact 2.4. For any Gaussian distribution $\mathcal{Q}$ on $\mathbb{R}^{m\times n}$,

(a) If $\dim(\operatorname{supp}(\mathcal{Q})) < \dim(\mathcal{M}_k)$, then $\mathcal{M}_k \not\subset \operatorname{supp}(\mathcal{Q})$, since a higher dimensional manifold cannot be a subset of a lower dimensional space.

(b) If $\dim(\operatorname{supp}(\mathcal{Q})) > \dim(\mathcal{M}_k)$ and $\mathcal{M}_k \subset \operatorname{supp}(\mathcal{Q})$, then $\mathcal{Q}(\mathcal{M}_k) = 0$, since a lower dimensional manifold has measure zero in a higher dimensional space (Lemma B.2).

(c) If $\dim(\operatorname{supp}(\mathcal{Q})) = \dim(\mathcal{M}_k)$, then $\mathcal{M}_k \not\subset \operatorname{supp}(\mathcal{Q})$, since there always exists a nonlinear manifold $\mathcal{M}_k \cap \mathcal{U}_{IJ} \subset \mathcal{M}_k$ with $\dim(\mathcal{M}_k \cap \mathcal{U}_{IJ}) = \dim(\mathcal{M}_k)$ such that $\mathcal{M}_k \cap \mathcal{U}_{IJ} \not\subset \operatorname{supp}(\mathcal{Q})$ for any linear manifold $\operatorname{supp}(\mathcal{Q})$ (Theorem B.7).

Hence, there is no Gaussian distribution $\mathcal{Q}$ on $\mathbb{R}^{m\times n}$ that covers $\mathcal{M}_k$.

2.3 Challenges in Covering $\mathcal{M}_k$ using Wrapped Gaussian

Since directly assigning a Gaussian on $\mathcal{M}_k$ does not guarantee the covering, another approach is to first define it in an Euclidean space, then push forward it to $\mathcal{M}_k$, yielding a *wrapped Gaussian*.

This approach has been applied to specific manifolds such as hyperbolic spaces (Nagano et al., 2019) and the manifold of all $m \times m$ symmetric positive definite matrices (De Surrel et al., 2025). In these works, a manifold $\mathcal{R}$ is obtained from the Euclidean space $\mathbb{R}^d$ by replacing its Euclidean metric with another Riemannian metric, such that $\mathcal{R}$ is *globally* diffeomorphic to $\mathbb{R}^d$, i.e., there exists a diffeomorphism E satisfying $E(\mathcal{R}) = \mathbb{R}^d$. Then, we can define a Gaussian distribution $\mathcal{Q}$ on $\mathbb{R}^d$ and push it forward to $\mathcal{R}$ through E^{-1}, obtaining the wrapped Gaussian $E^{-1}(\mathcal{Q})$. Suppose $\mathsf{X} \sim E^{-1}(\mathcal{Q})$ and $\mathsf{Y} \sim \mathcal{Q}$, and let $f_{\mathsf{X}}(X)$ and $f_{\mathsf{Y}}(Y)$ be their densities respectively. By change of variables,

$$f_{\mathsf{X}}(X) = f_{\mathsf{Y}}(E(X)) \left| \det \left(\frac{dE(z)}{dz}\bigg|_{z=X} \right) \right| \tag{1}$$

where $\frac{dE(z)}{dz} \in \mathbb{R}^{d\times d}$ is the Jacobian matrix of E. In this case, $\mathrm{supp}(\mathsf{X}) = \mathcal{R}$.

In the above cases, the obtained $\mathcal{R}$ is specific since it can be covered by one chart $(\mathcal{R}, E)$. However, general smooth manifolds like $\mathcal{M}_k$ may not be covered by one chart, i.e., there may not exist a diffeomorphism E such that $\mathcal{M}_k = E^{-1}(\mathbb{R}^{(m+n-k)k})$. This brings a main challenge: we may not be able to define a wrapped Gaussian in a single chart to cover $\mathcal{M}_k$.

Since a single chart may not work, our solution is as follows: use multiple charts to cover $\mathcal{M}_k$, then construct a wrapped Gaussian in each chart; by collecting wrapped Gaussians in all charts, we can obtain a mixture distribution that covers $\mathcal{M}_k$. The issue remained is how to obtain a valid multiple-chart covering for $\mathcal{M}_k$. We will illustrate this in Section 3.

2.4 The Consequence for not Covering $\mathcal{M}_k$

We show that distributions whose support are unable to cover the entire $\mathcal{M}_k$, such as the Gaussian in Section 2.2 and the single-chart wrapped Gaussian in Section 2.3, can cause serious problem in computing KL divergence – the KL divergence may go infinity.

In classic LoRA fine-tuning, we take $W = PQ$ where $P \in \mathbb{R}^{m\times k}$ and $Q \in \mathbb{R}^{k\times n}$ can be arbitrary, so W can be any element in $\mathcal{M}_k$. Since the LoRA weights are typically randomly initialized and stochastically optimized, we can assign a distribution over W. Let W_0 and W_1 represent the weight matrices before and after training. Following the PAC-Bayes analysis by (Dziugaite & Roy, 2017), let $\mathcal{T}$ be a family of distributions, $\pi \in \mathcal{T}$ be the prior such that $W_0 \in \mathrm{supp}(\pi)$, and $\rho \in \mathcal{T}$ be the posterior such that $W_1 \in \mathrm{supp}(\rho)$. Then, the KL divergence of π relative to ρ is defined as

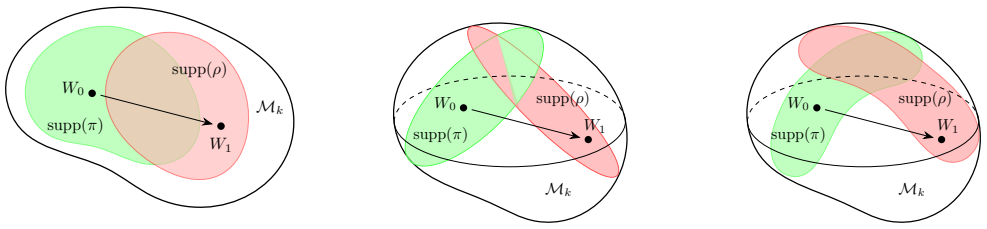
$$KL(\rho \,||\, \pi) = \int_{\mathrm{supp}(\rho)} f_\rho(W) \log \frac{f_\rho(W)}{f_\pi(W)}\, dW \tag{2}$$

where f_π and f_ρ are the densities of π and ρ, respectively. We say ρ is absolutely continuous with π if $\mathrm{supp}(\rho) \subset \mathrm{supp}(\pi)$, denoted as $\rho \ll \pi$. If not $\rho \ll \pi$, then $KL(\rho \,||\, \pi) = \infty$.

However, if both $\mathrm{supp}(\pi)$ and $\mathrm{supp}(\rho)$ does not equal to $\mathcal{M}_k$, we may not have $\rho \ll \pi$. For example, if $W_0 \in \mathrm{supp}(\pi), \mathcal{M}_k \setminus \mathrm{supp}(\pi) \neq \varnothing$ and $W_1 \in \mathcal{M}_k \setminus \mathrm{supp}(\pi)$, then for any ρ satisfying $W_1 \in \mathrm{supp}(\rho)$, we have $W_1 \in \mathrm{supp}(\rho) \setminus \mathrm{supp}(\pi)$, implying that ρ is not absolutely continuous with π and $KL(\rho \,||\, \pi) = \infty$. This happens if we let π and ρ be Gaussian or single-chart wrapped Gaussian, since they cannot cover $\mathcal{M}_k$. An illustration of this issue is shown in Figure 1.

One way to address this issue is to consider a simpler problem, e.g., defining π and ρ only on a subset of $\mathcal{M}_k$. For example, some simplified LoRA variants like (Zhu et al., 2024) freeze P and train Q only. We can assign a Gaussian to Q, such that W becomes the Gaussian matrix W in Example 2.3. Any π and ρ on W will satisfy $\mathrm{supp}(\pi) = \mathrm{supp}(\rho) = \mathrm{range}(I_n \otimes P) \subsetneq \mathcal{M}_k$, where $\mathrm{range}(I_n \otimes P)$ denotes the linear subspace spanned by the columns of $I_n \otimes P$. In this case, $KL(\rho \,||\, \pi)$ is always finite. However, the simplified LoRA restricts the obtainable models in a small linear subspace within $\mathcal{M}_k$, making these models less expressive than classic LoRA models.

To address this limitation, we focus on the classic LoRA setting without any simplification, and our goal is to find a family $\mathcal{T}$ such that $\mathrm{supp}(\pi) = \mathrm{supp}(\rho) = \mathcal{M}_k$ for any $\pi, \rho \in \mathcal{T}$. In this case, $\rho \ll \pi$ always holds, which prevents $KL(\rho \,||\, \pi)$ going infinity. As we will show in Section 3, such $\mathcal{T}$ can be obtained by constructing a wrapped Gaussian mixture on $\mathcal{M}_k$.



(a) Set view (b) Gaussian π, ρ (c) Wrapped Gaussian π, ρ

Figure 1: Illustration of the issue that $\rho \ll \pi$ may not hold if π and ρ do not cover $\mathcal{M}_k$. Subfigure (a) views $\mathrm{supp}(\pi)$, $\mathrm{supp}(\rho)$ and $\mathcal{M}_k$ as sets, while subfigures (b) and (c) show their geometry.

3 Constructing a Wrapped Gaussian Mixture Covering $\mathcal{M}_k$

A *Gaussian mixture* is a distribution whose density is a weighted average of Gaussian densities. In this section, we propose a wrapped Gaussian mixture that covers $\mathcal{M}_k$ based on differential-geometric construction.

The geometric structure of $\mathcal{M}_k$ in Section 2.1 provides two important ideas: (1) $\mathcal{M}_k$ can be covered by a collection of open subsets $\{\mathcal{M}_k \cap \mathcal{U}_{IJ}\}_{I \in \mathcal{I}_k, J \in \mathcal{J}_k}$. (2) For each I, J, there is a smooth submersion $\phi : \mathcal{U}_{IJ} \to \mathbb{R}^{(m-k)\times(n-k)}$ defined by $\phi(X) = D - CA^{-1}B$, such that $\mathcal{M}_k \cap \mathcal{U}_{IJ} \subset \mathcal{U}_{IJ}$ satisfies $\phi(\mathcal{M}_k \cap \mathcal{U}_{IJ}) = 0$ (see the proof of Proposition 2.1). The first idea guarantees the covering, and the second idea allows us to construct wrapped distributions on $\mathcal{M}_k$. Based on these ideas, our construction is as follows.

