Convergent Stochastic Training of Attention and Understanding LoRA

Anonymous Authors[1]

Abstract

Transformers have revolutionized machine learning and deploying attention layers in the model is increasingly standard across a myriad of applications. Further, for large models, it is common to implement Low Rank Adaptation (LoRA), whereby a factorized parameterization of them is trained, to achieve a surprisingly beneficial accuracy-size trade-off. In this work, via a unified framework we rigorously establish trainability of such models under stochastic methods. We prove that for any mild regularization, the empirical regression loss on a attention layer and LoRA on a shallow neural net, both induce Poincaré inequality for the corresponding Gibbs's measure. Then it follows via invoking recent results that a certain SDE, which mimics the SGD, minimizes the corresponding losses. In both the cases, our first-of-its-kind results of trainability on attention and nets, do not rely on any assumptions on the data or the size of the architecture.

1. Introduction

The remarkable empirical success of attention mechanisms have fundamentally reshaped modern machine learning, most prominently through the transformer architecture – which is the backbone of Large Language Models (LLMs) (Radford et al., 2018). By allowing models to dynamically weight interactions between "tokens"/data fragments, attention layers enable modeling of complicated distributions. Central to the attention mechanism are the query and key matrices, $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ which occur in the model as the product $\boldsymbol{W}_Q\boldsymbol{W}_K$ and one such pair are trained in each "attention head" of which there are many in each of the many attention layers in any commonly used transformer. Despite their central role in the practice of modern AI, our theoretical understanding remains limited of how these matrices evolve during successful training.

While modern attention mechanisms are now the standard, they originally emerged to address the bottleneck of fixed-length representations in early sequence modeling. Early sequence-to-sequence models were limited by fixed-length representations. (Bahdanau et al., 2016) introduced attention to dynamically aggregate encoder states, allowing the decoder to focus on relevant input positions. Building on this, (Luong et al., 2015) proposed alternative attention variants, including global attention over all positions and local attention over a subset.

These developments ultimately led to the transformer architecture (Vaswani et al., 2017), which removes recurrence entirely and instead models sequence interactions purely through stacked self-attention layers, where given an input $\boldsymbol{X} \in \mathbb{R}^{t\times d}$ each layer computes $\mathrm{RowSoftMax}_\beta\left(\frac{\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top}{\sqrt{d}}\right)\boldsymbol{X}\boldsymbol{W}_V$, where $\boldsymbol{W}_Q, \boldsymbol{W}_K$ and $\boldsymbol{W}_V$ are learned weight matrices where the RowSoftMax operator is defined as $[\mathrm{RowSoftMax}_\beta(\boldsymbol{M})]_{ij} \coloneqq \frac{e^{\beta M_{ij}}}{\sum_{k=1}^t e^{\beta M_{ik}}}$.

The transformer architecture (Vaswani et al., 2017) has become the dominant paradigm for sequence modeling, replacing recurrence with stacked self-attention and feed-forward layers. Numerous variants have since been proposed, namely Sparse Transformer (Child et al., 2019), Longformer (Beltagy et al., 2020), Linformer (Wang et al., 2020), Transformer-XL (Dai et al., 2019), ALBERT (Lan et al., 2020), and Vision Transformer (Dosovitskiy et al., 2021), among many others. A foundational reason explaining these successes is given in works like (Yun et al., 2020) that have show that transformers are universal approximators of sequence-to-sequence matrix functions.

The Rising Importance of Doing Regression on Transformers The success of transformers has led to their widespread adoption in scientific machine learning, where many tasks can be naturally formulated as regression problems over high-dimensional discretizations of function spaces. In particular, applications in fluid dynamics and weather prediction — ranging from operator learning for PDEs to data-driven forecasting — often reduce to learning mappings between input and output fields by doing regression using a attention-based architecture, FourCastNet (Pathak et al., 2022), GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), Poseidon (Herde et al., 2024) and GenCFD (Molinaro et al., 2024). This perspec-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

tive motivates studying attention not only as a representation mechanism, but as a regression operator whose properties govern generalization and efficiency in continuous domains.

Q1: *With no assumptions on data or architecture, can a attention layer be trained by a stochastic algorithm?*

On the other hand, the rise of large-scale pretraining has motivated parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA), which constrains updates to lie in low-dimensional subspaces while preserving the pretrained backbone. More precisely, the idea is to freeze the pre-trained weight matrix $\boldsymbol{W}_{\text{pre}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and only train a small update. This update is represented as the product of two small matrices, $\boldsymbol{A}$ and $\boldsymbol{B}$. The modified weights $\boldsymbol{W}'$ are parameterized as, $\boldsymbol{W}' = \boldsymbol{W}_{\text{pre}} + \frac{\alpha}{r}\boldsymbol{B}\boldsymbol{A}$ where $\boldsymbol{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ and $\boldsymbol{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ are the trainable factors and α being a scaling constant. By training only $\boldsymbol{A}$ and $\boldsymbol{B}$, we significantly reduce the number of parameters to update.

As showed in (Shuttleworth et al., 2025), LoRA works best when applied to all weight matrices and small-to-medium datasets, and its optimal learning rate is largely independent of rank due to the $\frac{1}{r}$ scaling. Overall, LoRA performs similar to full-finetuning in the "low-regret regime" making it a parameter-efficient alternative for post-training adaptation. However, the product $\boldsymbol{B}\boldsymbol{A}$ introduces a specific scaling problem : we can multiply $\boldsymbol{A}$ by a constant and divide $\boldsymbol{B}$ by the same constant without changing the final output. This redundancy makes the training landscape "flat" in certain directions, which is seemingly a natural obstacle for gradient based algorithms to succeed. And yet LoRA has proven strikingly effective in practice, dramatically reducing memory and computational overhead while retaining performance. Thus we posit that the optimization dynamics of neural models under LoRA have remained unclear from a theoretical standpoint.

Q2: *With no assumptions on data or width, can a implementation of LoRA on a net be trained by a stochastic algorithm?*

In this work, *firstly* we make progress towards uncovering hitherto unknown mathematical properties about the attention map and thus uncover a first-of-its-kind provably convergent training mechanism for the query and key matrices. *Secondly,* this work also initiates a theoretical study of training dynamics for standard neural networks under LoRA parameterization — which evidently shares a mathematical similarity of training a factorized weight parameterization as while training the key and query matrices in the attention, as introduced above.

Whether for training the key and query matrices of an attention head or for training standard nets under LoRA, we consider training through stochastic differential equations (SDEs), capturing the continuous-time limit of stochastic gradient methods commonly used in practice. *In both cases — for any number of parameters and for any data — we prove convergence of a risk function for the models under a SDE flow.*

1.1. Summary of Results

In this work, we establish that for both attention and depth-2 neural networks under LoRA, a mildly regularized regression loss (say $\tilde{V}$) is a Villani function (which will be precisely defined in the next section) — which in turn implies that the corresponding Gibbs' measure ($\sim e^{-\gamma \tilde{V}}$) satisfies the Poincare inequality. Then invoking recent results on isoperimetry based Stochastic Differential Equation (SDE) convergence (Shi et al., 2023) we can establish convergence in both settings for the following continuous-time stochastic gradient dynamics for the weights $\boldsymbol{T}$, given by the SDE,

$$\mathrm{d}\boldsymbol{T}_t = -\nabla \tilde{V}(\boldsymbol{T}_t)\,\mathrm{d}t + \sqrt{s}\,\mathrm{d}B_t\,, \tag{1}$$

where $\tilde{V}(\boldsymbol{T})$ denotes the regularized loss function, $s > 0$ is a temperature parameter and $(B_t)_{t\geq 0}$ is the Brownian motion.[1] We provide an informal restatement of our main results below.

Theorem 1.1 (Informal Statement of Provable Learning for Attention-Based Regression)**.** *Consider a single attention layer with key and query matrices $\boldsymbol{W}_K$ and $\boldsymbol{W}_Q$, being trained using the ℓ_2-loss function with either a logarithmically amplified $2-$norm regularization or a super-quadratic polynomial regularization. Then, for any arbitrarily low regularization, for any data and size of architecture, the loss function satisfies the Villani condition.*

As a consequence, for any $\varepsilon > 0$, there exists an appropriate step size s such that the SDE for $\boldsymbol{T} = (\boldsymbol{W}_Q, \boldsymbol{W}_K)$ converges, in expectation, in $\mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$ to within ε of the global minimum of the training loss.

The above informal restatement combines Theorem 3.1 with the relevant part of Theorem 3.3.

Theorem 1.2 (Informal Statement of Provable Learning for Depth-2 Neural Net Based Regression under LoRA)**.** *Consider a depth-2 neural network with weight matrix $\boldsymbol{W}$ factorized as $\boldsymbol{U}\boldsymbol{V}$, being trained using the ℓ_2-loss function with either a logarithmically amplified $2-$norm regularization or a super-quadratic polynomial regularization. Then, for any arbitrarily low regularization, for any data and size of architecture, the loss function satisfies the Villani condition.*

As a consequence, for any $\varepsilon > 0$, there exists an appropriate step size s such that the SDE for $\boldsymbol{T} = (\boldsymbol{U}, \boldsymbol{V})$ converges, in expectation, in $\mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$ to within ε of the global minimum of the training loss.

[1] The subscript t denotes continuous time.

The above informal restatement combines Theorem 3.2 with the relevant part of Theorem 3.3.

Remark 1.3. As shown in (Kumar et al., 2025), Theorems 1.1 and 1.2 can be extended to establish convergence to the global minima of the population risk, defined as $\mathbb{E}_{\mathcal{S}_n}[V_{\mathcal{S}_n}(\boldsymbol{T})]$, where $V_{\mathcal{S}_n}(\boldsymbol{T})$ denotes the loss evaluated on the dataset $\mathcal{S}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$, under the Langevin Monte Carlo algorithm.

1.2. Literature Review

A theoretical analysis of transformers in the infinite-width limit by deriving their Neural Network Gaussian Process (NNGP) and Neural Tangent Kernel (NTK) equivalents was provided in (Hron et al., 2020). They showed that self-attention layers admit well-defined kernel limit and in this regime, gradient descent training of a Transformer is equivalent to kernel regression with the corresponding NTK.

A global convergence framework for transformers by analyzing training dynamics in the mean-field limit was established in (Gao et al., 2024). By treating model width and depth as approaching infinity, the authors demonstrate that discrete gradient descent converges to a Wasserstein gradient flow on the distribution of parameters. Albeit the use of an infinite-width limit it strictly requires the inclusion of a weight decay parameter $\lambda > 0$, for the convergence proof to work.

A formal proof that transformers can converge to the functional behavior of near-optimal Reinforcement Learning (RL) algorithms through the optimization of a log-likelihood objective was given by (Lin et al., 2023). By viewing the attention mechanism as an iterative optimizer, they demonstrate that supervised pre-training on offline trajectories allows the model to implement algorithms like LinUCB and Thompson Sampling directly. However, these convergence guarantees rely on non-standard architectures, most notably the use of ReLU-based attention to facilitate exact linear algebraic operations. Furthermore, the authors assume model realizability, implying that the Transformer's capacity must be sufficient to encapsulate the expert's decision-making logic.

A multi-layer transformer trained on n-gram data, where each token depends on the preceding n tokens, is analyzed in (Chen et al., 2024), and it is shown that gradient flow converges to a model exhibiting induction head behavior. Specifically, induction here refers to the phenomenon whereby, if a token at position i matches a previous occurrence at position j, the model attends to the token following position j to predict the next token at position i. Their results provide a rigorous characterization of how attention layers, feed-forward networks, and normalization interact to learn features from context. This work advances prior studies that focused on linear or single-layer models by handling richer architectures and more realistic data distributions.

Fundamental algorithmic limits of Multi-Head Attention (MHA) restricted to a discrete Boolean input distribution $X \in \{\pm 1\}^{k\times d}$ was established in (Chen & Li, 2025). Theorem 1.2 in (Chen & Li, 2025) establishes that, under a non-degeneracy condition on the attention and projection matrices and realizability assumption of the samples, there exists an algorithm that for $m-$headed attention estimates the parameters in $(kd)^{O(m^3)}$ time, using $(kd)^{\Theta(m)}$ samples, and achieves predictions that are $(kd)^{-\Omega(m)}$ close to the true values in expectation. This result identifies the number of heads m as the dominant factor in computational scaling.

1.2.1. Review of Existing Attempts at Provable Training of LoRA on Neural Networks — with Weight Regularization

LoRA was first introduced by (Hu et al., 2022) where it was asserted that the weight updates for task-specific adaptation in attention based models reside in a manifold of low intrinsic dimension (Aghajanyan et al., 2021). By reparameterizing the update matrix $\Delta\boldsymbol{W}$ as the product of two low-rank matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, (Hu et al., 2022) demonstrated that optimization can converge to high-performance solutions with significantly fewer trainable parameters. Crucially, their initialization strategy — setting one matrix to zero — ensures a stable starting point at the pre-trained state, effectively bridging the gap between training efficiency and the convergence stability typically observed in full fine-tuning.

A rigorous analysis of LoRA on neural networks in the generic non-linear regime was given by (Kim et al., 2025), establishing a formal dichotomy between global convergence and parameter divergence. By characterizing the optimization as governed by a global Restricted Strong Convexity, they prove that the combination of zero-initialization and weight decay induces an implicit bias toward a low-rank global minimum. Crucially, they demonstrate that while the non-linear landscape may harbor spurious local minima, these points are spectrally isolated in high-rank regions and do not intersect with the stable optimization trajectory.

In the Neural Tangent Kernel (NTK) framework LoRA was analyzed by (Jang et al., 2024), where the neural network's optimization can be treated as a linearized system. In this regime, the training dynamics are governed by a quadratic objective subject to a low-rank structural constraint. Alongside a non-standard regularizer — derived from the Rademacher complexity of the low-rank bottleneck — the authors prove that the non-convex BA reparameterization does not introduce spurious local minima. (Jang et al., 2024) further demonstrates, if the rank is above a certain threshold, gradient-based methods like stochastic gradient descent (SGD) converge to a low-rank global minimizer.