Construct the Diffeomorphism. A wrapped distribution on $\mathcal{M}_k$ can be constructed through density transformation (1), which depends on a diffeomorphism (bijective). The submersion ϕ is surjective and cannot be directly used in (1). We need to convert ϕ into a diffeomorphism.

In any chart with given I and J,

(1) For any $X \in \mathcal{U}_{IJ}$, X_{IJ} is invertible. We define a map $H_{IJ} : X \mapsto \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ that shifts X_{IJ} to the upper-left corner by permuting rows and columns (this permutation depends on I, J), such that $A = X_{IJ}$. This map is a diffeomorphism since it is a permutation of the coordinates of $\mathbb{R}^{m\times n}$.

(2) Since $\begin{bmatrix} A & B \\ C & D \end{bmatrix} \mapsto D - CA^{-1}B$ is a smooth submersion, we define a map $G : \begin{bmatrix} A & B \\ C & D \end{bmatrix} \mapsto \begin{bmatrix} A & B \\ C & D - CA^{-1}B \end{bmatrix}$, then G is a diffeomorphism. The Jacobians of both G and G^{-1} are triangular $mn \times mn$ matrices with all diagonal entries equal to 1, so their determinants are 1.

(3) Since both H_{IJ} and G are diffeomorphisms, $G \circ H_{IJ} : X \mapsto \begin{bmatrix} A & B \\ C & D - CA^{-1}B \end{bmatrix}$ for all $X \in \mathcal{U}_{IJ}$ is a diffeomorphism. Moreover, for any $X \in \mathcal{M}_k \cap \mathcal{U}_{IJ} \subset \mathcal{U}_{IJ}$, $D - CA^{-1}B = 0$. Let

$$\mathcal{S} = \left\{ \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} : A \in \mathbb{R}^{k\times k} \text{ invertible}, B \in \mathbb{R}^{k\times(n-k)}, C \in \mathbb{R}^{(m-k)\times k} \right\} \subset \mathbb{R}^{m\times n}$$

then $\mathcal{M}_k \cap \mathcal{U}_{IJ} = (G \circ H_{IJ})^{-1}(\mathcal{S})$. Since $\mathcal{S}$ lies in an $(m+n-k)k$ dimensional Euclidean subspace, $(\mathcal{M}_k \cap \mathcal{U}_{IJ}, G \circ H_{IJ})$ is one chart of $\mathcal{M}_k$.

Construct the Gaussian in a Single Chart. We first define a conditional Gaussian distribution $\mathcal{Q}$ on $\mathbb{R}^{m\times n}$. Let $\mathsf{Y} = \begin{bmatrix}\mathsf{A} & \mathsf{B}\\ \mathsf{C} & \mathsf{0}\end{bmatrix} \sim \mathcal{Q}$, where every entry in $\mathsf{A}, \mathsf{B}, \mathsf{C}$ is an independent Gaussian random variable, and every entry in $\mathsf{0}$ equals to 0 with probability 1. Since A is a $k \times k$ Gaussian matrix with independent entries, by Corollary B.4, A is invertible with probability 1. This implies that the set $\{\begin{bmatrix}A & B\\ C & 0\end{bmatrix} : A \text{ not invertible}\}$ has measure zero in $\{\begin{bmatrix}A & B\\ C & 0\end{bmatrix} : \text{any } A\}$. Hence, $\mathcal{Q}(\mathcal{S}) = 1$, and $\mathcal{Q}(\mathbb{R}^{m\times n} \setminus \mathcal{S}) = 0$.

Now, we push forward $\mathcal{Q}$ from $\mathcal{S}$ to $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ via the diffeomorphism $\mathcal{M}_k \cap \mathcal{U}_{IJ} = (G \circ H_{IJ})^{-1}(\mathcal{S})$. Recall that the determinant of the Jacobian of G is 1, and the Jacobian of H_{IJ} is a permutation matrix, whose determinant has absolute value 1. Let $(G \circ H_{IJ})^{-1}(\mathcal{Q})$ be the distribution on $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ push forwarded from $\mathcal{Q}$, and let $\mathsf{X} \sim (G \circ H_{IJ})^{-1}(\mathcal{Q})$. By (1), we have

$$f_{\mathsf{X}}(X)/f_{\mathsf{Y}}((G \circ H_{IJ})(X)) = \left|\det\left(\frac{d(G \circ H_{IJ})(z)}{dz}\bigg|_{z=X}\right)\right| = \left|\det\left(\frac{dG(r)}{dr}\bigg|_{r=H_{IJ}(X)}\right)\right| \left|\det\left(\frac{dH_{IJ}(z)}{dz}\bigg|_{z=X}\right)\right| = 1 \quad (3)$$

(3) implies that f_{X} is a density of wrapped Gaussian *transferred* from the Gaussian density f_{Y}. Note that their ratio being 1 does not imply that X is Gaussian: $\mathsf{X} = (G \circ H_{IJ})^{-1}(\begin{bmatrix}\mathsf{A} & \mathsf{B}\\ \mathsf{C} & \mathsf{0}\end{bmatrix}) = H_{IJ}^{-1}(\begin{bmatrix}\mathsf{A} & \mathsf{B}\\ \mathsf{C} & \mathsf{CA}^{-1}\mathsf{B}\end{bmatrix})$, where $\mathsf{CA}^{-1}\mathsf{B}$ is not a Gaussian matrix. The goal of (3) is to establish a connection that allow us to compute f_{X} using f_{Y}, and, as we will show in Section 4, f_{Y} being Gaussian allows us to obtain a KL divergence bound for f_{X} in *closed form*.

Construct the Gaussian Mixture with Multiple Charts. Recall that $\{\mathcal{M}_k \cap \mathcal{U}_{IJ}\}_{I\in\mathcal{I}_k, J\in\mathcal{J}_k}$ covers $\mathcal{M}_k$, each $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ corresponds to one chart, and the total number of charts is $\binom{m}{k}\binom{n}{k}$.

We first assign a discrete distribution over the charts. Let $\mathcal{V} = \{1, 2, ..., \binom{m}{k}\binom{n}{k}\}$ be the index set of charts, and let C be a random variable on $\mathcal{V}$ such that $P(\mathsf{C} = i) = \alpha_i$, $\alpha_i \geq 0$ for all i and $\sum_{i=1}^{|\mathcal{V}|} \alpha_i = 1$. For simplicity, we write H_{IJ} as H_i and $\mathcal{U}_{IJ}$ as $\mathcal{U}_i$ if I, J corresponds to the ith chart.

Let X and Y in each chart be conditioned on C, and denote them by $\mathsf{X}|\mathsf{C}$ and $\mathsf{Y}|\mathsf{C}$, respectively. For any $\mathsf{C} = i$, $f_{\mathsf{Y}|\mathsf{C}}$ is a Gaussian density, and $f_{\mathsf{X}|\mathsf{C}}$ is a wrapped Gaussian density transferred from $f_{\mathsf{Y}|\mathsf{C}}$ through (3).

Let XC be the joint distribution of $\mathsf{X}|\mathsf{C}$ and C, then the support of XC lies in $\mathcal{M}_k \times \mathcal{V}$, and the density of XC is

$$f_{\mathsf{XC}}(X, i) = f_{\mathsf{X}|\mathsf{C}}(X|i)P(\mathsf{C} = i) = f_{\mathsf{Y}|\mathsf{C}}((G \circ H_i)(X)|i)P(\mathsf{C} = i) \quad (4)$$

for any $i \in \mathcal{V}$ and $X|i \in \mathcal{M}_k \cap \mathcal{U}_i$. Consequently, let X be the marginal distribution of XC, then the density of X is

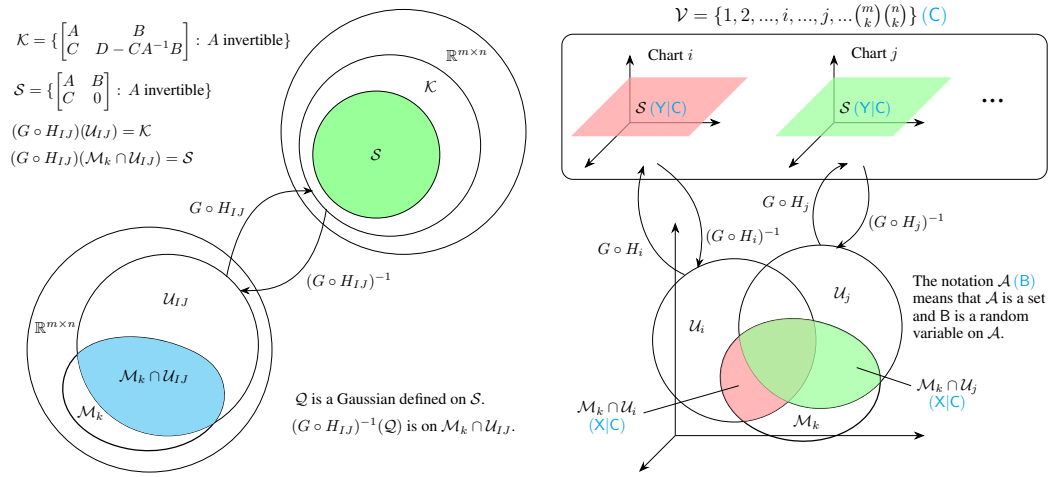
$$f_{\mathsf{X}}(X) = \sum_{i=1}^{|\mathcal{V}|} f_{\mathsf{XC}}(X, i) = \sum_{i=1}^{|\mathcal{V}|} f_{\mathsf{X}|\mathsf{C}}(X|i)P(\mathsf{C} = i) = \sum_{i=1}^{|\mathcal{V}|} f_{\mathsf{Y}|\mathsf{C}}((G \circ H_i)(X)|i)P(\mathsf{C} = i) \quad (5)$$

for any $X \in \mathcal{M}_k$. In (5), $\sum f_{\mathsf{Y}|\mathsf{C}}$ is a weighted sum of Gaussian densities, which is a Gaussian mixture on $\mathbb{R}^{m\times n}$; $f_{\mathsf{X}} = \sum f_{\mathsf{X}|\mathsf{C}}$ is a mixture of wrapped Gaussian densities on $\mathcal{M}_k$, transferred from $\sum f_{\mathsf{Y}|\mathsf{C}}$ by (3). Hence, we call the distribution of X a *wrapped Gaussian mixture*. It is easy to check that f_{X} is a valid density function since $\int_{\mathcal{M}_k} f_{\mathsf{X}}(X)\, dX = 1$. Moreover, for any $X \in \mathcal{M}_k$, there must exist some t such that $X \in \mathcal{M}_k \cap \mathcal{U}_t$. If taking $P(\mathsf{C} = i) > 0$ for all i, i.e., each chart is chosen with a positive probability, then by (5),

$$f_{\mathsf{X}}(X) \geq f_{\mathsf{Y}|\mathsf{C}}((G \circ H_t)(X)|t)P(\mathsf{C} = t) > 0$$

This implies $f_{\mathsf{X}}(X) > 0$ for any $X \in \mathcal{M}_k$, so $\operatorname{supp}(\mathsf{X}) = \mathcal{M}_k$.

Therefore, we have obtained a wrapped Gaussian mixture X whose support covers $\mathcal{M}_k$. An illustration of the above construction is shown in Figure 2.



(a) Diffeomorphism and wrapped Gaussian constructed in a single chart. (set view)

(b) Diffeomorphisms and wrapped Gaussian mixture constructed on multiple charts. (geometric view)

Figure 2: An illustration of the construction of the wrapped Gaussian mixture.

4 Practical KL Divergence Bound for the Wrapped Gaussian Mixture

Suppose we have two different wrapped Gaussian mixtures on $\mathcal{M}_k$, represented by densities f_{X} and g_{X} respectively. Since the KL divergence of Gaussian mixtures typically has no closed-form (Hershey & Olsen, 2007; Durrieu et al., 2012), it is difficult to compute $KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}})$ directly. Our goal is to estimate it using an upper bound that has a closed-form and is practical to compute.

One way to obtain the upper bound is to use the chain rule of KL divergence (Theorem 2.5.3, (Cover, 2006)). We utilize the joint distribution XC. Let f_{XC} and g_{XC} be the densities of XC, and f_{C} and g_{C} be the discrete probability masses of C, then

$$KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) = \mathbb{E}_{f_{\mathsf{X}}} \left[KL(f_{\mathsf{C}|\mathsf{X}} \,||\, g_{\mathsf{C}|\mathsf{X}}) \right] + KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}}) \tag{6}$$

$$= \mathbb{E}_{f_{\mathsf{C}}} \left[KL(f_{\mathsf{X}|\mathsf{C}} \,||\, g_{\mathsf{X}|\mathsf{C}}) \right] + KL(f_{\mathsf{C}} \,||\, g_{\mathsf{C}}) \tag{7}$$

(6) allows us to establish an upper bound. Since $\mathbb{E}_{f_{\mathsf{X}}} \left[KL(f_{\mathsf{C}|\mathsf{X}} \,||\, g_{\mathsf{C}|\mathsf{X}}) \right] \geq 0$, we have

$$KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) \geq KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}}) \tag{8}$$

This implies that $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ serves as an upper bound of $KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}})$.

(7) allows us to obtain a closed-form expression for $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$. By the density transformation (3), on every chart $\mathsf{C} = i$, we have

$$KL(f_{\mathsf{X}|\mathsf{C}} \,||\, g_{\mathsf{X}|\mathsf{C}}) = \int_{\mathcal{M}_k \cap \mathcal{U}_i} f_{\mathsf{X}|\mathsf{C}}(X|i) \log \frac{f_{\mathsf{X}|\mathsf{C}}(X|i)}{g_{\mathsf{X}|\mathsf{C}}(X|i)} \, dX$$

$$= \int_{\mathcal{S}} f_{\mathsf{Y}|\mathsf{C}}((G \circ H_i)(X)|i) \log \frac{f_{\mathsf{Y}|\mathsf{C}}(G \circ H_i)(X)|i)}{g_{\mathsf{Y}|\mathsf{C}}(G \circ H_i)(X)|i)} \, d(G \circ H_i)(X) = KL(f_{\mathsf{Y}|\mathsf{C}} \,||\, g_{\mathsf{Y}|\mathsf{C}})$$

Since $KL(f_{\mathsf{Y}|\mathsf{C}} \,||\, g_{\mathsf{Y}|\mathsf{C}})$ is the KL divergence of two multivariate Gaussian distributions, it is finite and has a closed-form. Consequently, the term

$$\mathbb{E}_{f_{\mathsf{C}}} \left[KL(f_{\mathsf{X}|\mathsf{C}} \,||\, g_{\mathsf{X}|\mathsf{C}}) \right] = \sum_{i=1}^{|\mathcal{V}|} KL(f_{\mathsf{X}|\mathsf{C}} \,||\, g_{\mathsf{X}|\mathsf{C}})\big|_{\mathsf{C}=i} P(\mathsf{C} = i) = \sum_{i=1}^{|\mathcal{V}|} KL(f_{\mathsf{Y}|\mathsf{C}} \,||\, g_{\mathsf{Y}|\mathsf{C}})\big|_{\mathsf{C}=i} P(\mathsf{C} = i)$$

is finite and has a closed-form. By choosing $f_{\mathsf{C}} = g_{\mathsf{C}}$, we can make $KL(f_{\mathsf{C}} \,||\, g_{\mathsf{C}}) = 0$, such that $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ is finite and has a closed-form.

The closed-form of $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ allows us to obtain its exact value though direct computation, avoiding the need for Monte Carlo sampling. However, it may still be impractical to compute, since f_{XC} and g_{XC} contain too many parameters, and the computational cost is proportional to the number of parameters used to characterize f_{XC} and g_{XC}. In each chart i, computing $KL(f_{\mathsf{Y}|\mathsf{C}} \,||\, g_{\mathsf{Y}|\mathsf{C}})\big|_{\mathsf{C}=i}$ involves two multivariate Gaussians of dimension $(m+n-k)k$, whose means and variances are represented by $4(m+n-k)k$ parameters, and the total cost is $O((m+n-k)k)$. Since we have $\binom{m}{k}\binom{n}{k}$ charts, if the KL divergence of each chart is computed individually, then the total cost would be $O((m+n-k)k\binom{m}{k}\binom{n}{k})$, which is huge. To reduce the cost, we propose a *parameter sharing* strategy: when the Gaussians in different charts share parameters (mean and variance), their KL divergence computations will overlap, and this overlap can reduce the total cost to $O(mn)$.

Practical KL Divergence based on Parameter Sharing.

We first discuss the parameter sharing of C. Take $P(\mathsf{C}=i) = 1/|\mathcal{V}|$ for both f_{C} and g_{C}, i.e., each chart is chosen with the same probability $1/|\mathcal{V}|$. Then, $KL(f_{\mathsf{C}} \,||\, g_{\mathsf{C}}) = 0$, and (7) is simplified to

$$KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} KL(f_{\mathsf{X}|\mathsf{C}} \,||\, g_{\mathsf{X}|\mathsf{C}})\big|_{\mathsf{C}=i} = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} KL(f_{\mathsf{Y}|\mathsf{C}} \,||\, g_{\mathsf{Y}|\mathsf{C}})\big|_{\mathsf{C}=i} \tag{9}$$

Now we discuss the parameter sharing of $f_{\mathsf{Y}|\mathsf{C}}$ and $g_{\mathsf{Y}|\mathsf{C}}$ across all charts. Let $\mathsf{v}_1, \mathsf{v}_2 \in \mathbb{R}^{m\times n}$ be two random matrices with independent Gaussian entries. Let $u_1 \in \mathbb{R}^{m\times n}$ and $s_1 \in \mathbb{R}_+^{m\times n}$ be the mean and variance of v_1, such that $\mathrm{vec}(\mathsf{v}_1) \sim \mathcal{N}_{mn}(\mathrm{vec}(u_1), \mathrm{diag}(\mathrm{vec}(s_1)))$; similarly, let $u_2 \in \mathbb{R}^{m\times n}$ and $s_2 \in \mathbb{R}_+^{m\times n}$ be the mean and variance of v_2. Then $KL(\mathsf{v}_1 \,||\, \mathsf{v}_2)$ is the KL divergence of two multivariate Gaussians of dimension mn:

$$KL(\mathsf{v}_1||\mathsf{v}_2) = \sum_{j=1}^{mn} KL((\mathsf{v}_1)_j||(\mathsf{v}_2)_j) = \frac{1}{2}\sum_{j=1}^{mn}\left[(\log\frac{(s_2)_j}{(s_1)_j} + \frac{(s_1)_j}{(s_2)_j}) + (s_2)_j^{\frac{1}{2}}(u_1-u_2)_j - 1\right] \tag{10}$$

For each chart i, we define the Gaussian densities $f_{\mathsf{Y}|\mathsf{C}}$ and $g_{\mathsf{Y}|\mathsf{C}}$ in the following way: take the permutations $H_i(\mathsf{v}_1) = \begin{bmatrix}\mathsf{A}_1 & \mathsf{B}_1\\ \mathsf{C}_1 & \mathsf{D}_1\end{bmatrix}$ and $H_i(\mathsf{v}_2) = \begin{bmatrix}\mathsf{A}_2 & \mathsf{B}_2\\ \mathsf{C}_2 & \mathsf{D}_2\end{bmatrix}$ and replace D_1 and D_2 with $\mathsf{0}$. Then, let $\begin{bmatrix}\mathsf{A}_1 & \mathsf{B}_1\\ \mathsf{C}_1 & \mathsf{0}\end{bmatrix}$ be the random variable of $f_{\mathsf{Y}|\mathsf{C}}$, and $\begin{bmatrix}\mathsf{A}_2 & \mathsf{B}_2\\ \mathsf{C}_2 & \mathsf{0}\end{bmatrix}$ be the random variable of $g_{\mathsf{Y}|\mathsf{C}}$. In this way, each chart takes a portion of the entries in v_1 and v_2 as its own entries, so the parameters in u_1, s_1, u_2, s_2 are shared across all charts.

In v_1(or v_2), each chart chooses k rows from m and k columns from n, so each entry $(\mathsf{v}_1)_j$(or $(\mathsf{v}_2)_j$) appears in $\binom{m-1}{k-1}\binom{n-1}{k-1}$ charts. Therefore, each $KL((\mathsf{v}_1)_j||(\mathsf{v}_2)_j)$ appears $\binom{m-1}{k-1}\binom{n-1}{k-1}$ times in the sum $\sum_{i=1}^{|\mathcal{V}|} KL(f_{\mathsf{Y}|\mathsf{C}} \,||\, g_{\mathsf{Y}|\mathsf{C}})\big|_{\mathsf{C}=i}$, i.e.,

$$\sum_{i=1}^{|\mathcal{V}|} KL(f_{\mathsf{Y}|\mathsf{C}} \,||\, g_{\mathsf{Y}|\mathsf{C}})\big|_{\mathsf{C}=i} = \binom{m-1}{k-1}\binom{n-1}{k-1}\sum_{j=1}^{mn} KL((\mathsf{v}_1)_j||(\mathsf{v}_2)_j) \tag{11}$$

Consequently, by combining (9), (10) and (11), we have

$$KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) = \frac{\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{m}{k}\binom{n}{k}} KL(\mathsf{v}_1||\mathsf{v}_2) = \frac{k^2}{mn} KL(\mathsf{v}_1||\mathsf{v}_2) \tag{12}$$

The $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ computed by (12) has the same computational cost as $KL(\mathsf{v}_1||\mathsf{v}_2)$, i.e., $O(mn)$. This means that $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ is practically computable under parameter sharing.