Limitations of LoRA It was demonstrated by (Shuttleworth et al., 2025) that the perceived equivalence between LoRA and full fine-tuning is an "illusion" maintained by surface-level metrics. Despite its parameter efficiency, LoRA can lead to catastrophic forgetting, thereby degrading performance in settings that require continual learning.

1.2.2. Organization

Section 2 introduces the analytic framework, including the underlying conditions, neural architecture, loss functions, and the assumptions required for the subsequent sections. Section 3 presents the main results of the paper, namely Theorem 3.1, Theorem 3.2, and Theorem 3.3. The proof of Theorem 3.1, along with the requisite auxiliary lemmas, is provided in Section A, while the detailed proofs of these lemmas are deferred to Section B. Section 4 presents experimental results on solving the 2D Darcy flow problem using the regularized loss functions introduced in Section 2. The proof of Theorem 3.2 is contained in Appendix C, with the corresponding supporting lemmas established in Appendix D. Lastly, Appendix E presents the proof of Theorem 3.3.

2. Mathematical Setup

In this section, we define the analytic conditions, architectures, and loss functions, alongside the core assumptions that underpin our subsequent analysis.

The Villani condition was introduced in (Villani, 2009) to guarantee that, when a function satisfies this condition, the associated Gibbs measure satisfies the Poincaré inequality. We recall that a distribution π is said to satisfy the Poincaré inequality for some constant C_{PI}, if for all smooth functions $h : \mathbb{R}^d \to \mathbb{R}$, $\mathrm{Var}_\pi(h) \le C_{PI}\mathbb{E}_\pi[\|\nabla h\|^2]$. (Shi et al., 2023) leverage the Poincaré inequality induced by the Villani condition to establish convergence results for certain stochastic differential equations (SDEs). We now proceed to formally define the corresponding analytic conditions.

Definition 2.1 (Confining Condition). A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be confining if it satisfies the following conditions, **(1.)** $f \in C^\infty$, **(2.)** $\lim_{\|\boldsymbol{x}\|\to+\infty} f(\boldsymbol{x}) = +\infty$, and **(3.)** $\int_{\mathbb{R}^d} e^{-\frac{2f(\boldsymbol{x})}{s}}\,\mathrm{d}\boldsymbol{x} < \infty \ \forall s > 0$.

Definition 2.2 (Villani Condition). A confining function f is said to satisfy the Villani condition if for all $s > 0$, $\frac{\|\nabla f(\boldsymbol{x})\|^2}{s} - \Delta f(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$:

In recent works (Gopalani et al., 2024; Gopalani & Mukherjee, 2025), it was shown that the Villani condition holds for depth-2 neural networks of arbitrary width and for both squared and logistic losses. Building on this, (Kumar et al., 2025) proved that Langevin Monte Carlo consequently achieves population risk minimization. In contrast to (Gopalani et al., 2024; Gopalani & Mukherjee, 2025; Kumar et al., 2025), our results do not require any lower bound on the regularization parameter.

Next we formally define the attention model and its regression loss that we choose to train.

Definition 2.3 (Attention Layer).

$$\begin{aligned}&\mathbb{R}^{t\times d} \ni \boldsymbol{X} \mapsto \mathrm{Attention}(\boldsymbol{X}) \\ &\coloneqq \mathrm{RowSoftMax}_\beta\left(\frac{\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top}{\sqrt{d}}\right)\boldsymbol{X}\boldsymbol{W}_v \in \mathbb{R}^{t\times d}\end{aligned} \tag{2}$$

where, $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d\times r}$ and $\boldsymbol{W}_V \in \mathbb{R}^{d\times d}$

and for any $\boldsymbol{M} \in \mathbb{R}^{t\times t}$,

$$[\mathrm{RowSoftMax}_\beta(\boldsymbol{M})]_{ij} = \frac{e^{\beta\boldsymbol{M}_{ij}}}{\sum_{k=1}^t e^{\beta\boldsymbol{M}_{ik}}}$$

A notable application of this framework is the Vision Transformer (ViT) (Dosovitskiy et al., 2021). In ViT, the input image is first divided into non-overlapping patches, commonly of size 16×16 pixels. Each patch is flattened and projected into a d-dimensional embedding, producing a sequence of t token embeddings that serve as input to the attention layers described above. This allows the model to capture long-range dependencies across the image while leveraging the same attention mechanism as in general Transformer architectures. The row-wise softmax scaling parameter β is typically set to 1.

We train the above model we consider two forms of factor-regularized potentials/loss functions for it : a non-polynomial/logarithmically amplified $2-$norm regularization and a polynomial regularization with exponent $2+\epsilon$. The following definitions formalize the corresponding regularized potentials used in our analysis.

Definition 2.4 (**Mean Square Loss on a Attention Layer with Non-Polynomial Factor-Regularization**). We define the potential $V_{\mathrm{ATT}}(\boldsymbol{T})$ in the factor space $\mathbb{R}^D$ for $\boldsymbol{T} = (\boldsymbol{W}_Q, \boldsymbol{W}_K)$ and $D = 2dr$ as,

$$\begin{aligned}\tilde{V}_{\mathrm{ATT}}(\boldsymbol{T}) &\coloneqq \hat{R}_A(\boldsymbol{T}) \\ &+ \frac{\lambda}{2}(\|\boldsymbol{W}_Q\|_F^2 + \|\boldsymbol{W}_K\|_F^2)\log\left(1 + \|\boldsymbol{W}_Q\|_F^2 + \|\boldsymbol{W}_K\|_F^2\right)\end{aligned} \tag{3}$$

where $\hat{R}_A(\boldsymbol{T}) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{W}_Q, \boldsymbol{W}_K)$ and $\ell_i(\boldsymbol{W}_Q, \boldsymbol{W}_K) \coloneqq \frac{1}{2}\left\|\boldsymbol{Y}_i - \mathrm{RowSoftMax}_\beta\left(\frac{\boldsymbol{X}_i\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}_i^\top}{\sqrt{d}}\right)\boldsymbol{X}\boldsymbol{W}_v\right\|_F^2$ corresponding to a choice of training data as, $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathbb{R}^{t\times d} \times \mathbb{R}^{t\times d} \mid i = 1, \ldots, n\}$.

Definition 2.5 (**Mean Square Loss on a Attention Layer with Polynomial Factor-Regularization**). We define the potential $V_{\epsilon,\mathrm{ATT}}(\boldsymbol{T})$ in the factor space $\mathbb{R}^D$ for $\boldsymbol{T} =$

$(\boldsymbol{W}_Q, \boldsymbol{W}_K)$ and $D = 2dr$ as,

$$V_{\epsilon,\mathrm{ATT}}(\boldsymbol{T}) \coloneqq \hat{R}_A(\boldsymbol{T}) + \frac{\lambda}{2}(\|\boldsymbol{W}_Q\|_F^{2+\epsilon} + \|\boldsymbol{W}_K\|_F^{2+\epsilon}) \quad (4)$$

where $\hat{R}_A(\boldsymbol{T})$ as given in Definition 2.4.

For the case of shallow neural networks, we assume a training in a space of weight matrices with a rank bound i.e we assume the trainable weight to be factorizable as, $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top$, with $\boldsymbol{U} \in \mathbb{R}^{p\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times r}$. This factorization implements low-rank adaptation (LoRA) approach for depth-2 nets — and we recall that this has been previously studied in (Jang et al., 2024; Kim et al., 2025) as a theoretical sandbox for the LoRA technique. As in the attention-based setting, regularization of the factor matrices is introduced to ensure well-behaved potentials in the weight space. Similar to the attention training setup above, we define two types of factor-regularized loss functions as follows,

Definition 2.6 (**Rank-Restricted Mean Square Loss on Shallow Nets with Non-Polynomial Factor-Regularization**). We define the potential $V(\boldsymbol{T})$ in the factor space $\mathbb{R}^D$ for $\boldsymbol{T} = (\boldsymbol{U}, \boldsymbol{V})$ and $D = (p+d)r$ as,

$$V(\boldsymbol{T}) \coloneqq \mathcal{L}(\boldsymbol{T}) + \frac{\lambda}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2)\log\left(1 + \|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right) \quad (5)$$

where $\mathcal{L}(\boldsymbol{T}) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{U}\boldsymbol{V}^\top)$ and $\ell_i(\boldsymbol{W}) \coloneqq \frac{1}{2}(y_i - \boldsymbol{a}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i))^2$.

Definition 2.7 (**Rank-Restricted Mean Square Loss on Shallow Nets with Polynomial Factor-Regularization**). We define the potential $V_\epsilon(\boldsymbol{T})$ in the factor space $\mathbb{R}^D$ for $\boldsymbol{T} = (\boldsymbol{U}, \boldsymbol{V})$ and $D = (p+d)r$ as,

$$V_\epsilon(\boldsymbol{T}) = \mathcal{L}(\boldsymbol{T}) + \frac{\lambda}{2}(\|\boldsymbol{U}\|_F^{2+\epsilon} + \|\boldsymbol{V}\|_F^{2+\epsilon}) \quad (6)$$

where $\mathcal{L}(\boldsymbol{T}) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{U}\boldsymbol{V}^\top)$ and $\ell_i(\boldsymbol{W}) \coloneqq \frac{1}{2}(y_i - \boldsymbol{a}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i))^2$.

To establish convergence results for the factorized, regularized losses defined above, we impose a few standard assumptions on the network and define certain properties of the training data.

Definition 2.8 (Training Data Bounds). The training data is bounded as follows:

1. For Attention : Each training example $(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathbb{R}^{t\times d} \times \mathbb{R}^{t\times d}$ satisfies $\|\boldsymbol{X}_i\|_F \le B_x, \|\boldsymbol{Y}_i\|_F \le B_y, i = 1, \ldots, n$.
2. For Neural Network : Each training example $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ satisfies $\|\boldsymbol{x}_i\|_2 \le B_x, |y_i| \le B_y, i = 1, \ldots, n$.

Depending on the architecture, the bounds B_x and B_y are interpreted according to the corresponding definitions given above.

Definition 2.9 (Attention-Specific Bound). For regression tasks on attention-based models, the weight matrix $\boldsymbol{W}_v$ is also bounded as $\|\boldsymbol{W}_v\|_F \le B_w$.

Assumption 2.10 (Activation Function Bounds). We assume that σ', σ' and σ'' are bounded by $\sup_{t\in\mathbb{R}}|\sigma(t)| = B_\sigma < \infty$, $\sup_{t\in\mathbb{R}}|\sigma'(t)| = B_{\sigma'} < \infty$ and $\sup_{t\in\mathbb{R}}|\sigma''(t)| = B_{\sigma''} < \infty$, respectively.

We can then characterize the global convergence of the SDE (1), motivated by (Shi et al., 2023) in their analysis of SGD on non-convex landscapes, for our regularized loss function for depth-2 nets under LoRA constraints and for the attention-based model. The SDE is modeled over $\boldsymbol{T}$. The corresponding invariant Gibbs measure $\mu_s(\mathrm{d}\boldsymbol{T})$ is defined as

$$\mu_s(d\boldsymbol{T}) \coloneqq \frac{1}{Z_s}\exp\left(-\frac{2}{s}\tilde{V}(\boldsymbol{T})\right)d\boldsymbol{T}, \quad (7)$$

where $\tilde{V}(\boldsymbol{T})$ represents any of the factor-regularized potentials defined above, $s > 0$ acts as the temperature parameter (proportional to the learning rate), and Z_s is the normalization constant.

3. Main Results

Given the formal setup in the previous section, we first state our key result showing that the regression loss functions associated with the softmax-attention layer, defined in Definitions 2.4 and 2.5, satisfy the Villani condition.

Theorem 3.1 (**Attention-Based Regression Loss is a Villani Function**). *Consider the regularized loss functions associated with the attention layer, $\tilde{V}_{\mathrm{ATT}}(\boldsymbol{T})$ and $V_{\epsilon,\mathrm{ATT}}(\boldsymbol{T})$, as defined in Definitions 2.4 and 2.5, respectively. Then, for any $\lambda, \epsilon > 0$, both $\tilde{V}_{\mathrm{ATT}}(\boldsymbol{T})$ and $V_{\epsilon,\mathrm{ATT}}(\boldsymbol{T})$, evaluated on the training data defined in Definition 2.8, satisfy the Villani condition (Definition 2.2).*

We next present our second key result, establishing that the regression loss functions for a depth-2 neural network with LoRA, as defined in Definitions 2.6 and 2.7, satisfy the Villani condition.

Theorem 3.2 (**Depth-2 Neural Net Based Regression Loss under LoRA is a Villani Function**). *Suppose that Assumptions 2.10 holds for the activation function σ. Consider the loss functions $\tilde{V}(\boldsymbol{T})$ and $V_\epsilon(\boldsymbol{T})$ for a depth-2 neural network with activation σ, as defined in Definitions 2.6 and 2.7, respectively. Then, for any $\lambda, \epsilon > 0$, both $\tilde{V}(\boldsymbol{T})$ and $V_\epsilon(\boldsymbol{T})$, evaluated on the training data defined in Definition 2.8, satisfy the Villani condition (Definition 2.2).*

By Theorems 3.1 and 3.2, all considered loss functions with their associated neural architectures satisfy the Villani condition. Consequently, we may invoke Theorem 1 of (Shi et al., 2023) to obtain the following convergence result for the SDE (1).

Theorem 3.3 (**Convergence of SDE for Depth-2 Neural Net Based Regression under LoRA and Attention-Based Regression**). *Suppose that Assumption 2.10 holds for the activation σ. Let $\tilde{L}^{(k)}(\boldsymbol{T})$ denote any of the four regularized potentials defined in Definition 2.4 ($\tilde{V}_{\mathrm{ATT}}$), Definition 2.5 ($V_{\epsilon,\mathrm{ATT}}$), Definition 2.6 ($V$), and Definition 2.7 ($V_\epsilon$), where $k \in \{1,2,3,4\}$ indexes the specific model and regularization choice. Suppose the initial probability density is $p_0 \in L^2((\mu_s^{(k)})^{-1})$ of the SDE (1) where $\mu_s^{(k)}$ is the corresponding Gibbs measure (7). For each $\tilde{L}^{(k)}(\boldsymbol{T})$ satisfying the Villani conditions, there exists a positive $\lambda_s^{(k)} > 0$ and a constant $D^{(k)}(s,p_0)$ such that:*

$$\mathbb{E}[\tilde{L}^{(k)}(\boldsymbol{T}_t)] - \tilde{L}^{(k)\star} \le \varepsilon^{(k)}(s) + D^{(k)}(s,p_0)e^{-\lambda_s^{(k)}t}, \quad (8)$$

where $\tilde{L}^{(k)\star} = \inf \tilde{L}^{(k)}(\boldsymbol{T})$ is the global minimum of the respective loss, and $\varepsilon^{(k)}(s) = \mathbb{E}_{\mu_s^{(k)}}[\tilde{L}^{(k)}(\boldsymbol{T})] - \tilde{L}^{(k)\star}$.

Then there exist constants $A^{(k)}, S^{(k)} > 0$ such that $\mathbb{E}_{\mu_s^{(k)}}[\tilde{L}^{(k)}(\boldsymbol{T})] - \tilde{L}^{(k)\star} \le A^{(k)}s$ for all $s \in (0, S^{(k)}]$. If we further choose the learning rate s such that $s \le \min\left\{\frac{\epsilon}{2A^{(k)}}, S^{(k)}\right\}$, and the time t satisfies, $t \ge \frac{1}{\lambda_s^{(k)}} \log\left(\frac{2D^{(k)}(s,p_0)}{\epsilon}\right)$, where $D^{(k)}(s,p_0) = C^{(k)}(s)\cdot\|p_0 - \mu_s^{(k)}\|_{L^2((\mu_s^{(k)})^{-1})}$ and $C^{(k)}(s)$ is a positive constant, then,

$$\mathbb{E}[\tilde{L}^{(k)}(\boldsymbol{T}_t)] - \tilde{L}^{(k)\star} \le \epsilon. \quad (9)$$

The proofs of the above are given in Section A.1, Appendix C and Appendix E, respectively.

We note that since λ and ϵ can be set to be arbitrarily small positive numbers for the above convergence, it follows that such a mild regularizer would have negligible effect at small/finite weight values w.r.t unregularized loss and that the regularization only appreciably affects the shape of the loss at infinity.

Remark 3.4 (**Necessity of Factor Regularization**). We note that the factorized loss function $\mathcal{L}(\boldsymbol{T}) = \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{U}\boldsymbol{V}^\top)$, where $\boldsymbol{T} = (\boldsymbol{U},\boldsymbol{V})$, exhibits a scaling invariance under the transformation, $g_{\boldsymbol{A}}(\boldsymbol{U},\boldsymbol{V}) = (\boldsymbol{U}\boldsymbol{A}, \boldsymbol{V}\boldsymbol{A}^{-1})$, $\boldsymbol{A} \in \mathbb{R}^{r\times r}$, $\det(\boldsymbol{A}) \neq 0$, since,

$$\mathcal{L}(\boldsymbol{U}\boldsymbol{A}, \boldsymbol{V}\boldsymbol{A}^{-1}) = \mathcal{L}(\boldsymbol{U},\boldsymbol{V}).$$

So the potential $\mathcal{L}(\boldsymbol{T})$ is constant along the non-compact orbits generated by the general linear group.

As a consequence, the Gibbs' measure $\mu_\gamma \propto e^{-\gamma\mathcal{L}(\boldsymbol{T})}$ is non-normalizable. Specifically, for any fixed rank-r matrix $\boldsymbol{W}_0 \neq 0$, $\mathcal{L}(\boldsymbol{T})$ is constant along the orbit $\mathcal{O} = \{(\boldsymbol{U}\boldsymbol{A}, \boldsymbol{V}\boldsymbol{A}^{-1}) \mid \det(\boldsymbol{A}) \neq 0\}$, which extends infinitely far from the origin. Hence, the partition function

$$Z = \int_{\mathbb{R}^D} e^{-\gamma\mathcal{L}(\boldsymbol{T})} d\boldsymbol{T}$$

diverges because it includes an integral of a non-zero constant density over an infinite-volume set. Consequently, the confining condition is violated, and the Poincaré Inequality cannot hold for the unregularized factorized loss.

4. An Empirical Study of Regularized Learning of Key and Query Matrices

Towards demonstrating an use of our regularized attention losses, we study the two-dimensional Darcy Flow PDE — which is popularly used as a benchmark in scientific-ML,

$$-\nabla\cdot\big(a(\mathbf{x})\,\nabla u(\mathbf{x})\big) = f(\boldsymbol{x}), \quad \mathbf{x} \in (0,1)^2, \quad (10)$$

where a is a spatially varying permeability (diffusion) coefficient, u is the unknown pressure field and f is a source function. The regression task is to learn a mapping $a \mapsto u$ from discretised input fields to discretised solution fields. Following the benchmark introduced by (Li et al., 2021), fields are discretised on a uniform 64×64 grid (bilinear-downsampled from the native 421×421 resolution). We use $N_{\mathrm{train}} = 900$ samples for training and $N_{\mathrm{test}} = 124$ held-out samples for evaluation. Both input and output fields are independently standardised using training-set mean and standard deviation.

Model Architecture

We use a patch-based single-head attention regressor defined as follows.

- (Tokenisation) Each 64×64 input field a is divided into non-overlapping 4×4 patches $\boldsymbol{p}_i$, yielding $t = (64/4)^2 = 256$ patches/tokens.
- (Embedding) Patches are projected to a d-dimensional representation via a two-stage convolutional encode: $\mathbf{x}_i = \mathrm{ConvEncoder}(\mathbf{p}_i)$, and then a flattening layer. $\mathbf{X}$ is then given by $[\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_t]^\top \in \mathbb{R}^{t\times d}$.
- (Positional encoding) Learnable 2-D positional bias $\boldsymbol{P}$ for conv-token grid is added to the patch embeddings: $\hat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{P}$.
- (Single-head attention) Query, key, and value projections are

$$\boldsymbol{Q} = \hat{\boldsymbol{X}}\boldsymbol{W}_Q, \quad \boldsymbol{K} = \hat{\boldsymbol{X}}\boldsymbol{W}_K, \quad \boldsymbol{V} = \hat{\boldsymbol{X}}\boldsymbol{W}_V,$$
$$\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d\times r}, \ \boldsymbol{W}_V \in \mathbb{R}^{d\times d}.$$

The attention map is computed with temperature fixed to 1 ($\beta = 1$), $\boldsymbol{A} = \mathrm{RowSoftMax}_\beta\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right) \in \mathbb{R}^{t\times t}$.

The representation $\mathrm{Attention}(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{V} \in \mathbb{R}^{t\times d}$ is projected back to patch space by a two-layer net and the resulting patches are rearranged into the predicted 64×64 output field $\hat{u}$. All dimensions are set to $r = d = 64$.

Two-Phase Training Protocol

Phase 1 (full pretraining): All parameters, embedding layers, projections $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$, and output MLP, are jointly optimised with plain mean-squared error (MSE) for 500 epochs using Adam (Kingma & Ba, 2014) with learning rate $\eta = 10^{-3}$ and batch size 32. This phase yields the *task-optimal* values for all weight matrices.

Phase 2 (ablation on ways of regularized training of $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$): The goal is to isolate the effect of norm regularisation on the query and key matrices. Apart from the key and the query weight all other parameters are set to their Phase 1 values. Then $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ are *re-initialised* identically across all three runs to their initial values used in Phase 1. Three objectives are then compared over 100 additional epochs,

1. **Unregularised:** (i.e., *none* in plots)

$$\mathcal{L}_0 = \hat{R}_*(\boldsymbol{T}),$$

2. **Log-amplified norm penalty:** (i.e., *log* in plots)

$$\mathcal{L}_{\log} = \hat{R}_*(\boldsymbol{T}) + \frac{\lambda}{2} S \log(1+S),\ \lambda > 0,$$
$$S = \|\boldsymbol{W}_Q\|_F^2 + \|\boldsymbol{W}_K\|_F^2.$$

3. **Super-quadratic norm penalty:** (i.e., *power* in plots)

$$\mathcal{L}_{2+\epsilon} = \hat{R}_*(\boldsymbol{T}) + \frac{\lambda}{2}\left(\|\boldsymbol{W}_Q\|_F^{2+\epsilon} + \|\boldsymbol{W}_K\|_F^{2+\epsilon}\right),\ \lambda, \epsilon > 0,$$

where $\boldsymbol{T} = (\boldsymbol{W}_Q, \boldsymbol{W}_K)$ and $\hat{R}_*(\boldsymbol{T})$ denotes the MSE loss with embedding layers, $\boldsymbol{W}_V$ and output MLP frozen at their optimal values.

In all Phase 2 runs we use Adam with $\eta = 10^{-3}$, batch size 32 (as in Phase-1), and hyper-parameters $\epsilon = 10^{-6}$, $\lambda = 10^{-5}$ for *log* and $\lambda = 10^{-4}$ for *power*.

Metrics

We track the following quantities per epoch and per run.

- **Train/Test RMSE:** $\sqrt{\mathbb{E}[\|\hat{u} - u\|^2]}$ in normalised target space.
- **Test Relative L2 Error:** $\varepsilon_{\mathrm{rel}} = \frac{\|\hat{u}-u\|_2}{\|u\|_2}$, the standard PDE-operator-learning benchmark metric.
- **Q/K Norm^2:** $S = \|\boldsymbol{W}_Q\|_F^2 + \|\boldsymbol{W}_K\|_F^2$.
- **Generalisation gap:** Δ = Test MSE – Train MSE.

4.1. Results analysis

In the very small regularization regime ($\epsilon = 10^{-6}$ and $\lambda = 10^{-4}$ or 10^{-5}), all methods, no regularization, "log" regularization, and "power" regularisation, exhibit nearly identical convergence behaviour in both normalised test/train RMSE and test relative L2 error. As shown in Figure 1, all configurations rapidly decay from their initial error and stabilise at approximately the same performance level (0.78), indicating that at this scale of regularization, there is no meaningful difference in predictive accuracy.

Despite the similarity in error metrics, substantial differences emerge in the internal dynamics of the model. In particular, the Frobenius norm ($\|\boldsymbol{W}_Q\|_F^2 + \|\boldsymbol{W}_K\|_F^2$) grows steadily in the absence of regularization, reaching significantly larger values over training. In contrast, both "log" and "power" regularization effectively constrain this growth, with power regularization enforcing the strongest suppression and log regularization yielding a slightly higher but still stable norm plateau. *Firstly,* these results demonstrate that even weak regularization of the kind studied here can meaningfully alter the scaling of the attention mechanism without impacting predictive performance.

In the generalization gap plot in Figure 1, all methods initially exhibit a rapid reduction in gap but the unregularized model shows a gradual increase over time, suggesting mild overfitting. Both log and power regularization mitigate this effect, maintaining a consistently lower and more stable gap throughout training, with power again providing the strongest control. *Secondly,* this result indicates that despite comparable test error improved generalization stability is induced by the regularizers being studied.

Overall, these findings suggest that the regularization studied in the presented theory primarily influences model stability. In low-λ regimes, log regularisation provides a robust default by stabilising the key and query weight norms with minimal intervention, while power regularisation is preferable when stronger suppression of attention weight growth is desired.

5. Conclusion

This work establishes that the SDE, that mimics the SGD, converges for both attention layers and depth-2 neural networks trained with LoRA, for arbitrary data and network sizes, even under arbitrarily low regularization, and we further provides non-asymptotic convergence rates.

A natural next question is whether the loss function on the attention layer with it's key-query-value matrices using LoRA, satisfy the Villani conditions in the corresponding

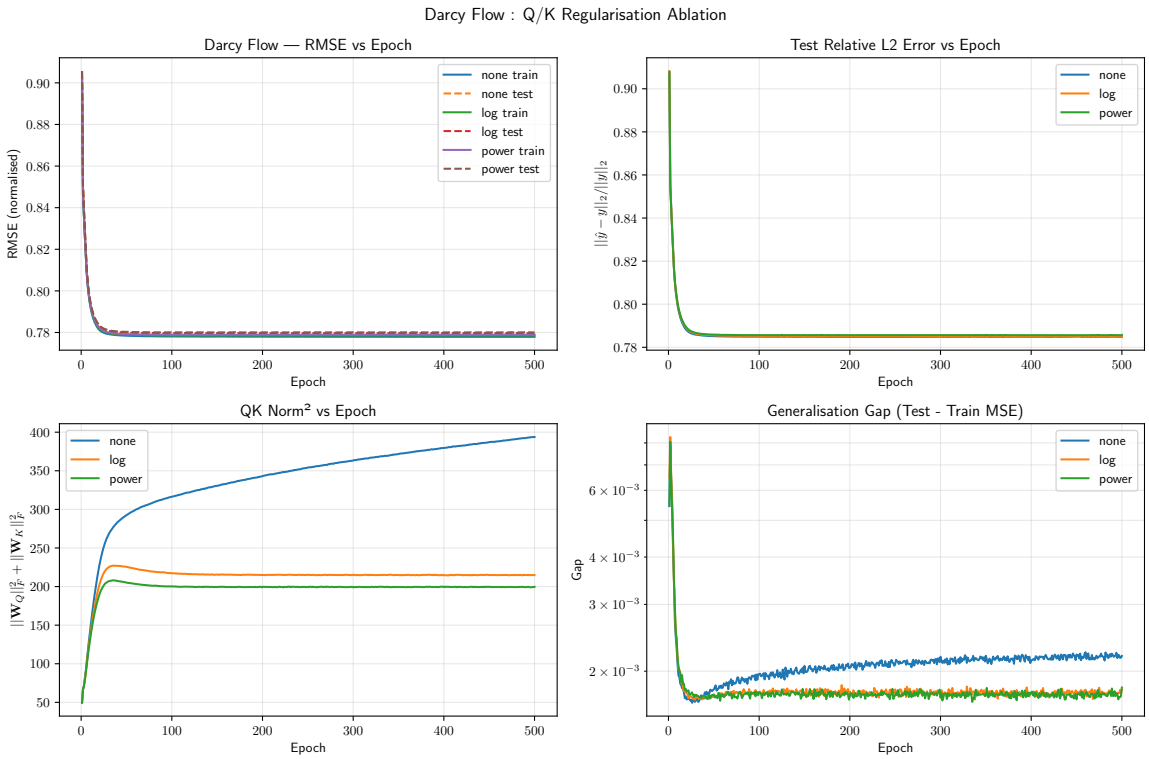

Figure 1. Per-epoch and per-run statistics for different loss functions

space of factor matrices. More generally, it remains open to prove convergence guarantees for a full transformer layer in which all three matrices per head ($\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V$) and the feedforward network weights are trained jointly, with one or more of these components potentially using LoRA.