5 Application: PAC-Bayes Bound for Stochastic LoRA Neural Networks

In classic neural networks, weight matrices are deterministic and optimized directly. However, these matrices can be considered random, since they are typically randomly initialized and stochastically trained. A neural network is stochastic if all its weights are modeled as random variables that follow a

weight distribution, and it is typically trained by optimizing the parameters of the weight distribution. By taking the initial weight distribution as the prior and the trained weight distribution as the posterior, we can establish a PAC-Bayes bound for stochastic neural networks (Dziugaite & Roy, 2017).

Stochastic LoRA Neural Networks (SLNNs). Classic LoRA neural networks represent a weight matrix as $W = W_0 + \Delta W$ where W_0 is fixed and of full rank, ΔW is trainable and of rank-k. Since ΔW can be any element in $\mathcal{M}_k$, we can obtain an SLNN by assigning a wrapped Gaussian mixture on $\mathcal{M}_k$ and replacing ΔW by X, yielding $\mathsf{W} = W_0 + \mathsf{X}$. If an SLNN contains multiple weight matrices $\mathsf{W}_1, \mathsf{W}_2, ..., \mathsf{W}_l$ where $\mathsf{W}_i = W_0^i + \mathsf{X}_i$, we take $\mathsf{X}' = [\mathsf{X}_1, \mathsf{X}_2, ..., \mathsf{X}_l]$ and consider each X_i drawn from an independent wrapped Gaussian mixture, such that the KL divergence of X' is the sum of the KL divergence of each X_i.

PAC-Bayes bound for SLNNs. SLNNs can be characterized by various PAC-bayes bounds. We choose Maurer's bound (Theorem 3.2, (Alquier, 2024); (Maurer, 2004)) for example. Suppose we use binary loss. Let g_{X} be the prior, f_{X} be the posterior, $\mathbb{E}_{f_{\mathsf{X}}}[R(\mathsf{X})]$ be the expected risk and $\mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})]$ be the expected empirical risk. Then, given any $\delta > 0$ and g_{X}, the bound is expressed by

$$P\left(\forall f_{\mathsf{X}},\ \mathbb{E}_{f_{\mathsf{X}}}[R(\mathsf{X})] \leq \mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})] + \sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}}\right) \geq 1 - \delta \tag{13}$$

Our goal is to find an f_{X} that minimizes (13). The wrapped Gaussian mixture X guarantees that $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ is always finite (Section 4), so the bound cannot become infinite. Moreover, (12) provides a practically computable form of $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$, enabling f_{X} being solved efficiently. Details of the computation of (13) are shown in Appendix C. The final bound is calculated by (21).

Experiment. We construct neural networks with the input, hidden, and output layer sizes of $784, 512$ and 1, respectively, and the number of hidden layers ranging from 1 to 3. The hidden layers use ReLU activation and the output layer uses sigmoid activation. We first train these neural networks on MNIST dataset, using SGD with learning rate 0.1, momentum 0.9 and batch size 64. All of them reach ≈ 0 training error and < 0.02 test error within 20 epochs. We then fine-tune them on FashionMNIST dataset using LoRA. For both datasets, we use the default train-test split, and map labels 0–4 to 1 and labels 5–9 to 0 for binary classification. We use binary loss for testing and BCE loss for training, where BCE loss serves as a differentiable surrogate for binary loss.

The PAC-Bayes bound (13) is defined on FashionMNIST test set with binary loss, where we take $\delta = 0.025$ and $N = 10000$ (test set size) and solve it using Adam optimizer with learning rate 0.001. Table 1 presents the results of our bound computed by (21), where a model (l, k) means l hidden layers with rank-k LoRA, and $\hat{\mathbb{E}}^v_{f_{\mathsf{X}}}[r(\mathsf{X})]$ is a sample mean used for estimating $\mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})]$. The results show that the bound (21) is < 1 for all models, indicating that it is non-vacuous. Moreover, for a given l, the bound tends to be tighter for larger k. This indicates that a LoRA neural network with a larger rank tends to generalize better, consistent with the observations of (Shuttleworth et al., 2026).

Table 1: Experimental results

Model (l, k)	$(1, 8)$	$(1, 16)$	$(1, 64)$	$(2, 8)$	$(2, 16)$	$(2, 64)$	$(3, 8)$	$(3, 16)$	$(3, 64)$
$\hat{\mathbb{E}}^v_{f_{\mathsf{X}}}[r(\mathsf{X})]$	0.447	0.395	0.204	0.464	0.406	0.308	0.505	0.426	0.278
$KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$	29.09	52.83	52.82	36.07	89.32	554.39	27.38	69.44	33.84
Bound (21)	0.586	0.546	0.356	0.607	0.572	0.572	0.644	0.584	0.421

6 Conclusion

We identify a fundamental issue in applying PAC-Bayes bounds to low-rank models: when choosing prior and posterior to be Gaussians over low-rank matrices, their supports may not coincide, since no Gaussian distribution can cover $\mathcal{M}_k$. This partial coverage can cause the KL divergence to be infinite, yielding vacuous PAC-Bayes bounds. To address this issue, we propose a wrapped Gaussian mixture whose support covers $\mathcal{M}_k$, and derive practically computable KL divergence bound for the mixture. We apply this mixture to establish a PAC-Bayes bound for stochastic LoRA neural networks and empirically show that the bound is non-vacuous and captures LoRA generalization behavior, suggesting that the mixture can be a useful statistic tool for analyzing low-rank model performance.

AI USE STATEMENT

In this work, we used generative AI tools for grammar correction, polishing writing and generating code for experiment. We have not used generative AI tools for generating insights, math proofs or references. We have reviewed all AI-assisted work, especially, the LLM-generated code was verified and tested for correctness by the primary author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work is mainly theoretical and not directly related to real world application. We do not foresee any harmful consequences of our work in the real world.

REPRODUCIBILITY STATEMENT

We have provided experimental details in Appendix C, and have attached the source code. We use two public datasets MNIST and FashionMNIST. Both of them can be downloaded by our code.

REFERENCES

Pierre Alquier. Bayesian methods for low-rank matrix estimation: short survey and theoretical study. In *International Conference on Algorithmic Learning Theory*, pp. 309–323. Springer, 2013.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 2006.

Thibault De Surrel, Fabien Lotte, Sylvain Chevallier, and Florian Yger. Wrapped Gaussian on the manifold of symmetric positive definite matrices. In *International Conference on Machine Learning*, pp. 12816–12839. PMLR, 2025.

J-L Durrieu, J-Ph Thiran, and Finnian Kelly. Lower and upper bounds for approximation of the kullback-leibler divergence between gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4833–4836, 2012.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

Xinlong Feng and Zhinan Zhang. The rank of a random matrix. *Applied mathematics and computation*, 185(1):689–694, 2007.

Ruixin Guo, Ruoming Jin, Xinyu Li, and Yang Zhou. PAC-Bayes bounds for multivariate linear regression and linear autoencoders. In *Advances in Neural Information Processing Systems*, 2026.

John R Hershey and Peder A Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pp. IV–317, 2007.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yerlan Idelbayev and Miguel A Carreira-Perpinán. Low-rank compression of neural nets: Learning the rank of each layer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8049–8059, 2020.

Ioan Mackenzie James. On category, in the sense of Lusternik-Schnirelmann. *Topology*, 1978.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Antoine Ledent, Rodrigo Alves, and Yunwen Lei. Generalization bounds for rank-sparse neural networks. In *Advances in Neural Information Processing Systems*, 2026.

John M Lee. *Introduction to Smooth Manifolds (Second Edition)*. Springer, 2012.

The Tien Mai. On a low-rank matrix single-index model. *Mathematics*, 11(9):2065, 2023.

Andreas Maurer. A note on the PAC Bayesian theorem. *arXiv preprint*, 2004.

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 230–234, 1998.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International conference on machine learning*, pp. 4693–4702. PMLR, 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Athanasios N Nikolakopoulos, Xia Ning, Christian Desrosiers, and George Karypis. Trust your neighbors: A comprehensive survey of neighborhood-based methods for recommender systems. *Recommender systems handbook*, pp. 39–89, 2021.

Calyampudi R Rao. *Linear statistical inference and its applications*. Wiley New York, 1973.

Walter Rudin. *Real and Complex Analysis (Third Edition)*. McGraw-Hill, 1987.

Uri Shalit, Daphna Weinshall, and Gal Chechik. Online learning in the embedded manifold of low-rank matrices. *Journal of Machine Learning Research*, 13(2), 2012.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*, 2026.

Harald Steck. Autoencoders that don't overfit towards the identity. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19598–19608, 2020.

André Uschmajew and Bart Vandereycken. Geometric methods on low-rank matrix and tensor manifolds. In *Handbook of variational methods for nonlinear geometric data*, pp. 261–313. Springer, 2020.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint*, 2012.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-One: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *International Conference on Machine Learning*, pp. 75513–75574. PMLR, 2025.

Jiacheng Zhu, Kristjan Greenewald, Nadjahi, et al. Asymmetry in low-rank adapters of foundation models. In *International Conference on Machine Learning*, pp. 62369 – 62385, 2024.

CONTENTS

1 Introduction **1**

2 Low-Rank Matrix Manifolds and Challenges in Distribution Covering **2**

2.1 The manifold of $\mathcal{M}_k$ 2

2.2 Gaussian Distribution cannot Cover $\mathcal{M}_k$ 3

2.3 Challenges in Covering $\mathcal{M}_k$ using Wrapped Gaussian 4

2.4 The Consequence for not Covering $\mathcal{M}_k$ 4

3 Constructing a Wrapped Gaussian Mixture Covering $\mathcal{M}_k$ **5**

4 Practical KL Divergence Bound for the Wrapped Gaussian Mixture **7**

5 Application: PAC-Bayes Bound for Stochastic LoRA Neural Networks **8**

6 Conclusion **9**

A Math Notation **13**

B Supplemental Propositions **13**

C Experimental Details **15**

D Related Work **17**

E Discussion and Limitations **18**

A MATH NOTATION

Symbol	Description
X, Y	Random variables, denoted in mathsf font. We use normal fonts X, Y or x, y to represent their realization or any deterministic variables.
$A \setminus B$	Set difference. $A \setminus B = \{x : x \in A, x \notin B\}$.
$A \subset B$	A is a subset of B, i.e., $x \in B$ if $x \in A$.
$A \subsetneq B$	A is a proper subset of B, i.e., $A \subset B$ and $A \neq B$.
$A \not\subset B$	A is not a subset of B, i.e., $A \setminus B \neq \varnothing$.
$\mathcal{M}_k$	The set of all rank-k matrices in $\mathbb{R}^{m\times n}$.
$\mathcal{V}$	The index set of charts. $\mathcal{V} = \{1, 2, ..., \binom{m}{k}\binom{n}{k}\}$.
I_n	The $n \times n$ identity matrix.
$X \otimes Y$	The Kronecker product of the matrices X and Y.
range	$\mathrm{range}(X)$ means the linear space spanned by the columns of the matrix X.
dim	The dimension of a set when viewed as an Euclidean space or a smooth manifold.
vec	Vectorization. For a matrix $X \in \mathbb{R}^{m\times n}$, $\mathrm{vec}(X) = [X_{*1}^T, X_{*2}^T, ..., X_{*n}^T]^T \in \mathbb{R}^{mn}$, which is an mn dimensional column vector.
vec^{-1}	The inverse operation of vec .
supp	The support of a distribution or a random variable. Let $\mathsf{X} \sim \mathcal{P}$ and f_{X} be the density of X, then $\mathrm{supp}(\mathcal{P}) = \mathrm{supp}(\mathsf{X}) = \{X : f_{\mathsf{X}}(X) > 0\}$.
diag	For $x \in \mathbb{R}^d$, $\mathrm{diag}(x) \in \mathbb{R}^{d\times d}$ is a diagonal matrix with all elements of x in the diagonal.
$\mathcal{N}_d(\mu, \Sigma)$	Multivariate Gaussian distribution in $\mathbb{R}^d$, with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$.

B SUPPLEMENTAL PROPOSITIONS

We present essential propositions that support our arguments in the main paper. These propositions are based on the foundations of topology, measure theory and differential geometry. Some propositions may lack explicit proofs in the existing literature. We provide proofs for all these propositions.

Proposition B.1. *For any I and J, $\mathcal{U}_{IJ}$ is an open subset of $\mathbb{R}^{m\times n}$.*

Proof: Given any I and J, define a function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ by $f(X) = \det(X_{IJ})$ for any $X \in \mathbb{R}^{m\times n}$, then f is continuous. Since $\mathcal{U}_{IJ} = f^{-1}(\mathbb{R} \setminus \{0\})$ and $\mathbb{R} \setminus \{0\}$ is open, by continuity, $\mathcal{U}_{IJ}$ is open. □

Recall that in measure theory, a probability space is defined as $(\mathcal{X}, \Sigma, \mu)$ where $\mathcal{X}$ is a set, Σ is the Borel σ-algebra of $\mathcal{X}$, and μ is a Borel measure defined on Σ such that $\mu(\mathcal{X}) = 1$. Here $\mathcal{X}$ can be any topological space, including the Euclidean space $\mathbb{R}^{m\times n}$ and the manifold $\mathcal{M}_k$. This is because, any topological space is equipped with a topology, and a Borel σ-algebra can always be generated from a topology (1.11, (Rudin, 1987)); the Borel σ-algebra allows a Borel measure μ to be defined. μ is referred to as a *probability measure*, and any probability distribution on $\mathcal{X}$ can be represented by some μ. We say $A \subset \mathcal{X}$ *has measure zero* in $\mathcal{X}$ if for any μ with $\mathrm{supp}(\mu) = \mathcal{X}$, $\mu(A) = 0$.

Lemma B.2. (Sard's Theorem) *Let $\mathcal{M}$ be a smooth manifold and $\mathcal{S} \subset \mathcal{M}$ be an immersed submanifold. If $\dim(\mathcal{S}) < \dim(\mathcal{M})$, then $\mathcal{S}$ has measure zero in $\mathcal{M}$.*

Proof: See Corollary 6.12, (Lee, 2012). □

Proposition B.3. *For any $k < \min\{m, n\}$, $\mathcal{M}_k$ has measure zero in $\mathbb{R}^{m\times n}$.*

Proof: For any $k < \min\{m, n\}$, $\mathcal{M}_k \subsetneq \mathbb{R}^{m\times n}$. By Proposition 2.1, we have $\dim(\mathcal{M}_k) = (m + n - k)k < mn = \dim(\mathbb{R}^{m\times n})$. By Lemma B.2, $\mathcal{M}_k$ has measure zero in $\mathbb{R}^{m\times n}$. □

Corollary B.4. *For any real random matrix $\mathsf{W} \in \mathbb{R}^{m\times n}$ with $\mathrm{supp}(\mathsf{W}) = \mathbb{R}^{m\times n}$, W has full rank with probability* 1.

Proof: An earlier proof based on determinant is provided by (Feng & Zhang, 2007). Here we provide an alternative proof based on Sard's theorem.

The distribution of W can be represented by a probability measure on $\mathbb{R}^{m\times n}$. Let T be a realization of W. If T is not of full rank, then $T \in \bigcup_{k=1}^{\min\{m,n\}-1} \mathcal{M}_k$. By Proposition B.3, each $\mathcal{M}_k$ with $1 \le k \le \min\{m,n\}-1$ has measure zero. Hence, $\bigcup_{k=1}^{\min\{m,n\}-1} \mathcal{M}_k$ has measure zero since it is a union of finite measure zero sets. This implies that the set of all full rank matrices, denoted as $\mathbb{R}^{m\times n} \setminus \left(\bigcup_{k=1}^{\min\{m,n\}-1} \mathcal{M}_k\right)$, is of measure 1. □

The following proposition demonstrates the relationship between entry-independence and non-singularity of a Gaussian matrix.

Proposition B.5. *Given a Gaussian matrix $\mathsf{W} \in \mathbb{R}^{m\times n}$, if all its entries are independent, then W is of full rank with probability 1. Equivalently, if W is not of full rank with positive probability, then the entries of W cannot be independent.*

Proof: All entries of W being independent implies that the covariance matrix of $\mathrm{vec}(\mathsf{W})$ has full rank, meaning that the support of the Gaussian distribution is the whole space $\mathbb{R}^{m\times n}$. By Corollary B.4, W is of full rank with probability 1. □

Let $u \in \mathbb{R}^d$ be a vector and $\mathcal{S} \subset \mathbb{R}^d$ be a vector subspace, then the set $u + \mathcal{S} = \{u + s : s \in \mathcal{S}\}$ is called an *affine subspace* (Lee, 2012). It is a set formed by shifting every element in $\mathcal{S}$ by a constant vector u. In particular, if $u \in \mathcal{S}$, then $u + \mathcal{S}$ contains 0, and it becomes a vector space.

Proposition B.6. *For any Gaussian distribution $\mathcal{Q}$ on $\mathbb{R}^{m\times n}$, whether its covariance matrix is singular or non-singular, $\mathrm{supp}(\mathcal{Q})$ is an affine subspace.*

Proof: Let $\mathsf{W} \sim \mathcal{Q}$. Since $\mathrm{vec}(\mathsf{W}) \in \mathbb{R}^{mn}$ is a Gaussian random vector, we suppose $\mathrm{vec}(\mathsf{W}) \sim \mathcal{N}_{mn}(\mu, \Sigma)$. The covariance matrix Σ can be either of singular or non-singular. Suppose $\mathrm{rank}(\Sigma) = t$ where $1 \le t \le mn$, there must exist a full-rank matrix $Q \in \mathbb{R}^{mn\times t}$ such that $\Sigma = QQ^T$. Hence, $\mathrm{vec}(\mathsf{W}) \in \mathcal{N}_{mn}(\mu, \Sigma)$ if and only if $\mathrm{vec}(\mathsf{W}) = Q\epsilon + \mu$, where $\epsilon \sim \mathcal{N}_t(0, I_t)$ (Rao, 1973). Let $\mathcal{S} = \mathrm{span}(Q_{*1}, Q_{*2}, ..., Q_{*k})$ be the vector space spanned by the columns of Q, we have $\mathrm{supp}(\mathrm{vec}(\mathsf{W})) = \mu + \mathcal{S}$. Let vec^{-1} be the inverse operation of vec (i.e., reorganizing the elements in a vector back to a matrix), then $\mathrm{supp}(\mathsf{W}) = \mathrm{vec}^{-1}(\mu) + \mathrm{vec}^{-1}(\mathcal{S})$. Here $\mathrm{vec}^{-1}(\mathcal{S})$ is a still vector space since vec^{-1} is isomorphic. □

Theorem B.7. *For any Gaussian $\mathcal{Q}$ on $\mathbb{R}^{m\times n}$ with $\dim(\mathrm{supp}(\mathcal{Q})) = \dim(\mathcal{M}_k)$, $\mathcal{M}_k \not\subset \mathrm{supp}(\mathcal{Q})$ holds, i.e., $\mathcal{M}_k$ cannot be covered by $\mathcal{Q}$.*

Proof: By Proposition B.6, for any $\mathcal{Q}$ with $\dim(\mathrm{supp}(\mathcal{Q})) = \dim(\mathcal{M}_k) = (m+n-k)k$, there must exist a full rank $Q \in \mathbb{R}^{mn\times(m+n-k)k}$ and a direction vector $\mu \in \mathbb{R}^{mn}$ such that the support

$$\mathrm{supp}(\mathrm{vec}(\mathcal{Q})) = \{Q\epsilon + \mu : \epsilon \in \mathbb{R}^{(m+n-k)k}\}$$

is an affine subpace. Here $\mathrm{vec}(\mathcal{Q})$ on $\mathbb{R}^{mn}$ is transferred from $\mathcal{Q}$ on $\mathbb{R}^{m\times n}$ by vec.

Let $\epsilon = \begin{bmatrix}\mathrm{vec}(A)\\ \mathrm{vec}(B)\\ \mathrm{vec}(C)\end{bmatrix}$ where $A \in \mathbb{R}^{k\times k}$, $B \in \mathbb{R}^{k\times(n-k)}$ and $C \in \mathbb{R}^{(m-k)\times k}$, then all invertible A and arbitrary B, C form a subset $\mathcal{S} \subset \mathbb{R}^{(m+n-k)k}$. By Proposition 2.1, we can define a smooth mapping

$$G^{-1}(\epsilon) = \begin{bmatrix}\mathrm{vec}(A)\\ \mathrm{vec}(B)\\ \mathrm{vec}(C)\\ \mathrm{vec}(CA^{-1}B)\end{bmatrix} := \mathrm{vec}\left(\begin{bmatrix}A & B\\ C & CA^{-1}B\end{bmatrix}\right)$$

such that $G^{-1}(\mathcal{S}) = \mathrm{vec}(\mathcal{M}_k \cap \mathcal{U}_{IJ})$ for some I, J.