An interesting direction for future work is to extend our guarantees to more advanced sequence models such as FlashAttention (Dao et al., 2022), Performers (Choromanski et al., 2021b), Mamba (Gu & Dao, 2023). In contrast to vanilla attention, which computes softmax attention, these methods modify the computation in different ways. FlashAttention (Dao et al., 2022) computes the same softmax attention but using a memory-efficient tiled algorithm, which retains the same mathematical structure as vanilla attention and is thus likely amenable to a similar analysis. Performers (Choromanski et al., 2021a) approximates the softmax kernel $\exp(\boldsymbol{Q}\boldsymbol{K}^\top)$ using positive random feature maps $\phi(\boldsymbol{Q})$ and $\phi(\boldsymbol{K})$, resulting in an approximate attention computation of the form $\mathrm{Attn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \approx \frac{\phi(\boldsymbol{Q})(\phi(\boldsymbol{K})^\top \boldsymbol{V})}{\phi(\boldsymbol{Q})(\phi(\boldsymbol{K})^\top \mathbf{1})}$, which reduces the computational complexity. Mamba (Gu & Dao, 2023) replaces attention entirely with an input-dependent state-space model (SSM), yielding a linear-time recurrence. Extending convergence guarantees to these alternative attention mechanisms remains an open problem.

Impact Statement

This paper presents work whose goal is to advance the theoretical underpinnings of the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aghajanyan, A., Gupta, S., and Zettlemoyer, L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 7319–7328, 2021.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2016. URL https://arxiv.org/abs/1409.0473.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538, 2023.

Chen, S. and Li, Y. Provably learning a multi-head attention layer. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, pp. 1744–1754, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718174. URL https://doi.org/10.1145/3717823.3718174.

Chen, S., Sheen, H., Wang, T., and Yang, Z. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 66479–66567. Curran Associates, Inc., 2024. doi: 10.52202/079017-2127.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., and Weller, A. Rethinking attention with performers. *ICLR*, 2021a.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. Rethinking attention with performers. *International Conference on Learning Representations*, 2021b.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *ACL*, 2019.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Gao, C., Cao, Y., Li, Z., He, Y., Wang, M., Liu, H., Klusowski, J., and Fan, J. Global convergence in training large-scale transformers. *Advances in Neural Information Processing Systems*, 37:29213–29284, 2024.

Gopalani, P. and Mukherjee, A. Global convergence of sgd on two layer neural nets. *Information and Inference: A Journal of the IMA*, 14(1):iaae035, 01 2025. ISSN 2049-8772. doi: 10.1093/imaiai/iaae035. URL https://doi.org/10.1093/imaiai/iaae035.

Gopalani, P., Jha, S., and Mukherjee, A. Global convergence of SGD for logistic loss on two layer neural nets. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=9TqAUYB6tC.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Herde, M., Raonić, B., Rohner, T., Käppeli, R., Molinaro, R., de Bézenac, E., and Mishra, S. Poseidon: Efficient foundation models for pdes. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 72525–72624. Curran Associates, Inc., 2024. doi: 10.52202/079017-2311.

Hron, J., Bahri, Y., Sohl-Dickstein, J., and Novak, R. Infinite attention: NNGP and NTK for deep attention networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4376–4386. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hron20a.html.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jang, U., Lee, J. D., and Ryu, E. K. Lora training in the ntk regime has no spurious local minima. In *International Conference on Machine Learning (ICML)*, 2024.

Kim, J., Kim, J., and Ryu, E. K. Lora training provably converges to a low-rank global minimum or it fails loudly. *arXiv preprint arXiv:2502.09376*, 2025.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kumar, D., Jha, S., and Mukherjee, A. Langevin montecarlo provably learns depth two neural nets at any size and data, 2025. URL https://arxiv.org/abs/2503.10428.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *ICLR*, 2020.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

Lin, L., Bai, Y., and Mei, S. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*, 2023.

Luong, T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In Màrquez, L., Callison-Burch, C., and Su, J. (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://aclanthology.org/D15-1166/.

Molinaro, R., Lanthaler, S., Raonič, B., Rohner, T., Armegioiu, V., Simonis, S., Grund, D., Ramic, Y., Wan, Z. Y., Sha, F., Mishra, S., and Zepeda-Núñez, L. Generative ai for fast and accurate statistical computation of fluids. *arXiv preprint arXiv:2409.18359*, 2024.

Ormaniec, W., Dangel, F., and Singh, S. P. What does it mean to be a transformer? insights from a theoretical hessian analysis. *arXiv preprint arXiv:2410.10986*, 2024.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Shi, B., Su, W., and Jordan, M. I. On learning rates and schrÃ¶dinger operators. *Journal of Machine Learning Research*, 24(379):1–53, 2023. URL http://jmlr.org/papers/v24/20-364.html.

Shuttleworth, R. S., Andreas, J., Torralba, A., and Sharma, P. LoRA vs full fine-tuning: An illusion of equivalence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=xp7B8rkh7L.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Villani, C. *Hypocoercivity*, volume 202. American Mathematical Society, 2009.

Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

A. Proof of Villani Conditions for Regression on Attention

Towards stating the proofs we note the following notations,

Definition A.1 (Defining $\boldsymbol{Y}_i, \hat{\boldsymbol{Y}}_i, \boldsymbol{S}_i$ and $\boldsymbol{E}_i$). For a choice of training data as, $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathbb{R}^{t\times d} \times \mathbb{R}^{t\times d} \mid i = 1, \ldots, n\}$ and $\hat{R}_A(\boldsymbol{T}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{T})$, the loss is $\ell_i(\boldsymbol{T}) = \frac{1}{2}\|\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i(\boldsymbol{T})\|_F^2$. The output is defined as $\hat{\boldsymbol{Y}}_i(\boldsymbol{T}) \coloneqq \boldsymbol{S}_i(\boldsymbol{T})\boldsymbol{X}_i\boldsymbol{W}_v$, where $\boldsymbol{S}_i(\boldsymbol{T}) \coloneqq \mathrm{RowSoftMax}_\beta(\boldsymbol{M}_i) \in \mathbb{R}^{t\times t}$ and $\boldsymbol{M}_i = \frac{1}{\sqrt{d}}\boldsymbol{X}_i\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}_i^\top \in \mathbb{R}^{t\times t}$. We also define the error as, $\boldsymbol{E}_i \coloneqq \hat{\boldsymbol{Y}}_i(\boldsymbol{T}) - \boldsymbol{Y}_i \in \mathbb{R}^{t\times d}$

Lemma A.2. *The norm of RowSoftMax$_\beta$ is bounded by $\|\boldsymbol{S}_i\|_F \le \sqrt{t}$. The norm of the Jacobian and Hessian of RowSoftMax$_\beta$ from Definition 2.3 are bounded by $\beta B_{s'}$ and $\beta^2 B_{s''}$, respectively. That is*

$$\|\mathrm{d}\boldsymbol{S}\|_F \le B_{s'}\beta\|\mathrm{d}\boldsymbol{M}\|_F, \tag{11}$$

and

$$\|\mathrm{d}^2\boldsymbol{S}_i\|_F \le \beta^2 B_{s''}\|\mathrm{d}\boldsymbol{M}_i\|_F^2 \tag{12}$$

where $B_{s'} = 2$ and $B_{s''} = 6t^2$ are finite constants, and β is the constant temperature parameter introduced from Definition 2.3. [2]

Lemma A.3. *The bound of gradient and laplacian of $\hat{R}_A(\boldsymbol{T})$ are given by*

$$\|\nabla_{\boldsymbol{T}}\hat{R}_A(\boldsymbol{T})\| = \sqrt{\|\nabla_{\boldsymbol{W}_Q}\hat{R}_A(\boldsymbol{T})\|_F^2 + \|\nabla_{\boldsymbol{W}_K}\hat{R}_A(\boldsymbol{T})\|_F^2} \le \left[(\sqrt{t}B_xB_w + B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\right]\|\boldsymbol{T}\| \tag{13}$$

and

$$|\Delta_{\boldsymbol{T}}\hat{R}_A(\boldsymbol{T})| = \Delta_{\boldsymbol{W}_Q}\hat{R}_A(\boldsymbol{T}) + \Delta_{\boldsymbol{W}_K}\hat{R}_A(\boldsymbol{T}) \le \frac{B_x^4}{d}\left(((\beta B_{s'})B_xB_w)^2 + (\sqrt{t}B_xB_w + B_y)B_xB_w(\beta^2 B_{s''})\right)\|\boldsymbol{T}\|^2. \tag{14}$$

The above lemmas are proved in Section B.

A.1. Proof of Theorem 3.1 for Loss in Definition 2.4

Proof. We note that,

$$V_{ATT}(\boldsymbol{T}) = \hat{R}_A(\boldsymbol{T}) + R(\boldsymbol{T}) = \hat{R}_A(\boldsymbol{T}) + \frac{\lambda}{2}\|\boldsymbol{T}\|^2\log\left(1 + \|\boldsymbol{T}\|^2\right), \tag{15}$$

where the regularization is defined as $R(\boldsymbol{T}) = \frac{\lambda}{2}\|\boldsymbol{T}\|^2\log\left(1 + \|\boldsymbol{T}\|^2\right)$.

We start with the gradient of the regularization term,

$$\nabla R(\boldsymbol{T}) = \lambda\boldsymbol{T}\log\left(1 + \|\boldsymbol{T}\|^2\right) + \lambda\boldsymbol{T}\frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2}, \tag{16}$$

so

$$\|\nabla R(\boldsymbol{T})\| = \lambda\|\boldsymbol{T}\|\left(\log\left(1 + \|\boldsymbol{T}\|^2\right) + \frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2}\right) \tag{17}$$

Using the expression of $\|\nabla R(\boldsymbol{T})\|$ from above and the upperbound on $\|\nabla_{\boldsymbol{T}}\hat{R}_A(\boldsymbol{T})\|$ from Lemma A.3, we have,

$$\begin{aligned}
\|\nabla V_{ATT}(\boldsymbol{T})\|^2 &= \|\nabla\hat{R}_A(\boldsymbol{T}) + \nabla R(\boldsymbol{T})\|^2 \ge \|\nabla R(\boldsymbol{T})\|^2 - 2\|\nabla R(\boldsymbol{T})\|\sup\|\nabla_{\boldsymbol{T}}\hat{R}_A(\boldsymbol{T})\| \\
&\ge \lambda^2\|\boldsymbol{T}\|^2\left(\log\left(1 + \|\boldsymbol{T}\|^2\right) + \frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2}\right)^2 \\
&\quad - 2\lambda\beta\|\boldsymbol{T}\|\left(\log\left(1 + \|\boldsymbol{T}\|^2\right) + \frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2}\right)\cdot\left[(\sqrt{t}B_xB_w + B_y)B_wB_xB_{s'}\frac{B_x^2}{\sqrt{d}}\right]\|\boldsymbol{T}\|
\end{aligned} \tag{18}$$

[2] For a matrix-valued function $\boldsymbol{S}(\boldsymbol{M})$, the first-order differential $\mathrm{d}\boldsymbol{S}$ represents the linear principal part of the change in $\boldsymbol{S}$ given an infinitesimal perturbation $\mathrm{d}\boldsymbol{M}$, defined via the Taylor expansion: $\boldsymbol{S}(\boldsymbol{M}+\mathrm{d}\boldsymbol{M}) = \boldsymbol{S}(\boldsymbol{M}) + \mathrm{d}\boldsymbol{S} + \mathcal{O}(\|\mathrm{d}\boldsymbol{M}\|^2)$. Similarly, the second-order differential $\mathrm{d}^2\boldsymbol{S}$ represents the quadratic variation, such that $\boldsymbol{S}(\boldsymbol{M} + \mathrm{d}\boldsymbol{M}) = \boldsymbol{S}(\boldsymbol{M}) + \mathrm{d}\boldsymbol{S} + \frac{1}{2}\mathrm{d}^2\boldsymbol{S} + \mathcal{O}(\|\mathrm{d}\boldsymbol{M}\|^3)$, encapsulating the action of the Hessian tensor.