Given $\epsilon_0 = \begin{bmatrix}\mathrm{vec}(A_0)\\ \mathrm{vec}(B_0)\\ \mathrm{vec}(C_0)\end{bmatrix} \in \mathcal{S}$, the Jacobian matrix of G^{-1} at ϵ_0 is

$$\left.\frac{\partial G^{-1}}{\partial \epsilon}\right|_{\epsilon=\epsilon_0} = \begin{bmatrix} I_{k^2} & 0 & 0\\ 0 & I_{k(n-k)} & 0\\ 0 & 0 & I_{(m-k)k}\\ -(B_0^T \otimes C_0)(A_0^{-T} \otimes A_0^{-1}) & I_{n-k} \otimes (C_0 A_0^{-1}) & (A_0^{-1}B_0)^T \otimes I_{m-k}\end{bmatrix}$$

Thus, the tangent space of G^{-1} at ϵ_0 is $\mathrm{range}(\frac{\partial G^{-1}}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0})$, whose basis depend on ϵ_0.

Since $\mathrm{supp}(\mathrm{vec}(\mathcal{Q}))$ is an affine subspace, its tangent space $\mathrm{range}(Q)$ is the same at every point. If there exists a $\mathcal{Q}$ such that $G^{-1}(\mathcal{S}) \subset \mathrm{supp}(\mathrm{vec}(\mathcal{Q}))$, then the tangent space at every point in $G^{-1}(\mathcal{S})$ is the same, i.e., $\mathrm{range}(\frac{\partial G^{-1}}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0}) = \mathrm{range}(Q)$ for any $\epsilon_0 \in \mathcal{S}$, which contradicts.

Hence, there exists $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ for some I, J such that $\mathcal{M}_k \cap \mathcal{U}_{IJ} \not\subset \mathrm{supp}(\mathcal{Q})$ for any $\mathcal{Q}$, implying that $\mathcal{M}_k \not\subset \mathrm{supp}(\mathcal{Q})$. □

Theorem B.7 demonstrates that $\mathcal{M}_k$ always contains a nonlinear submanifold $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ that cannot fit into any linear manifold of the same dimension. This also implies that $\mathcal{M}_k$ is not an affine subspace for any $k < \min\{m, n\}$.

Let B be a topological space and $A \subset B$ be a subset. We say A is *everywhere dense* in B if the closure of A is B, and A is *nowhere dense* in B if the closure of A contains no nonempty open subset of B. It is known that A is closed and nowhere dense if and only if A^c is everywhere dense.

Proposition B.8. *For any I and J, $\mathcal{U}_{IJ}$ is everywhere dense in $\mathbb{R}^{m\times n}$.*

Proof: Following the proof of Proposition B.1, we know that $\mathcal{U}_{IJ}^c = f^{-1}(0)$, which means that $\mathcal{U}_{IJ}^c$ is a level set of f. By Submersion Level Set Theorem (Corollary 5.13, (Lee, 2012)), $\mathcal{U}_{IJ}^c$ is a $mn-1$ dimensional submanifold embedded in $\mathbb{R}^{m\times n}$. By Sard's Theorem, $\mathcal{U}_{IJ}^c$ has measure zero in $\mathbb{R}^{m\times n}$, implying that it contains no nonempty open subset in $\mathbb{R}^{m\times n}$ (Corollary C.25, (Lee, 2012)). Hence, $\mathcal{U}_{IJ}^c$ is nowhere dense in $\mathbb{R}^{m\times n}$.

Since $\mathcal{U}_{IJ}$ is open, $\mathcal{U}_{IJ}^c$ is closed, implying that $\mathcal{U}_{IJ}$ is everywhere dense. □

Corollary B.9. *For any I and J, $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ is everywhere dense in $\mathcal{M}_k$.*

Proof: Recall that we denote Ω as the subspace topology of $\mathcal{M}_k$ in Section 2.1. By Proposition B.8, since $\mathcal{U}_{IJ}$ is everywhere dense in $\mathbb{R}^{m\times n}$, the closure of $\mathcal{U}_{IJ}$ is $\mathbb{R}^{m\times n}$. Under Ω, the closure of $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ is the whole topological subspace $\mathcal{M}_k \cap \mathbb{R}^{m\times n} = \mathcal{M}_k$. Thus, $\mathcal{M}_k \cap \mathcal{U}_{IJ}$ is everywhere dense in $\mathcal{M}_k$. □

C Experimental Details

This section presents the details of our experiment in Section 5, including the notation, deduction and computation the PAC-Bayes bound for SLNNs (13) and its corresponding hyperpatameter choice.

Notation of the PAC-Bayes bound.

We focus on the supervised learning setting for a binary classification task. Let $S = \{x_i, y_i\}_{i=1}^N$ be the dataset where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ for all i. Since a LoRA weight matrix is expressed by $W = W_0 + \Delta W$ where only ΔW is trainable, we can denote a LoRA neural network for binary classification as $h_{\Delta W} : \mathbb{R}^d \to [0, 1]$. For each input x_i, let $\hat{y}_i = h_{\Delta W}(x_i)$ be the prediction and $\mathbb{I}_{[|y_i - h_{\Delta W}(x_i)| \le 0.5]} \in \{0, 1\}$ be the binary loss ($\mathbb{I}_P = 0$ if the condition P holds, otherwise $\mathbb{I}_P = 1$), then the empirical risk is defined as $r(\Delta W) = \frac{1}{N}\sum_{i=1}^N \mathbb{I}_{[|y_i - h_{\Delta W}(x_i)| \le 0.5]}$. Suppose each data pair (x_i, y_i) in S is sampled from a distribution $\mathcal{D}$, the risk is defined as $R(\Delta W) = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[\mathbb{I}_{[|y_i - h_{\Delta W}(x_i)| \le 0.5]}]$. It is clear that $R(\Delta W) = \mathbb{E}_{S\sim\mathcal{D}^N}[r(\Delta W)]$.

When converting a classic LoRA neural network into a stochastic LoRA neural network, we assume that ΔW is sampled from a wrapped Gaussian mixture, and replace ΔW with the random matrix X. Suppose the distribution of X is represented by the density f_{X}, we denote $\mathbb{E}_{f_{\mathsf{X}}}[R(\mathsf{X})]$ as the expected risk and $\mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})]$ as the expected empirical risk, then the gap $\mathbb{E}_{f_{\mathsf{X}}}[R(\mathsf{X})] - \mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})]$ can be estimated by PAC-Bayes bounds.

Derive the PAC-Bayes Bound.

Applying the stochastic LoRA weight matrix $\mathsf{W} = W_0 + \mathsf{X}$ to Maurer's bound (Theorem 3.2, (Alquier, 2024); (Maurer, 2004)): given any $\delta > 0$ and g_{X}, we have

$$P\left(\forall f_{\mathsf{X}},\ \mathbb{E}_{f_{\mathsf{X}}}[R(\mathsf{X})] \le \mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})] + \sqrt{\frac{KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}}) + \log(2\sqrt{N}/\delta)}{2N}}\right) \ge 1 - \delta$$

Since $\text{supp}(\mathsf{X}) = \mathcal{M}_k$, f_X is always absolutely continuous with respect to g_X, which avoids $KL(f_\mathsf{X} \,||\, g_\mathsf{X})$ being infinity. But $KL(f_\mathsf{X} \,||\, g_\mathsf{X})$ has no closed form and is difficult to compute directly. To address this, we apply $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) \geq KL(f_\mathsf{X} \,||\, g_\mathsf{X})$ (8) and get

$$P\left(\forall f_\mathsf{X},\ \mathbb{E}_{f_\mathsf{X}}[R(\mathsf{X})] \leq \mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})] + \sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}}\right) \geq 1 - \delta \tag{14}$$

which is (13) in the main paper. Here, $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ is finite and has closed form, but it may not be practically computable. We can make the computation practical by introducing an extra parameter sharing assumption, under which $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ is computed by (12).

For simplicity, the bound is expressed with a single mixture X. If an SLNN contains multiple independent mixtures $\mathsf{X}' = [\mathsf{X}_1, \mathsf{X}_2, ..., \mathsf{X}_l]$ (Section 5), we replace X in the bound by X'.

Compute the PAC-Bayes bound.

Our approach of computing PAC-Bayes bounds mainly follows from (Dziugaite & Roy, 2017). The goal is to find f_X that minimizes the bound (14), i.e., given g_X, δ and N, we solve

$$\min_{f_\mathsf{X}}\ \mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})] + \sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}} \tag{15}$$

In (15), the $\sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}}$ part is easy to compute based on our work, while the term $\mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})]$ remains difficult to compute – this issue is common in computing PAC-Bayes bounds (Dziugaite & Roy, 2017). There are two difficulties: (a) $r(\mathsf{X})$ is based on binary loss, which is discrete and not differentiable, making it impossible to compute the gradients. (b) The expectation $\mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})]$ is typically intractable due to the irregularity of f_X and r. We address (a) using binary cross-entropy (BCE) loss as a differentiable surrogate, and address (b) using Monte Carlo sampling.

Step 1. We first obtain an f_X for (15) using stochastic gradient descent.

Recall the setting of practical KL divergence in Section 4: f_X is parameterized by $u_1 \in \mathbb{R}^{m\times n}$ and $s_1 \in \mathbb{R}_+^{m\times n}$, and g_X is parameterized by $u_2 \in \mathbb{R}^{m\times n}$ and $s_2 \in \mathbb{R}_+^{m\times n}$. In this case, (15) becomes fixing u_2, s_2 and solving for u_1, s_1. We set $s_1 = \alpha/(1 + e^{s_1'})$ where $s_1' \in \mathbb{R}^{m\times n}$, and $s_2 = \alpha/(1 + e^{s_2'})$ where $s_2' \in \mathbb{R}^{m\times n}$, such that all entries in s_1, s_2 are ranged in $(0, \alpha)$. Here α is a tunable hyperparameter and its choice affects the tightness of the bound, as shown in Table 2. Under this representation, we can solve for s_1 by optimizing s_1'.