Next we note that,

$$\begin{aligned}\Delta R(\boldsymbol{T}) &= \sum_{k=1}^{D} \frac{\partial}{\partial T_k}\left(\lambda T_k\left[\log\left(1+\|\boldsymbol{T}\|^2\right)+\frac{\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}\right]\right)\\ &= D\lambda\left[\log\left(1+\|\boldsymbol{T}\|^2\right)+\frac{\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}\right]+\sum_{k=1}^{D}\lambda T_k\left[\frac{2T_k}{1+\|\boldsymbol{T}\|^2}+\frac{2T_k}{(1+\|\boldsymbol{T}\|^2)^2}\right]\\ &= D\lambda\left[\log\left(1+\|\boldsymbol{T}\|^2\right)+\frac{\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}\right]+\frac{2\lambda\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}+\frac{2\lambda\|\boldsymbol{T}\|^2}{(1+\|\boldsymbol{T}\|^2)^2}. \end{aligned} \tag{19}$$

Also recall the upper bound of$|\Delta_T \hat{R}_A(\boldsymbol{T})|$

$$\frac{1}{s}\|\nabla V_{ATT}\|^2-\Delta V_{ATT}\geq\frac{1}{s}\|\nabla V_{ATT}\|^2-|\Delta_T\hat{R}_A|-\Delta R(T)$$

given Lemma A.3, substitute equation 18 and 19 into Villani condition, we have,

$$\begin{aligned}&\lim_{\|\boldsymbol{T}\|\to\infty}\left(\frac{1}{s}\|\nabla V_{ATT}\|^2-\Delta V_{ATT}\right)\geq\lim_{\|\boldsymbol{T}\|\to\infty}\frac{1}{s}\|\nabla V_{ATT}\|^2-|\Delta_T\hat{R}_A|-\Delta R(T)\\ &=\lim_{\|\boldsymbol{T}\|\to\infty}\|\boldsymbol{T}\|^2\Bigg[\underbrace{\frac{\lambda^2}{s}\left(\log\left(1+\|\boldsymbol{T}\|^2\right)+\frac{\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}\right)^2}_{\text{from equation 18}\quad\to+\infty\quad\text{at}\quad O(\log\|\boldsymbol{T}\|)^2}\\ &\underbrace{-\frac{2\lambda\beta}{s}\left[(\sqrt{t}B_xB_w+B_y)B_wB_xB_{s'}\frac{B_x^2}{\sqrt{d}}\right]\left(\log\left(1+\|\boldsymbol{T}\|^2\right)+\frac{\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}\right)}_{\text{from equation 18}\quad\to+\infty\quad\text{at}\quad O(\log\|\boldsymbol{T}\|)}\\ &\underbrace{-\beta^2\frac{B_x^4}{d}\left((B_{s'}B_xB_w)^2+(\sqrt{t}B_xB_w+B_y)B_xB_wB_{s''}\right)}_{\text{from equation 14}\quad\text{constant}}\underbrace{-\frac{D\lambda\left[\log\left(1+\|\boldsymbol{T}\|^2\right)+\frac{\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}\right]+\frac{2\lambda\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}+\frac{2\lambda\|\boldsymbol{T}\|^2}{(1+\|\boldsymbol{T}\|^2)^2}}{\|\boldsymbol{T}\|^2}}_{\text{from equation 19}\quad\to0}\Bigg].\end{aligned} \tag{20}$$

So Villani condition is satisfied for all β and r(which is contained in D), this is because the leading term from the gradient of regularization dominates, which is independent on β and r. □

A.2. Proof of Theorem 3.1 for Loss in Definition 2.5

Proof. We note that,

$$V_{ATT,\varepsilon}(\boldsymbol{T})=\hat{R}_{A,\varepsilon}(\boldsymbol{T})+R_\varepsilon(\boldsymbol{T})=\hat{R}_A(\boldsymbol{T})+\left[\frac{\lambda}{2}(\|\boldsymbol{W}_Q\|_F^{2+\epsilon}+\|\boldsymbol{W}_K\|_F^{2+\epsilon})\right], \tag{21}$$

where the regularization term is defined as $R_\varepsilon(\boldsymbol{T})=\left[\frac{\lambda}{2}(\|\boldsymbol{W}_Q\|_F^{2+\epsilon}+\|\boldsymbol{W}_K\|_F^{2+\epsilon})\right]$. Since we have,

$$\begin{aligned}\nabla R_\varepsilon(\boldsymbol{T})=\nabla\left[\frac{\lambda}{2}(\|\boldsymbol{W}_Q\|_F^{2+\epsilon}+\|\boldsymbol{W}_K\|_F^{2+\epsilon})\right]&=\left(\nabla_{\boldsymbol{W}_Q}\left[\frac{\lambda}{2}(\|\boldsymbol{W}_Q\|_F^2)^{\frac{2+\epsilon}{2}}\right],\nabla_{\boldsymbol{W}_K}\left[\frac{\lambda}{2}(\|\boldsymbol{W}_K\|_F^2)^{\frac{2+\epsilon}{2}}\right]\right)\\ &=\frac{\lambda}{2}\left(\frac{2+\epsilon}{2}(\|\boldsymbol{W}_Q\|_F^2)^{\frac{\epsilon}{2}}\cdot\nabla_{\boldsymbol{W}_Q}(\|\boldsymbol{W}_Q\|_F^2),\frac{2+\epsilon}{2}(\|\boldsymbol{W}_K\|_F^2)^{\frac{\epsilon}{2}}\cdot\nabla_{\boldsymbol{W}_K}(\|\boldsymbol{W}_K\|_F^2)\right)\\ &=\frac{\lambda}{2}\left(\frac{2+\epsilon}{2}\|\boldsymbol{W}_Q\|_F^{\epsilon}\cdot(2\boldsymbol{W}_Q),\frac{2+\epsilon}{2}\|\boldsymbol{W}_K\|_F^{\epsilon}\cdot(2\boldsymbol{W}_K)\right)\\ &=\frac{\lambda}{2}\left((2+\epsilon)\|\boldsymbol{W}_Q\|_F^{\epsilon}\boldsymbol{W}_Q,(2+\epsilon)\|\boldsymbol{W}_K\|_F^{\epsilon}\boldsymbol{W}_K\right)\end{aligned} \tag{22}$$

so

$$\|\nabla R_\varepsilon(\boldsymbol{T})\|^2=\left(\frac{\lambda}{2}\right)^2(2+\varepsilon)^2\left(\|\boldsymbol{W}_Q\|_F^{2+2\varepsilon}+\|\boldsymbol{W}_K\|_F^{2+2\varepsilon}\right)\geq\left(\frac{\lambda}{2}\right)^2(2+\varepsilon)^2 2^{-\varepsilon}\|\boldsymbol{T}\|^{2+2\varepsilon}, \tag{23}$$

where we have used, $\frac{(\|\boldsymbol{W}_Q\|^2)^{1+\varepsilon}+(\|\boldsymbol{W}_K\|^2)^{1+\varepsilon}}{2} \geq \left(\frac{\|\boldsymbol{W}_Q\|^2+\|\boldsymbol{W}_K\|^2}{2}\right)^{1+\varepsilon} = 2^{-1-\varepsilon}\|\boldsymbol{T}\|^{2+2\varepsilon}$ by Jensen Inequality.

Since $\frac{\partial}{\partial \boldsymbol{W}_{Q,ij}}\|\boldsymbol{W}_Q\|_F^{2+\varepsilon} = (2+\varepsilon)\|\boldsymbol{W}_Q\|_F^{\varepsilon}\boldsymbol{W}_{Q,ij}$ from above, we have

$$\frac{\partial^2}{\partial \boldsymbol{W}_{Q,ij}^2}\|\boldsymbol{W}_Q\|_F^{2+\varepsilon} = (2+\varepsilon)\varepsilon\|\boldsymbol{W}_Q\|_F^{\varepsilon-2}W_{Q,ij}^2 + (2+\varepsilon)\|\boldsymbol{W}_Q\|_F^{\varepsilon}. \tag{24}$$

By summing over all $d \times r$ elements of $\boldsymbol{W}_Q$:

$$\begin{aligned}\Delta_{\boldsymbol{W}_Q}\|\boldsymbol{W}_Q\|_F^{2+\varepsilon} &= (2+\varepsilon)\varepsilon\|\boldsymbol{W}_Q\|_F^{\varepsilon-2}\cdot\left(\sum_{i=1}^{d}\sum_{j=1}^{r}W_{Q,ij}^2\right) + \sum_{i=1}^{d}\sum_{j=1}^{r}\left[(2+\varepsilon)\|\boldsymbol{W}_Q\|_F^{\varepsilon}\right] \\ &= (2+\varepsilon)\|\boldsymbol{W}_Q\|_F^{\varepsilon}\cdot\varepsilon + (2+\varepsilon)\|\boldsymbol{W}_Q\|_F^{\varepsilon}\cdot dr = (2+\varepsilon)(\varepsilon+dr)\|\boldsymbol{W}_Q\|_F^{\varepsilon}.\end{aligned} \tag{25}$$

By symmetry, the Laplacian for $\boldsymbol{W}_K \in \mathbb{R}^{d\times r}$ follows the exact same form: $\Delta_{\boldsymbol{W}_K}\|\boldsymbol{W}_K\|_F^{2+\varepsilon} = (2+\varepsilon)(\varepsilon+dr)\|\boldsymbol{W}_K\|_F^{\varepsilon}$ We have

$$\Delta R_\varepsilon(\boldsymbol{T}) = \frac{\lambda}{2}(2+\varepsilon)\left[(\varepsilon+dr)\|\boldsymbol{W}_Q\|_F^{\varepsilon} + (\varepsilon+dr)\|\boldsymbol{W}_K\|_F^{\varepsilon}\right] \leq \frac{\lambda}{2}(2+\varepsilon)(2\varepsilon+D)\|\boldsymbol{T}\|^{\varepsilon}, \tag{26}$$

where $D = 2dr$.

For the potential, together with the upper bound of $\|\nabla\hat{R}_A\|$ from Lemma A.3 we have,

$$\begin{aligned}\|\nabla V_{ATT,\epsilon}\|^2 &= \|\nabla\hat{R}_A + \nabla R_\epsilon\|^2 \geq \|\nabla R_\epsilon\|^2 - 2\|\nabla R_\epsilon\|\|\nabla\hat{R}_A\| \\ &\geq \left(\frac{\lambda}{2}\right)^2(2+\epsilon)^2 2^{-\epsilon}\|\boldsymbol{T}\|^{2+2\epsilon} - 2\left(\frac{\lambda}{2}(2+\epsilon)\|\boldsymbol{T}\|^{1+\epsilon}\right)\left(\left[(\sqrt{t}B_xB_w+B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\right]\|\boldsymbol{T}\|\right) \\ &= \left(\frac{\lambda}{2}\right)^2(2+\epsilon)^2 2^{-\epsilon}\|\boldsymbol{T}\|^{2+2\epsilon} - \lambda(2+\epsilon)\left[(\sqrt{t}B_xB_w+B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\right]\|\boldsymbol{T}\|^{2+\epsilon}.\end{aligned} \tag{27}$$

The last inequality we have used $\|\boldsymbol{W}_Q\|_F^{2+2\varepsilon} + \|\boldsymbol{W}_K\|_F^{2+2\varepsilon} \leq (\|\boldsymbol{W}_Q\|_F^2 + \|\boldsymbol{W}_K\|_F^2)^{1+\varepsilon} = \|\boldsymbol{T}\|^{2+2\varepsilon}$.

Substituting this, along with the upper bound of $|\Delta_T\hat{R}_A|$ by Lemma A.3 and upper bound of $\Delta R_\epsilon(\boldsymbol{T})$ from equation 26, into the Villani limit expression for any given $s > 0$,

$$\begin{aligned}&\lim_{\|\boldsymbol{T}\|\to\infty}\left(\frac{1}{s}\|\nabla V_{ATT,\varepsilon}\|^2 - \Delta V_{ATT,\varepsilon}\right) \geq \lim_{\|\boldsymbol{T}\|\to\infty}\frac{1}{s}\|\nabla V_{ATT,\varepsilon}\|^2 - |\Delta_{\boldsymbol{T}}\hat{R}_A| - \Delta R_\varepsilon(\boldsymbol{T}) \\ &\geq \lim_{\|\boldsymbol{T}\|\to\infty}\left[\frac{1}{s}\left(\left(\frac{\lambda}{2}\right)^2(2+\epsilon)^2 2^{-\epsilon}\|\boldsymbol{T}\|^{2+2\epsilon} - \lambda(2+\epsilon)\left[(\sqrt{t}B_xB_w+B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\right]\|\boldsymbol{T}\|^{2+\epsilon}\right)\right. \\ &\quad\left. - \left((\beta B_{s'}B_xB_w)^2 + (\sqrt{t}B_xB_w+B_y)B_xB_w\beta^2B_{s''}\right)\frac{B_x^4}{d}\|\boldsymbol{T}\|^2 - \frac{\lambda}{2}(2+\epsilon)(2\epsilon+D)\|\boldsymbol{T}\|^{\epsilon}\right] \\ &= \lim_{\|\boldsymbol{T}\|\to\infty}\|\boldsymbol{T}\|^2\left[\underbrace{\frac{1}{s}\left(\frac{\lambda}{2}\right)^2(2+\epsilon)^2 2^{-\epsilon}\|\boldsymbol{T}\|^{2\epsilon}}_{\text{from equation 27 } \to+\infty \text{ at } \|\boldsymbol{T}\|^{2\epsilon}} - \underbrace{\frac{\lambda(2+\epsilon)\left[(\sqrt{t}B_xB_w+B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\right]}{s}\|\boldsymbol{T}\|^{\epsilon}}_{\text{from equation 27 } \to+\infty \text{ at } \|\boldsymbol{T}\|^{\epsilon}}\right. \\ &\quad\left. - \underbrace{\left((\beta B_{s'}B_xB_w)^2 + (\sqrt{t}B_xB_w+B_y)B_xB_w\beta^2B_{s''}\right)\frac{B_x^4}{d}}_{\text{from equation 14 Constant}} - \underbrace{\frac{\lambda}{2}(2+\epsilon)(2\epsilon+D)\|\boldsymbol{T}\|^{\epsilon-2}}_{\text{from equation 26 } \to 0}\right].\end{aligned} \tag{28}$$

The leading order term $O(\|\boldsymbol{T}\|^{2\epsilon})$ dominates. Since $\epsilon > 0$, regardless of the choices of dimensions d, t, key-query inner dimension r, and Softmax temperature β. Consequently, the limit diverges to $+\infty$, proving that the ϵ-regularized attention loss unconditionally satisfies the Villani condition. □

B. Proofs of Intermediate Lemmas for Theorem 3.1

Proof of Lemma A.2. By the definition of the row-wise softmax function, each element $S_{i,j,k}$ (the element in the j-th row and k-th column of S_i) represents a valid probability. Therefore, all elements are positive, $S_{i,j,k} \in (0,1)$, and the sum of the elements across any given row j is exactly 1:$\sum_{k=1}^{t} S_{i,j,k} = 1$. The squared Frobenius norm of the matrix S_i is defined as the sum of its squared entries:

$$\|S_i\|_F^2 = \sum_{j=1}^{t}\sum_{k=1}^{t} S_{i,j,k}^2 \le \left(\sum_{k=1}^{t} S_{i,j,k}\right)^2 = 1^2 = 1,$$

so$\|S_i\|_F^2 \le \sum_{j=1}^{t} 1 = t$, which means the bound of $\boldsymbol{S} = \mathrm{RowSoftMax}_\beta(\boldsymbol{M})$ is given by $\|S_i\|_F \le \sqrt{t}$.