Moreover, we restrict $u_1, u_2 \in \mathcal{M}_k$ such that, when variances s_1, s_2 are small, f_X and g_X concentrate around u_1 and u_2, respectively, analogous to Gaussian distributions with small variances (Appendix E). In this case, (15) can be written as

$$u_1^*, {s_1'}^* = \underset{u_1\in\mathcal{M}_k,\, s_1'\in\mathbb{R}^{m\times n}}{\arg\min}\ \mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})] + \sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}} \tag{16}$$

We still need to address the non-differentiability and intractability of $\mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})]$ to make (16) computable. First, we take one sample of $r(\mathsf{X})$, and let it be $r(u_1)$, so that $r(u_1)$ is a single-sample estimate of $\mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})]$. Then, we replace $r(u_1)$ by $\hat{r}(u_1)$, where $\hat{r}$ is obtained from r by replacing the binary loss with the BCE loss. In this way, (16) is converted to

$$u_1^*, {s_1'}^* = \underset{u_1\in\mathcal{M}_k,\, s_1'\in\mathbb{R}^{m\times n}}{\arg\min}\ \hat{r}(u_1) + \sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}} \tag{17}$$

Now, (17) is computable and differentiable. We solve it using Adam, where u_1 is optimized using LoRA gradient updates. The obtained $u_1^*, {s_1'}^*$ are the parameters of f_X. Note that this f_X is in general not an optimal solution of (15), but a sub-optimal one. We consider this sub-optimal solution acceptable as long as it can give a tight bound.

Step 2. After f_X is obtained, we plug it into (15) to compute the minimum. Still, $\mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})]$ is intractable with the given f_X, so we approximate it using Monte Carlo sampling.

Let $\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_v$ be v samples of X, we define the sample mean $\hat{\mathbb{E}}_{f_\mathsf{X}}^v[r(\mathsf{X})] = \frac{1}{v}\sum_{i=1}^v r(\mathsf{x}_i)$. Then, $\hat{\mathbb{E}}_{f_\mathsf{X}}^v[r(\mathsf{X})]$ is tractable, and the error of using $\hat{\mathbb{E}}_{f_\mathsf{X}}^v[r(\mathsf{X})]$ to estimate $\mathbb{E}_{f_\mathsf{X}}[r(\mathsf{X})]$ can be characterized

by Hoeffding's bound:

$$\forall f_{\mathsf{X}},\ P\left(\mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})] \geq \hat{\mathbb{E}}^{v}_{f_{\mathsf{X}}}[r(\mathsf{X})] + \sqrt{\frac{\log(1/\delta')}{2v}}\right) \leq \delta' \tag{18}$$

However, (18) is not compatible with the uniform bound (14), which holds for an intersection of events $\forall f_{\mathsf{X}}$. We first convert (14) into a single-event bound:

$$\forall f_{\mathsf{X}},\ P\left(\mathbb{E}_{f_{\mathsf{X}}}[R(\mathsf{X})] \leq \mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})] + \sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}}\right) \geq 1-\delta \tag{19}$$

Then, combining (18) and (19), we get

$$\forall f_{\mathsf{X}},\ P\left(\mathbb{E}_{f_{\mathsf{X}}}[R(\mathsf{X})] \leq \hat{\mathbb{E}}^{v}_{f_{\mathsf{X}}}[r(\mathsf{X})] + \sqrt{\frac{\log(1/\delta')}{2v}} + \sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}}\right) \geq 1-\delta-\delta' \tag{20}$$

Thus, all terms in the bound (20) are computable. We plug the f_{X} obtained from (17) into

$$\hat{\mathbb{E}}^{v}_{f_{\mathsf{X}}}[r(\mathsf{X})] + \sqrt{\frac{\log(1/\delta')}{2v}} + \sqrt{\frac{KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) + \log(2\sqrt{N}/\delta)}{2N}} \tag{21}$$

and let (21) be the final bound that we report in Table 1. We set $\delta' = 0.025$ and $v = 200$, such that $\sqrt{\frac{\log(1/\delta')}{2v}} \approx 0.096$. Note that this quantity can be reduced if using a larger number of samples v, but the computation will be more expensive.

Hyperparameter Choice.

Remember that we set all entries in the variance matrices s_1, s_2 to range in $(0, \alpha)$, where α is a tunable hyperparameter. Using the model $(1, 8)$ in Table 1 for example, we investigate how different choices of α affect the tightness of the bound.

In Table 2, both $r(u_1)$ and $\hat{\mathbb{E}}^{v}_{f_{\mathsf{X}}}[r(\mathsf{X})] = \frac{1}{v}\sum_{i=1}^{v} r(\mathsf{x}_i)$ approximate $\mathbb{E}_{f_{\mathsf{X}}}[r(\mathsf{X})]$, where the former is a fixed estimate and the latter is a v-sample estimator. The results show that when α decreases, $\hat{\mathbb{E}}^{v}_{f_{\mathsf{X}}}[r(\mathsf{X})]$ tends to decrease and become closer to $r(u_1)$. The reason is that, u_1 locates at the center of the density f_{X}, while the v samples $\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_v$ locate around u_1. When α is large, $\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_v$ tend to be far away from u_1, and each x_i tends to perform worse, making $r(\mathsf{x}_i)$ larger than $r(u_1)$; when α decreases, $\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_v$ concentrate to u_1, making each $r(\mathsf{x}_i)$ closer to $r(u_1)$. When α changes from 10^{-7} to 10^{-8}, $\hat{\mathbb{E}}^{v}_{f_{\mathsf{X}}}[r(\mathsf{X})]$ increases because $r(u_1)$ increases.

On the other hand, $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ tends to increase as α decreases. This is because, as α converges to 0, on each chart $\mathsf{C} = i$, the Gaussian densities $f_{\mathsf{X}|\mathsf{C}}$ and $g_{\mathsf{X}|\mathsf{C}}$ will converge to two Dirac delta functions centered at two different points, whose KL divergence is infinite.

The optimal bound (21) occurs at the trade-off point between decreasing $\hat{\mathbb{E}}^{v}_{f_{\mathsf{X}}}[r(\mathsf{X})]$ and increasing $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$, which corresponds to $\alpha = 10^{-3}$. We consider this to be the optimal hyperparameter and test all models in Table 1 under $\alpha = 10^{-3}$.

Table 2: The bound obtained from different α in the $(1, 8)$ model.

α	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
$r(u_1)$	0.063	0.063	0.066	0.072	0.085	0.104	0.148	0.234
$\hat{\mathbb{E}}^{v}_{f_{\mathsf{X}}}[r(\mathsf{X})]$	0.495	0.463	0.447	0.432	0.420	0.348	0.254	0.289
$KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$	4.89	12.35	29.09	108.61	412.89	2336.8	14108	79876
Bound (21)	0.617	0.591	0.586	0.604	0.661	0.786	1.190	2.384

D Related Work

Low-rank models have been popularized in deep learning in recent years through LoRA-based parameter-efficient fine-tuning (PEFT) techniques (Hu et al., 2022; Meng et al., 2024; Zhu et al.,

2024; Zhang et al., 2025). However, from a statistical learning perspective, the generalizability of these models has not been well studied. LoRA-based approaches can, in general, be viewed as introducing a trainable low-rank perturbation matrix ΔW to a fixed base matrix W_0, resulting in $W_0 + \Delta W$. Existing works use spectral norm-based generalization bounds to characterize such perturbations, such as (Bartlett et al., 2017; Neyshabur et al., 2018). However, these bounds cannot capture rank, since the spectral norm *does not characterize rank*: a full-rank matrix can be converted to a low-rank matrix through truncated SVD while its spectral norm remains unchanged, as long as its largest singular value is not truncated. The norm that most closely relates to rank is the Schatten-p norm as $p \to 0$ (although it becomes a semi-norm in this limit). Perturbation bounds based on the Schatten-p norm have also been proposed (Ledent et al., 2026), but these bounds can become vacuous even in low-rank regimes unless the norm is aggressively constrained. Due to these difficulties, we consider norm-based approaches unsuitable for analyzing low-rank models, and instead propose a distribution-based approach for their analysis.

This distribution-based approach can be applied to another category of generalization bounds called PAC-Bayes bounds (McAllester, 1998; Dziugaite & Roy, 2017; Alquier, 2024). These bounds use information-theoretic quantities such as KL divergence $KL(\rho \,||\, \pi)$ to characterize generalizability. They are typically tighter than norm-based bounds and have been empirically shown to be non-vacuous on neural networks (Dziugaite & Roy, 2017). There are a few studies of PAC-Bayes bounds on low-rank models, which focus on training a low-rank matrix to recover the ground-truth low-rank matrix in a regression setting (Alquier, 2013; Mai, 2023). However, these works do not consider the geometry of $\mathcal{M}_k$, and their analysis relies on complex low-rank distributions that are typically characterized less precisely by introducing unspecified constants (Appendix E). These bounds are not directly computable since the unspecified constants prevent the evaluation of their exact values, making them less usable in practice. Different from these works, we utilize the geometric property of $\mathcal{M}_k$ to obtain a simpler low-rank distribution (wrapped Gaussian mixture). This distribution has a closed-form density expression and does not rely on unspecified constants, which simplifies the low-rank analysis and yields a practically computable bound.

Currently in machine learning research, the low-rank manifold $\mathcal{M}_k$ has been a popular tool for manifold optimization (Shalit et al., 2012; Vandereycken, 2013). Our work does not focus on optimization. Instead, we use $\mathcal{M}_k$ in a novel way: introducing a distribution on $\mathcal{M}_k$ and using it as a tool for performance evaluation. It can be used to evaluate the performance of LoRA as part of a PAC-Bayes bound, but it does not inherently serve as a LoRA variant for optimization. On the other hand, existing works have studied introducing distributions on simple manifolds that can be covered by a single chart (Nagano et al., 2019; De Surrel et al., 2025), but they have not yet studied introducing distributions on $\mathcal{M}_k$, which is a more complex manifold and in general cannot be covered by a single chart. The main difficulty is that, constructing a distribution that covers $\mathcal{M}_k$ requires *a valid multiple-chart covering*, which previous works have not obtained. We address this issue through a careful analysis based on the work of (Uschmajew & Vandereycken, 2020), which provides the differential-geometric foundation for $\mathcal{M}_k$.

E Discussion and Limitations

On the Definition of Support.

The definition of support varies in literature. In information theory, it is typically defined as the set of all points with positive density, i.e., $\{X : f_{\mathsf{X}}(X) > 0\}$ (Cover, 2006). In real analysis, it is typically defined as the closure of the previous set (Rudin, 1987), i.e., it may involve points with zero density. We use the first definition in our paper, since it is more convenient for analyzing KL divergence.

Under the first definition, $\rho \ll \pi$ if and only if $\mathrm{supp}(\rho) \subset \mathrm{supp}(\pi)$, indicating that the absolute continuity can be directly represented by set inclusion. Under the second definition, $\rho \ll \pi$ only if $\mathrm{cl}(\mathrm{supp}(\rho)) \subset \mathrm{cl}(\mathrm{supp}(\pi))$, where cl denotes the closure. Note that the converse may not be true. For example, let ρ, π be single-chart wrapped Gaussians on $\mathcal{M}_k$, with $\mathrm{supp}(\rho) = \mathcal{M}_k \cap \mathcal{U}_1$ and $\mathrm{supp}(\pi) = \mathcal{M}_k \cap \mathcal{U}_2$ respectively, then $\mathrm{cl}(\mathrm{supp}(\rho)) = \mathrm{cl}(\mathrm{supp}(\pi)) = \mathcal{M}_k$ (Corollary B.9). In this case, we have $\mathrm{cl}(\mathrm{supp}(\rho)) \subset \mathrm{cl}(\mathrm{supp}(\pi))$, but $\rho \ll \pi$ does not hold.