To determine the exact bounds for the Jacobian and Hessian of $\boldsymbol{S} = \mathrm{RowSoftMax}_\beta(\boldsymbol{M})$, we adapt the structural findings from (Ormaniec et al., 2024).

As established in Appendix C.1, Lemma C.1 of (Ormaniec et al., 2024), since the row-wise softmax applies independently to each row, the cross-row derivatives are strictly zero. This decoupling endows the full Jacobian matrix and the Hessian tensor with a block-diagonal structure, allowing us to analyze them row by row.

Jacobian Bound For any single row i, the output probability is defined element-wise as $\boldsymbol{S}_{i,j} = \frac{\exp(\beta \boldsymbol{M}_{i,j})}{\sum_l \exp(\beta \boldsymbol{M}_{i,l})}$. To compute the derivative with respect to the input $\boldsymbol{M}_{i,k}$ we invoke Lemma B.1, Equation 19 of (Ormaniec et al., 2024) to obtain the exact local Jacobian matrix,

$$\frac{\partial \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,k}} = \frac{\partial}{\partial(\beta \boldsymbol{M}_{i,k})}\left(\frac{\exp(\beta \boldsymbol{M}_{i,j})}{\sum_l \exp(\beta \boldsymbol{M}_{i,l})}\right)\cdot \frac{\partial(\beta \boldsymbol{M}_{i,k})}{\partial \boldsymbol{M}_{i,k}} = \boldsymbol{S}_{i,j}(\delta_{j,k} - \boldsymbol{S}_{i,k})\cdot\beta \tag{29}$$

where $\delta_{j,k}$ is the Kronecker delta. Expressing this element-wise relationship in matrix form for the entire i-th row $\boldsymbol{S}_i \in \mathbb{R}^t$,:

$$\mathcal{J}_i = \frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{M}_i} = \beta\left(\mathrm{diag}(\boldsymbol{S}_i) - \boldsymbol{S}_i^\top \boldsymbol{S}_i\right) \tag{30}$$

where $\mathcal{J}_i \in \mathbb{R}^{t\times t}$ is used to denote the Jacobian and $\boldsymbol{S}_i$ is the i-th row of the attention probability matrix. Since $\boldsymbol{S}_{i,k} \in (0,1)$ and $\sum_k \boldsymbol{S}_{i,k} = 1$, $\mathrm{diag}(\boldsymbol{S}_i) - \boldsymbol{S}_i^\top \boldsymbol{S}_i$ represents the exact covariance matrix of a categorical distribution. We have the bound

$$\|\mathrm{diag}(\boldsymbol{S}_i) - \boldsymbol{S}_i^\top \boldsymbol{S}_i\|_2 \le \|\mathrm{diag}(\boldsymbol{S}_i)\|_2 + \|\boldsymbol{S}_i^\top \boldsymbol{S}_i\|_2 \le \max_k \boldsymbol{S}_{i,k} + \sum_{k=1}^{t} \boldsymbol{S}_{i,k}^2 \le 2. \tag{31}$$

In the above step, the first inequality applies the triangle inequality. The subsequent equality is exact for the spectral norm ($\|\cdot\|_2$): the norm of the diagonal matrix equals its maximum entry, and the norm of the rank-1 positive semi-definite matrix $\boldsymbol{S}_i^\top \boldsymbol{S}_i$ equals its trace. The final strict inequality holds because $\boldsymbol{S}_i$ is a probability vector which dictates that $\max_k \boldsymbol{S}_{i,k} < 1$ and $\sum_{k=1}^{t} \boldsymbol{S}_{i,k}^2 < 1$.

We denote $\mathrm{d}\boldsymbol{M}$ and $\mathrm{d}\boldsymbol{S}$ as the first-order matrix differentials, representing an arbitrary infinitesimal perturbation in the input pre-activation matrix and the corresponding induced perturbation in the output probability matrix, respectively. For a single row i, $\mathrm{d}\boldsymbol{M}_i$ and $\mathrm{d}\boldsymbol{S}_i$ represent their respective row vector differentials. Since the full Jacobian is block-diagonal, the differential mapping from $\mathrm{d}\boldsymbol{M}$ to $\mathrm{d}\boldsymbol{S}$ operates independently on each row. For any individual row i, the Euclidean norm (2-norm) of the differential vector satisfies $\|\mathrm{d}\boldsymbol{S}_i\|_2 \le 2\beta\|\mathrm{d}\boldsymbol{M}_i\|_2$ based on the spectral norm bound derived above. By definition, the squared Frobenius norm of a matrix is the sum of the squared 2-norms of its row vectors. Summing over all t rows, we obtain:

$$\|\mathrm{d}\boldsymbol{S}\|_F^2 = \sum_{i=1}^{t}\|\mathrm{d}\boldsymbol{S}_i\|_2^2 \le \sum_{i=1}^{t}(2\beta)^2\|\mathrm{d}\boldsymbol{M}_i\|_2^2 = 4\beta^2\sum_{i=1}^{t}\|\mathrm{d}\boldsymbol{M}_i\|_2^2 = 4\beta^2\|\mathrm{d}\boldsymbol{M}\|_F^2 \tag{32}$$

Taking the square root of both sides directly yields the global bound for the differential:

$$\|\mathrm{d}\boldsymbol{S}\|_F \le B_{s'}\beta\|\mathrm{d}\boldsymbol{M}\|_F \tag{33}$$

Therefore, the Jacobian norm is bounded by $\beta B_{s'}$, where the constant is explicitly evaluated as $B_{s'} = 2$.

Hessian Bound Differentiating the local Jacobian element $\frac{\partial \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,k}} = \beta(\boldsymbol{S}_{i,j}\delta_{j,k} - \boldsymbol{S}_{i,j}\boldsymbol{S}_{i,k})$ with respect to another input $\boldsymbol{M}_{i,l}$ yields an additional factor of β via the chain rule.

$$\begin{aligned}\frac{\partial^2 \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,k}\partial \boldsymbol{M}_{i,l}} &= \beta\left(\frac{\partial \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,l}}\delta_{j,k} - \frac{\partial \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,l}}\boldsymbol{S}_{i,k} - \boldsymbol{S}_{i,j}\frac{\partial \boldsymbol{S}_{i,k}}{\partial \boldsymbol{M}_{i,l}}\right)\\ &= \beta^2\left((\boldsymbol{S}_{i,j}\delta_{j,l} - \boldsymbol{S}_{i,j}\boldsymbol{S}_{i,l})\delta_{j,k} - (\boldsymbol{S}_{i,j}\delta_{j,l} - \boldsymbol{S}_{i,j}\boldsymbol{S}_{i,l})\boldsymbol{S}_{i,k} - \boldsymbol{S}_{i,j}(\boldsymbol{S}_{i,k}\delta_{k,l} - \boldsymbol{S}_{i,k}\boldsymbol{S}_{i,l})\right)\\ &= \beta^2\boldsymbol{S}_{i,j}\left(\delta_{j,k}\delta_{j,l} - \boldsymbol{S}_{i,l}\delta_{j,k} - \delta_{j,l}\boldsymbol{S}_{i,k} + \boldsymbol{S}_{i,l}\boldsymbol{S}_{i,k} - \boldsymbol{S}_{i,k}\delta_{k,l} + \boldsymbol{S}_{i,k}\boldsymbol{S}_{i,l}\right)\\ &= \beta^2\boldsymbol{S}_{i,j}\left(2\boldsymbol{S}_{i,k}\boldsymbol{S}_{i,l} + \delta_{j,k}\delta_{j,l} - \delta_{k,l}\boldsymbol{S}_{i,k} - \delta_{j,k}\boldsymbol{S}_{i,l} - \delta_{j,l}\boldsymbol{S}_{i,k}\right)\end{aligned} \tag{34}$$

This scalar element-wise formulation is the direct expansion of the matrix-level second derivative derived in Lemma C.1 of (Ormaniec et al., 2024). This expression exclusively comprises attention probability values $\boldsymbol{S}_{i,\cdot} \in (0,1)$ and Kronecker deltas $\delta \in \{0,1\}$. By the triangle inequality, the absolute value of each individual entry in this local 3D Hessian tensor $\mathcal{H}_i \in \mathbb{R}^{t\times t\times t}$, defined as the local 3D Hessian tensor $\mathcal{H}_i \in \mathbb{R}^{t\times t\times t}$ as the collection of all second-order partial derivatives of the i-th row of the output $\boldsymbol{S}$ with respect to the i-th row of the input $\boldsymbol{M}$. For $i \in \{1,2,\ldots,t\}$, its element at index (j,k,l) is defined as:

$$(\mathcal{H}_i)_{j,k,l} := \frac{\partial^2 \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,k}\partial \boldsymbol{M}_{i,l}} \quad \text{for } j,k,l \in \{1,\ldots,t\},$$

which is strictly bounded as,

$$\left|\frac{\partial^2 \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,k}\partial \boldsymbol{M}_{i,l}}\right| < \beta^2\cdot 1\cdot(2\cdot 1\cdot 1 + 1 + 1 + 1 + 1) = 6\beta^2 \tag{35}$$

For a single row i, the squared Frobenius norm of its local Hessian tensor $\mathcal{H}_i$ is the sum of its t^3 squared entries. We bound this local tensor norm as,

$$\|\mathcal{H}_i\|_F^2 = \sum_{j,k,l=1}^{t}\left(\frac{\partial^2 \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,k}\partial \boldsymbol{M}_{i,l}}\right)^2 \le \sum_{j,k,l=1}^{t}(6\beta^2)^2 = t^3(36\beta^4) \implies \|\mathcal{H}_i\|_F \le 6\beta^2 t^{1.5} \tag{36}$$

For mapping $\boldsymbol{S} = \mathrm{RowSoftMax}_\beta(\boldsymbol{M})$, the complete global 6th-order Hessian tensor $H_{full} \in \mathbb{R}^{t\times t\times t\times t\times t\times t}$ is defined as the derivative of any output entry matrix $\boldsymbol{S}_{i,j}$ with respect to any two input entries $\boldsymbol{M}_{p,k}$ and $\boldsymbol{M}_{q,l}$:

$$(H_{full})_{i,j,p,k,q,l} := \frac{\partial^2 \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{p,k}\partial \boldsymbol{M}_{q,l}} \quad \text{for } i,j,p,k,q,l \in \{1,\ldots,t\}. \tag{37}$$

As discussed in Appendix C.1, Lemma C.1 of (Ormaniec et al., 2024), the full Hessian tensor for the entire matrix-to-matrix mapping is block-diagonal, all cross-row second derivatives evaluate to zero. We have

$$\frac{\partial^2 \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{p,k}\partial \boldsymbol{M}_{q,l}} = \begin{cases}(\mathcal{H}_i)_{j,k,l} & \text{if } i = p = q,\\ 0 & \text{otherwise.}\end{cases} \tag{38}$$

This simplifies its total squared Frobenius norm being simply the sum of the squared norms of the t independent row blocks. Thus, the global Hessian norm $\|H_{full}\|_F$,satisfies,

$$\begin{aligned}\|H_{full}\|_F^2 &= \sum_{i=1}^{t}\sum_{j=1}^{t}\sum_{p=1}^{t}\sum_{k=1}^{t}\sum_{q=1}^{t}\sum_{l=1}^{t}\left(\frac{\partial^2 \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{p,k}\partial \boldsymbol{M}_{q,l}}\right)^2\\ &= \sum_{i=1}^{t}\sum_{j,k,l=1}^{t}\left(\frac{\partial^2 \boldsymbol{S}_{i,j}}{\partial \boldsymbol{M}_{i,k}\partial \boldsymbol{M}_{i,l}}\right)^2 = \sum_{i=1}^{t}\|\mathcal{H}_i\|_F^2 \le \sum_{i=1}^{t}(36\beta^4 t^3) = 36\beta^4 t^4.\end{aligned} \tag{39}$$

Taking the square root of both sides, the global Hessian norm is bounded by $6\beta^2 t^2$. Therefore, the bound can be denoted as $\beta^2 B_{s''}$, that is

$$\|\mathrm{d}^2\boldsymbol{S}_i\|_F \le \beta^2 B_{s''}\|\mathrm{d}\boldsymbol{M}_i\|_F^2, \tag{40}$$

where the constant is explicitly evaluated as $B_{s''} = 6t^2$. □

Proof of Lemma A.3. We begin by bounding the norm of the gradient $\nabla_{\boldsymbol{T}}\hat{R}_A$, where $\boldsymbol{T} = (\boldsymbol{W}_Q, \boldsymbol{W}_K)$. Recall from Definitions 2.4 and 2.5 that$\hat{R}_A \coloneqq \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{T})$. Taking gradients and applying the triangle inequality yields $\|\nabla_{\boldsymbol{T}}\hat{R}_A\| \le \frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{T}}\ell_i(\boldsymbol{T})\|$. Thus, it suffices to bound $\|\nabla_{\boldsymbol{T}}\ell_i(\boldsymbol{T})\|$, where $\|\nabla_{\boldsymbol{T}}\ell_i(\boldsymbol{T})\|^2 = \|\nabla_{\boldsymbol{W}_Q}\ell_i\|_F^2 + \|\nabla_{\boldsymbol{W}_K}\ell_i\|_F^2$.