Therefore, the first definition is better in characterizing $\rho \ll \pi$, which is critical in analyzing the finiteness of KL divergence.

Note that $KL(\rho \,||\, \pi) = \infty$ if $\text{supp}(\rho) \setminus \text{supp}(\pi) \neq \varnothing$. In (2), $\int_{\text{supp}(\rho)}$ is a Riemann integral. If there exists a $W' \in \text{supp}(\rho) \setminus \text{supp}(\pi)$, then $f_\rho(W') \log \frac{f_\rho(W')}{f_\pi(W')} = \infty$. The singular point W' in $\text{supp}(\rho)$ will cause the upper integral of $\int_{\text{supp}(\rho)}$ to be infinity.

On the Definition of Distribution Covering.

Definition 2.2 allows that $\text{supp}(\mathcal{P}) \setminus \mathcal{M}_k$ be nonempty but has measure zero. This imposes fewer constraints when defining $\mathcal{P}$, as $\text{supp}(\mathcal{P})$ can contain matrices with rank less than k, which do not lie in $\mathcal{M}_k$ but typically have measure zero. For example, in Example 2.3, the Gaussian $\mathcal{Q}$ covers $\text{supp}(\mathcal{Q}) \cap \mathcal{M}_k$ since $\mathcal{Q}(\text{supp}(\mathcal{Q}) \cap \mathcal{M}_k) = 1$, implying that $\mathcal{Q}(\text{supp}(\mathcal{Q}) \setminus \mathcal{M}_k) = 0$. But $\text{supp}(\mathcal{Q}) \setminus \mathcal{M}_k$ is nonempty: since the density on $\mathsf{Q} = 0$ is positive, the density on $\mathsf{W} = 0$ is also positive, so $0 \in \text{supp}(\mathcal{Q})$; since $0 \notin \mathcal{M}_k$, we know that $0 \in \text{supp}(\mathcal{Q}) \setminus \mathcal{M}_k$.

On Estimating KL Divergence with an Upper Bound.

In Section 4, since $KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}})$ has no closed from, we estimate it with an upper bound $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ that has closed form. This is a common approach in estimating the KL divergence of a Gaussian mixture (Hershey & Olsen, 2007; Durrieu et al., 2012). Note that the gap $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}}) - KL(f_{\mathsf{X}} \,||\, g_{\mathsf{X}})$ is not directly computable, but can be indirectly estimated through $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$, as $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ being small implies that the gap is small. Tables 1 and 2 show that $KL(f_{\mathsf{XC}} \,||\, g_{\mathsf{XC}})$ can be small with appropriate hyperparameter choice.

On Low-Rank Forms: PQ versus $\begin{bmatrix} A & B \\ C & CA^{-1}B \end{bmatrix}$.

Rank-k matrices are most commonly expressed in the form PQ where $P \in \mathbb{R}^{m \times k}$ and $Q \in \mathbb{R}^{k \times n}$. Under this form, a natural way to construct a random rank-k matrix is to assign Gaussian distributions on P and Q respectively, yielding PQ. This approach has been studied by (Alquier, 2013): since the distribution of PQ is complex, they simplify its analysis by introducing unspecified constants to characterize only its tail behavior (a common approach in random matrix analysis, see (Vershynin, 2012)), and additionally assumes that the norms of P and Q are bounded by a constant; in these ways they obtain a KL divergence for PQ. There are several issues in this approach: first, their KL divergence expression contains unspecified constants, making it not directly computable; second, the distribution of PQ does not cover $\mathcal{M}_k$ due to the norm truncation; third, their low-rank assumptions are too specific to be applied to a broader practical low-rank models like classic LoRA, which requires the distribution to cover $\mathcal{M}_k$.

The low-rank form $\begin{bmatrix} A & B \\ C & CA^{-1}B \end{bmatrix}$ for invertible $A \in \mathbb{R}^{k \times k}$ and arbitrary $B \in \mathbb{R}^{k \times (n-k)}$ and $C \in \mathbb{R}^{(m-k) \times k}$ comes from the differential-geometric analysis of the low-rank manifold $\mathcal{M}_k$ by (Uschmajew & Vandereycken, 2020) (Proposition 2.1). We find this form much easier to construct an analyzable low-rank distribution: as shown in our work, we assign Gaussians on A, B, C, yielding a wrapped Gaussian matrix $\begin{bmatrix} \mathsf{A} & \mathsf{B} \\ \mathsf{C} & \mathsf{CA}^{-1}\mathsf{B} \end{bmatrix}$. This wrapped Gaussian has nice properties that both its density and KL divergence can be expressed explicitly without using uspecified constants or norm truncation. Moreover, it is associated with the geometry of $\mathcal{M}_k$ by corresponding to one chart in a multi-chart covering of $\mathcal{M}_k$. In contrast, the distribution of PQ is obtained without considering the geometry of $\mathcal{M}_k$.

It is difficult for a low-rank matrix in the PQ form to utilize the differential-geometric properties of $\mathcal{M}_k$ to simplify the analysis, since these properties are directly derived from the $\begin{bmatrix} A & B \\ C & CA^{-1}B \end{bmatrix}$ form, rather than the PQ form.

Hence, we consider the wrapped Gaussian a more suitable statistical modeling for practical low-rank models like classic LoRA, since it is easier to analyze and depends less on specific assumptions.

On the Heavy-tailed Nature of the Wrapped Gaussian Mixture.

We show that the wrapped Gaussian mixture X is heavy-tailed and has no mean. In our construction in Section 4, in each chart, $\mathsf{Y}|\mathsf{C}$ is a Gaussian on Euclidean space, and $\mathsf{X}|\mathsf{C}$ is a wrapped Gaussian on $\mathcal{M}_k$. We show that $\mathsf{X}|\mathsf{C}$ has no mean. Taking

$$\mathbb{E}[\mathsf{X}|\mathsf{C}] = \mathbb{E}[(G \circ H_{IJ})^{-1}(\begin{bmatrix}\mathsf{A} & \mathsf{B}\\ \mathsf{C} & 0\end{bmatrix})] = \mathbb{E}[H_{IJ}^{-1}(\begin{bmatrix}\mathsf{A} & \mathsf{B}\\ \mathsf{C} & \mathsf{C}\mathsf{A}^{-1}\mathsf{B}\end{bmatrix})]$$
$$= H_{IJ}^{-1}(\mathbb{E}\begin{bmatrix}\mathsf{A} & \mathsf{B}\\ \mathsf{C} & \mathsf{C}\mathsf{A}^{-1}\mathsf{B}\end{bmatrix}) = H_{IJ}^{-1}(\begin{bmatrix}\mathbb{E}[\mathsf{A}] & \mathbb{E}[\mathsf{B}]\\ \mathbb{E}[\mathsf{C}] & \mathbb{E}[\mathsf{C}]\mathbb{E}[\mathsf{A}^{-1}]\mathbb{E}[\mathsf{B}]\end{bmatrix})$$

In the third equality, we swap $\mathbb{E}$ and H_{IJ}^{-1} because the permutation function H_{IJ}^{-1} is linear. The last equality gives $\mathbb{E}[\mathsf{C}\mathsf{A}^{-1}\mathsf{B}] = \mathbb{E}[\mathsf{C}]\mathbb{E}[\mathsf{A}^{-1}]\mathbb{E}[\mathsf{B}]$ since we assume that A, B and C are independent Gaussian matrices. Note that $\mathbb{E}[\mathsf{A}^{-1}]$ does not exist (i.e., is infinite): in the simplest case where A is a 1×1 matrix, A^{-1} follows a reciprocal Gaussian distribution, which is heavy tailed and has no mean. Therefore, $\mathbb{E}[\mathsf{X}|\mathsf{C}]$ does not exist for any chart $\mathsf{C} = i$. Since $\mathbb{E}[\mathsf{X}] = \mathbb{E}_{\mathsf{C}}[\mathbb{E}[\mathsf{X}|\mathsf{C}]]$, we know that $\mathbb{E}[\mathsf{X}]$ does not exist, so X is heavy-tailed.

When converting a fixed weight matrix W_0 to a stochastic weight matrix W, (Dziugaite & Roy, 2017) lets W follow a Gaussian distribution satisfying $\mathbb{E}[\mathsf{W}] = W_0$, such that samples from W are near W_0 with high probability. In general, the distribution of W does not need to be Gaussian, and it does not even need to have a mean. All we need is that the distribution has some adjustable parameters that allow the density to concentrate around W_0. The Cauchy distribution and the wrapped Gaussian mixture fall into this category.

On the Minimum Number of Charts Covering $\mathcal{M}_k$.

In Section 3, we have shown that $\mathcal{M}_k$ can be covered by charts $(\mathcal{M}_k \cap \mathcal{U}_i, G \circ H_i)$ for $i \in \{1, 2, ..., \binom{m}{k}\binom{n}{k}\}$. And in Section 4 we have explained that the number of charts $\binom{m}{k}\binom{n}{k}$ can significantly increase the computational complexity of KL divergence unless parameter sharing is used. Regarding this issue, several questions arise: (a) Can $\mathcal{M}_k$ be covered by fewer charts? (b) What is the minimum number of charts to cover $\mathcal{M}_k$? (c) Can $\mathcal{M}_k$ be covered by one chart, i.e., it is diffeomorphic to a subset of $\mathbb{R}^{(m+n-k)k}$?

The questions (b) and (c) are challenging if the open sets used to cover $\mathcal{M}_k$ can be arbitrary. For (b), if we additionally assume that all open sets are contractible, then the minimum number of charts is known as the Lyusternik-Schnirelmann(LS) category (James, 1978). For (c), if $\mathcal{M}_k$ itself is contractible, then its LS category is 1, meaning that it can be covered by one chart.

So far, regarding (c), we have not yet found a proof showing whether $\mathcal{M}_k$ is contractible. Regarding (a), we have not yet found an alternative chart covering $\{(S_i, \psi_i)\}_{i=1}^t$ for $\mathcal{M}_k$. It is possible that such a covering with $t < \binom{m}{k}\binom{n}{k}$ exists, but the corresponding diffeomorphisms ψ_i may not have explicit expressions. We are not certain about the answer yet, so we leave these questions as open problems for future work.