From Definition A.1, recall that $\boldsymbol{M}_i = \frac{1}{\sqrt{d}}\boldsymbol{X}_i\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}_i^\top$. In equation (54), we will show that the parameter gradients $\nabla_{\boldsymbol{W}_Q}\ell_i$ and $\nabla_{\boldsymbol{W}_K}\ell_i$ can be expressed in terms of the intermediate gradient $\nabla_{\boldsymbol{M}_i}\ell_i$ as

$$\nabla_{\boldsymbol{W}_Q}\ell_i = \frac{1}{\sqrt{d}}\boldsymbol{X}_i^\top(\nabla_{\boldsymbol{M}_i}\ell_i)\,\boldsymbol{X}_i\boldsymbol{W}_K,$$
$$\nabla_{\boldsymbol{W}_K}\ell_i = \frac{1}{\sqrt{d}}\boldsymbol{X}_i^\top(\nabla_{\boldsymbol{M}_i}\ell_i)\,\boldsymbol{X}_i\boldsymbol{W}_Q.$$

Therefore, it suffices to bound $\|\nabla_{\boldsymbol{M}_i}\ell_i\|_F$. Next, we will express the intermediate gradient in terms of the upstream gradient $(\nabla_{\boldsymbol{S}_i}\ell_i)$ in equation (48), $\nabla_{\boldsymbol{M}_i}\ell_i = \left(\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{M}_i}\right)^\top \nabla_{\boldsymbol{S}_i}\ell_i$, where $\boldsymbol{S}_i$ is defined in Definition A.1. In equation (44) we will show that the upstream gradient admits the form $\nabla_{\boldsymbol{S}_i}\ell_i = \boldsymbol{E}_i\boldsymbol{W}_v^\top\boldsymbol{X}_i^\top$, where $\boldsymbol{E}_i = \hat{\boldsymbol{Y}}_i - \boldsymbol{Y}_i$ is the error matrix.

Thus, we proceed in four steps, (Step 1) **Bounding the Error Matrix** (i.e. bounding $\|\boldsymbol{E}_i\|_F$), which in turn allows us to (Step 2) **Bounding the Upstream Gradient** (i.e. control $\|\nabla_{\boldsymbol{S}_i}\ell_i\|_F$), then (Step 3) **Bounding the Intermediate Gradient** (i.e. $\|\nabla_{\boldsymbol{M}_i}\ell_i\|_F$) and finally (Step 4) **Bounding the Parameter Gradients**.

• **Bounding the Error Matrix:** First using the triangle inequality and the Softmax output bound $\|\boldsymbol{S}_i\|_F \le \sqrt{t}$ to bound $\boldsymbol{E}_i = \hat{\boldsymbol{Y}}_i - \boldsymbol{Y}_i$ to obtain $\|\boldsymbol{E}_i\|_F \le \sqrt{t}B_xB_w + B_y$ (see Equation 41).

Recalling $\boldsymbol{E}_i = \hat{\boldsymbol{Y}}_i(T) - \boldsymbol{Y}_i$ we have,

$$\|\boldsymbol{E}_i\|_F \le \|\hat{\boldsymbol{Y}}_i\|_F + \|\boldsymbol{Y}_i\|_F \le \|\boldsymbol{S}_i\|_F\|\boldsymbol{X}_i\|_F\|\boldsymbol{W}_v\|_F + B_y \le \sqrt{t}B_xB_w + B_y, \tag{41}$$

where we have used the bound of $\|\boldsymbol{S}_i\|_F$ proved in Lemma A.2. Recall the other constants used in the RHS come from definitions 2.3, 2.8 and 2.9.

• **Bounding the Upstream Gradient:** Based on the explicit expression for the loss gradient with respect to the output probability matrix, $\nabla_{\boldsymbol{S}_i}\ell_i = \boldsymbol{E}_i\boldsymbol{W}_v^\top\boldsymbol{X}_i^\top$, we apply sub-multiplicativity to bound the upstream gradient: $\|\nabla_{\boldsymbol{S}_i}\ell_i\|_F \le (\sqrt{t}B_xB_w + B_y)B_wB_x$ (see Equation 45).

Recalling $\hat{\boldsymbol{Y}}_i(\boldsymbol{T}) = \boldsymbol{S}_i(\boldsymbol{T})\boldsymbol{X}_i\boldsymbol{W}_v$, for the loss $\ell_i(\boldsymbol{T}) = \frac{1}{2}\|\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i(\boldsymbol{T})\|_F^2 = \frac{1}{2}\operatorname{Tr}\big(\boldsymbol{E}_i^\top\boldsymbol{E}_i\big)$, take the differential

$$\mathrm{d}\ell_i = \frac{1}{2}\operatorname{Tr}((\mathrm{d}\boldsymbol{E}_i)^\top\boldsymbol{E}_i + \boldsymbol{E}_i^\top\mathrm{d}\boldsymbol{E}_i) = \operatorname{Tr}(\boldsymbol{E}_i^\top\mathrm{d}\boldsymbol{E}_i). \tag{42}$$

Since the target $\boldsymbol{Y}_i$ is a constant, the is simply $\mathrm{d}\boldsymbol{E}_i = \mathrm{d}\hat{\boldsymbol{Y}}_i$. Recalling $\hat{\boldsymbol{Y}}_i(\boldsymbol{T}) = \boldsymbol{S}_i(\boldsymbol{T})\boldsymbol{X}_i\boldsymbol{W}_v$, $\boldsymbol{X}_i$ and $\boldsymbol{W}_v$ are treated as constants with respect to the pre-activations, we have $\mathrm{d}\hat{\boldsymbol{Y}}_i = (\mathrm{d}\boldsymbol{S}_i)\boldsymbol{X}_i\boldsymbol{W}_v$. So the differential of the loss ℓ_i with respect to $\boldsymbol{S}_i$ is given by:

$$\mathrm{d}\ell_i = \operatorname{Tr}(\boldsymbol{E}_i^\top\,\mathrm{d}\hat{\boldsymbol{Y}}_i) = \operatorname{Tr}(\boldsymbol{E}_i^\top(\mathrm{d}\boldsymbol{S}_i)\boldsymbol{X}_i\boldsymbol{W}_v). \tag{43}$$

Using the cyclic property of the trace, we can rearrange this as $\mathrm{d}\ell_i = \operatorname{Tr}(\boldsymbol{X}_i\boldsymbol{W}_v\boldsymbol{E}_i^\top\,\mathrm{d}\boldsymbol{S}_i) = \operatorname{Tr}((\boldsymbol{E}_i\boldsymbol{W}_v^\top\boldsymbol{X}_i^\top)^\top\,\mathrm{d}\boldsymbol{S}_i)$. By identifying this with the standard Frobenius inner product $\mathrm{d}\ell_i = \operatorname{Tr}((\nabla_{\boldsymbol{S}_i}\ell_i)^\top\,\mathrm{d}\boldsymbol{S}_i)$, we extract the exact gradient,

$$\nabla_{\boldsymbol{S}_i}\ell_i = \boldsymbol{E}_i\boldsymbol{W}_v^\top\boldsymbol{X}_i^\top. \tag{44}$$

So we have,

$$\|\nabla_{\boldsymbol{S}_i}\ell_i\|_F = \|\boldsymbol{E}_i\boldsymbol{W}_v^\top\boldsymbol{X}_i^\top\|_F \le \|\boldsymbol{E}_i\|_F\|\boldsymbol{W}_v\|_F\|\boldsymbol{X}_i\|_F \le (\sqrt{t}B_xB_w + B_y)B_wB_x. \tag{45}$$

• **Bounding the Intermediate Gradient:** Utilizing the exact chain rule $\nabla_{\boldsymbol{M}_i}\ell_i = (\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{M}_i})^\top\nabla_{\boldsymbol{S}_i}\ell_i$ and substituting the Softmax Jacobian bound $\|(\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{M}_i})\|_F \le \beta B_{s'}$, we obtain the intermediate gradient bound: $\|\nabla_{\boldsymbol{M}_i}\ell_i\|_F \le (\sqrt{t}B_xB_w + B_y)B_wB_x\beta B_{s'}$ (see Equation 49). Recall that the local Jacobian $(\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{M}_i}) \in \mathbb{R}^{t\times t}$ and its bound are defined in Lemma A.2.

Since the loss differential can be equivalently expressed in terms of either $\boldsymbol{M}_i$ or $\boldsymbol{S}_i$

$$\mathrm{d}\ell_i = \langle \nabla_{\boldsymbol{M}_i} \ell_i, \mathrm{d}\boldsymbol{M}_i \rangle = \langle \nabla_{\boldsymbol{S}_i} \ell_i, \mathrm{d}\boldsymbol{S}_i \rangle, \tag{46}$$

express this in trace form, we have

$$\mathrm{Tr}\left((\nabla_{\boldsymbol{M}_i} \ell_i)^\top \mathrm{d}\boldsymbol{M}_i\right) = \mathrm{Tr}\left((\nabla_{\boldsymbol{S}_i} \ell_i)^\top \mathrm{d}\boldsymbol{S}_i\right) = \mathrm{Tr}\left((\nabla_{\boldsymbol{S}_i} \ell_i)^\top (\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{M}_i}) \mathrm{d}\boldsymbol{M}_i\right) = \mathrm{Tr}\left((\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{M}_i})^\top \nabla_{\boldsymbol{S}_i} \ell_i)^\top \mathrm{d}\boldsymbol{M}_i\right), \tag{47}$$

we can extract

$$\nabla_{\boldsymbol{M}_i} \ell_i = (\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{M}_i})^\top \nabla_{\boldsymbol{S}_i} \ell_i. \tag{48}$$

Take Frobenius norm of both sides and apply the Cauchy-Schwarz inequality, we have

$$\|\nabla_{\boldsymbol{M}_i} \ell_i\|_F \leq (\sqrt{t} B_x B_w + B_y) B_w B_x \beta B_{s'}. \tag{49}$$

• **Bounding the Parameter Gradients:** By extracting the gradient with respect to the query and key weight matrices, $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$, we obtain the bound of $\|\nabla_{\boldsymbol{W}_Q} \ell_i\|_F$ and $\|\nabla_{\boldsymbol{W}_K} \ell_i\|_F$ using the bound of $\|\nabla_{\boldsymbol{M}_i} \ell_i\|_F$, then we substitute the intermediate bound to achieve the final bound $\|\nabla_{\boldsymbol{T}} \ell_i\|^2 = \|\nabla_{\boldsymbol{W}_Q} \ell_i\|_F^2 + \|\nabla_{\boldsymbol{W}_K} \ell_i\|_F^2$, and hence the gradient squared bound (see Equation 58).

Next we fix $\boldsymbol{W}_k$ to analyze the partial derivative with respect to $\boldsymbol{W}_Q$. To see that first consider the change in $\boldsymbol{M}_i$ due to variation in $\boldsymbol{W}_Q$,

$$\mathrm{d}_{\boldsymbol{W}_Q} \boldsymbol{M}_i = \frac{1}{\sqrt{d}} \boldsymbol{X}_i (\mathrm{d}\boldsymbol{W}_Q) \boldsymbol{W}_K^\top \boldsymbol{X}_i^\top, \tag{50}$$

Since,

$$\mathrm{d}\ell_i = \langle \nabla_{\boldsymbol{M}_i} \ell_i, \mathrm{d}_{\boldsymbol{W}_Q} \boldsymbol{M}_i \rangle = \mathrm{Tr}((\nabla_{\boldsymbol{M}_i} \ell_i)^\top \mathrm{d}_{\boldsymbol{W}_Q} \boldsymbol{M}_i) = \mathrm{Tr}((\nabla_{\boldsymbol{M}_i} \ell_i)^\top \frac{1}{\sqrt{d}} \boldsymbol{X}_i (\mathrm{d}\boldsymbol{W}_Q) \boldsymbol{W}_K^\top \boldsymbol{X}_i^\top), \tag{51}$$

apply the cyclic property of trace, we have

$$\mathrm{d}\ell_i = \frac{1}{\sqrt{d}} \mathrm{Tr}\left(\boldsymbol{W}_K^\top \boldsymbol{X}_i^\top (\nabla_{\boldsymbol{M}_i} \ell_i)^\top \boldsymbol{X}_i (\mathrm{d}\boldsymbol{W}_Q)\right). \tag{52}$$

Compare this with the differential directly expressed in terms of the gradient with respect to $\boldsymbol{W}_Q$ as

$$\mathrm{d}\ell_i = \langle \nabla_{\boldsymbol{W}_Q} \ell_i, \mathrm{d}\boldsymbol{W}_Q \rangle = \mathrm{Tr}\left((\nabla_{\boldsymbol{W}_Q} \ell_i)^\top \mathrm{d}\boldsymbol{W}_Q\right), \tag{53}$$

we have $(\nabla_{\boldsymbol{W}_Q} \ell_i)^\top = \frac{1}{\sqrt{d}} \boldsymbol{W}_K^\top \boldsymbol{X}_i^\top (\nabla_{\boldsymbol{M}_i} \ell_i)^\top \boldsymbol{X}_i$, or

$$\nabla_{\boldsymbol{W}_Q} \ell_i = \frac{1}{\sqrt{d}} \boldsymbol{X}_i^\top (\nabla_{\boldsymbol{M}_i} \ell_i) \boldsymbol{X}_i \boldsymbol{W}_K. \tag{54}$$

By taking Frobenius norm of both sides and applying the Cauchy-Schwarz inequality, we have:

$$\|\nabla_{\boldsymbol{W}_Q} \ell_i\|_F \leq \frac{1}{\sqrt{d}} \|\boldsymbol{X}_i^\top\|_F \cdot \|\nabla_{\boldsymbol{M}_i} \ell_i\|_F \cdot \|\boldsymbol{X}_i\|_F \cdot \|\boldsymbol{W}_K\|_F = \|\nabla_{\boldsymbol{M}_i} \ell_i\|_F \frac{B_x^2}{\sqrt{d}} \|\boldsymbol{W}_K\|_F. \tag{55}$$

Finally, substituting the upper bound $\|\nabla_{\boldsymbol{M}_i} \ell_i\|_F$ derived from equation 49, we have:

$$\|\nabla_{\boldsymbol{W}_Q} \ell_i\|_F \leq (\sqrt{t} B_x B_w + B_y) B_w B_x \beta B_{s'} \frac{B_x^2}{\sqrt{d}} \|\boldsymbol{W}_K\|_F. \tag{56}$$

By symmetry,

$$\|\nabla_{\boldsymbol{W}_K}\ell_i\|_F \le (\sqrt{t}B_xB_w + B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\|\boldsymbol{W}_Q\|_F. \tag{57}$$

Combining equations 56 and 57, we have,

$$\|\nabla_{\boldsymbol{T}}\ell_i\|^2 = \|\nabla_{\boldsymbol{W}_Q}\ell_i\|_F^2 + \|\nabla_{\boldsymbol{W}_K}\ell_i\|_F^2 \le \left[(\sqrt{t}B_xB_w + B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\right]^2(\|\boldsymbol{W}_K\|_F^2 + \|\boldsymbol{W}_Q\|_F^2) \tag{58}$$

$$= \left[(\sqrt{t}B_xB_w + B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\right]^2\|\boldsymbol{T}\|^2. \tag{59}$$

So for $\|\nabla_{\boldsymbol{T}}\hat{R}_A\|$, we have,

$$\|\nabla_{\boldsymbol{T}}\hat{R}_A\| \le \frac{1}{n}\sum_{i=1}^n\|\nabla_{\boldsymbol{T}}\ell_i(\boldsymbol{T})\| \le \left((\sqrt{t}B_xB_w + B_y)B_wB_x\beta B_{s'}\frac{B_x^2}{\sqrt{d}}\right)\|\boldsymbol{T}\|. \tag{60}$$

Thus we have proven equation 13 and next we prove equation 14.

• **Bounding the Laplacian of the Loss** For the second-order variations, we expand the Laplacian using the chain rule. By substituting both the Jacobian bound ($\beta B_{s'}$) and the Softmax Hessian bound ($\beta^2 B_{s''}$) from Lemma A.2 into $|\Delta_{\boldsymbol{W}_Q}\ell_i|$ and $|\Delta_{\boldsymbol{W}_K}\ell_i|$, we obtain the Laplacian bound.

Recalling, $\ell_i(\boldsymbol{T}) = \frac{1}{2}\|\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i(\boldsymbol{T})\|_F^2 = \frac{1}{2}\operatorname{Tr}\big(\boldsymbol{E}_i^\top\boldsymbol{E}_i\big)$, firstly we observe that,

$$\frac{\partial\ell_i}{\partial\boldsymbol{W}_{Q\,jk}} = \frac{1}{2}\sum_{a,b}2\boldsymbol{E}_{i,ab}\cdot\frac{\partial\boldsymbol{E}_{i,ab}}{\partial\boldsymbol{W}_{Q\,jk}} = \sum_{a,b}\boldsymbol{E}_{i,ab}\cdot\frac{\partial(\hat{\boldsymbol{Y}}_{i,ab} - \boldsymbol{Y}_{i,ab})}{\partial\boldsymbol{W}_{Q\,jk}}, \tag{61}$$

where $\boldsymbol{W}_{Q\,jk}$ is the $(j,k)-$th element of the matrix $\boldsymbol{W}_Q$, which is a scalar.

Note that $\boldsymbol{Y}_{i,ab}$ is a constant and hence,

$$\frac{\partial^2\ell_i}{\partial\boldsymbol{W}_{Q\,jk}^2} = \sum_{a,b}\left(\left[\frac{\partial\boldsymbol{E}_{i,ab}}{\partial\boldsymbol{W}_{Q\,jk}}\right]\cdot\frac{\partial\hat{\boldsymbol{Y}}_{i,ab}}{\partial\boldsymbol{W}_{Q\,jk}} + \boldsymbol{E}_{i,ab}\cdot\left[\frac{\partial^2\hat{\boldsymbol{Y}}_{i,ab}}{\partial\boldsymbol{W}_{Q\,jk}^2}\right]\right)$$

$$= \sum_{a,b}\left(\frac{\partial\hat{\boldsymbol{Y}}_{i,ab}}{\partial\boldsymbol{W}_{Q\,jk}}\right)^2 + \sum_{a,b}\left(\boldsymbol{E}_{i,ab}\cdot\frac{\partial^2\hat{\boldsymbol{Y}}_{i,ab}}{\partial\boldsymbol{W}_{Q\,jk}^2}\right) = \left\|\frac{\partial\hat{\boldsymbol{Y}}_i}{\partial\boldsymbol{W}_{Q\,jk}}\right\|_F^2 + \operatorname{Tr}\left(\boldsymbol{E}_i^\top\frac{\partial^2\hat{\boldsymbol{Y}}_i}{\partial\boldsymbol{W}_{Q\,jk}^2}\right). \tag{62}$$

Towards analyzing the second term in the RHS above, consider the following derivative of $\boldsymbol{S}_i = \text{RowSoftMax}(\boldsymbol{M}_i)$.

$$\frac{\partial\boldsymbol{S}_{i,ab}}{\partial\boldsymbol{W}_{Q\,jk}} = \sum_{c,d}\left(\frac{\partial\boldsymbol{S}_{i,ab}}{\partial\boldsymbol{M}_{i,cd}}\cdot\frac{\partial\boldsymbol{M}_{i,cd}}{\partial\boldsymbol{W}_{Q\,jk}}\right) =: \nabla\text{Softmax}\left[\frac{\partial\boldsymbol{M}_i}{\partial\boldsymbol{W}_{Q\,jk}}\right]. \tag{63}$$

We have denoted the RHS to be $\nabla\text{Softmax}\left[\frac{\partial\boldsymbol{M}_i}{\partial\boldsymbol{W}_{Q\,jk}}\right]$. This means the perturbation of the underlying weight $\boldsymbol{W}_{Q\,jk}$ induces a directional change in the intermediate pre-activation matrix $\boldsymbol{M}_i$. By passing this direction through the Jacobian tensor of the Softmax operator and performing a tensor contraction, we compute the exact resulting variation in the output probability matrix $\boldsymbol{S}_i$.

To compute the second derivative $\frac{\partial^2\boldsymbol{S}_i}{\partial\boldsymbol{W}_{Q\,jk}^2}$, we differentiate the above result with respect to $\boldsymbol{W}_{Q\,jk}$ again to get,

$$\frac{\partial^2\boldsymbol{S}_{i,ab}}{\partial\boldsymbol{W}_{Q\,jk}^2} = \sum_{c,d}\sum_{e,f}\left(\frac{\partial^2\boldsymbol{S}_{i,ab}}{\partial\boldsymbol{M}_{i,cd}\partial\boldsymbol{M}_{i,ef}}\cdot\frac{\partial\boldsymbol{M}_{i,cd}}{\partial\boldsymbol{W}_{Q\,jk}}\cdot\frac{\partial\boldsymbol{M}_{i,ef}}{\partial\boldsymbol{W}_{Q\,jk}}\right) + \sum_{c,d}\left(\frac{\partial\boldsymbol{S}_{i,ab}}{\partial\boldsymbol{M}_{i,cd}}\cdot\frac{\partial^2\boldsymbol{M}_{i,cd}}{\partial\boldsymbol{W}_{Q\,jk}^2}\right). \tag{64}$$

As before, we denote the RHS in tensor notation as "$\nabla^2\text{Softmax}\left[\frac{\partial\boldsymbol{M}_i}{\partial\boldsymbol{W}_{Q\,jk}}, \frac{\partial\boldsymbol{M}_i}{\partial\boldsymbol{W}_{Q\,jk}}\right] + \nabla\text{Softmax}\left[\frac{\partial^2\boldsymbol{M}_i}{\partial\boldsymbol{W}_{Q\,jk}^2}\right]$"— the first term represents the second-order sensitivity of the Softmax operator (a 6th-order Hessian tensor) acting simultaneously as a

bilinear map on two identical first-order directional perturbations and the second term represents the first-order sensitivity (Jacobian tensor) acting on the second-order perturbation of the intermediate matrix $\boldsymbol{M}_i$. Recalling $\hat{\boldsymbol{Y}}_i(\boldsymbol{T}) = \boldsymbol{S}_i(\boldsymbol{T})\boldsymbol{X}_i\boldsymbol{W}_v$, and using the above notation we have

$$\frac{\partial^2 \hat{\boldsymbol{Y}}_i}{\partial \boldsymbol{W}_{Q_{jk}}^2} = \frac{\partial^2 \hat{\boldsymbol{S}}_i}{\partial \boldsymbol{W}_{Q_{jk}}^2}\boldsymbol{X}_i\boldsymbol{W}_v = \frac{\partial}{\partial \boldsymbol{W}_{Q_{jk}}}\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{W}_{Q_{jk}}}\boldsymbol{X}_i\boldsymbol{W}_v = \frac{\partial}{\partial \boldsymbol{W}_{Q_{jk}}}\left(\nabla \text{Softmax}\left[\frac{\partial \boldsymbol{M}_i}{\partial \boldsymbol{W}_{Q_{jk}}}\right]\right)\boldsymbol{X}_i\boldsymbol{W}_v$$
$$= \left(\nabla^2 \text{Softmax}\left[\frac{\partial \boldsymbol{M}_i}{\partial \boldsymbol{W}_{Q_{jk}}}, \frac{\partial \boldsymbol{M}_i}{\partial \boldsymbol{W}_{Q_{jk}}}\right] + \nabla \text{Softmax}\left[\frac{\partial^2 \boldsymbol{M}_i}{\partial \boldsymbol{W}_{Q_{jk}}^2}\right]\right)\boldsymbol{X}_i\boldsymbol{W}_v. \tag{65}$$

Note that $\boldsymbol{M}_i = \frac{1}{\sqrt{d}}\boldsymbol{X}_i\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}_i^\top$ is linear in $\boldsymbol{W}_Q$, so $\frac{\partial^2 \boldsymbol{M}_i}{\partial \boldsymbol{W}_{Q_{jk}}^2} \equiv 0$. From equation 64, using equation 50 and the Hessian bound from Lemma A.2, we have

$$\sum_{j,k}\left\|\frac{\partial^2 \boldsymbol{S}_i}{\partial \boldsymbol{W}_{Q_{jk}}^2}\right\|_F \le \sum_{j,k}\|\nabla^2 \text{Softmax}(\boldsymbol{M}_i)\|_F \cdot \left\|\frac{\partial \boldsymbol{M}_i}{\partial \boldsymbol{W}_{Q_{jk}}}\right\|_F \cdot \left\|\frac{\partial \boldsymbol{M}_i}{\partial \boldsymbol{W}_{Q_{jk}}}\right\|_F \le \beta^2 B_{s''}\sum_{j,k}\left\|\frac{\partial \boldsymbol{M}_i}{\partial \boldsymbol{W}_{Q_{jk}}}\right\|_F^2 \le \beta^2 B_{s''}\frac{B_x^4}{d}\|\boldsymbol{W}_K\|_F^2. \tag{66}$$

So

$$|\Delta_{\boldsymbol{W}_Q}\ell_i| \le (\beta B_{s'}B_xB_w)^2\left[\frac{B_x^2}{\sqrt{d}}\right]^2\|\boldsymbol{W}_K\|_F^2 + (\sqrt{t}B_xB_w + B_y)B_xB_w\beta^2 B_{s''}\frac{B_x^4}{d}\|\boldsymbol{W}_K\|_F^2. \tag{67}$$

By symmetry,

$$|\Delta_{\boldsymbol{W}_K}\ell_i| \le (\beta B_{s'}B_xB_w)^2\left[\frac{B_x^2}{\sqrt{d}}\right]^2\|\boldsymbol{W}_Q\|_F^2 + (\sqrt{t}B_xB_w + B_y)B_xB_w\beta^2 B_{s''}\frac{B_x^4}{d}\|\boldsymbol{W}_Q\|_F^2. \tag{68}$$

We have,

$$\|\Delta_{\boldsymbol{T}}\hat{R}_A\| \le \|\frac{1}{n}\sum_{i=1}^n \Delta_{\boldsymbol{T}}\ell_i(T)\| \le \|\Delta_T\ell_i(T)\| \le \frac{B_x^4}{d}\left((\beta B_{s'}B_xB_w)^2 + (\sqrt{t}B_xB_w + B_y)B_xB_w\beta^2 B_{s''}\right)\|\boldsymbol{T}\|^2. \tag{69}$$

□

C. Proof of Villani Conditions for Regression on Shallow Nets Under LoRA Constraints

This appendix provides the formal proof of Theorem 3.2, which establishes that the factor-regularized loss landscapes of shallow neural networks under LoRA constraints satisfy the Villani condition. By isolating the upper bound of the neural network's data fitting term, we can demonstrate how the factor regularization dominates as the parameters grow to infinity.

Lemma C.1. *The corrected potential $V_\epsilon(\boldsymbol{T})$ satisfies the confining condition for all $\lambda > 0$.*

Lemma C.2. *For the i-th sample, the input is $\boldsymbol{x}_i \in \mathbb{R}^d$. Define $\boldsymbol{h}_i := \boldsymbol{V}^\top\boldsymbol{x}_i \in \mathbb{R}^r$, $\boldsymbol{s}_i := \boldsymbol{U}\boldsymbol{h}_i \in \mathbb{R}^p$, and*

$$\boldsymbol{z}_i(\boldsymbol{U},\boldsymbol{V}) := \boldsymbol{a}^\top\sigma(\boldsymbol{s}_i) = \sum_{j=1}^p a_j\sigma(s_{i,j}),$$

where σ acts component-wise. Defining $\boldsymbol{g}_i := \boldsymbol{a}\odot\sigma'(\boldsymbol{s}_i) \in \mathbb{R}^p$, it follows that the gradients of $\boldsymbol{z}_i$ with respect to $\boldsymbol{U} \in \mathbb{R}^{p\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ are $\nabla_{\boldsymbol{U}}\boldsymbol{z}_i = \boldsymbol{g}_i\boldsymbol{h}_i^\top$ and $\nabla_{\boldsymbol{V}}\boldsymbol{z}_i = \boldsymbol{x}_i(\boldsymbol{U}^\top\boldsymbol{g}_i)^\top$, that is

$$\nabla_{\boldsymbol{U}}\boldsymbol{z}_i = [\boldsymbol{a}\odot\sigma'(\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{x}_i)](\boldsymbol{V}^\top\boldsymbol{x}_i)^\top \quad \textit{and} \quad \nabla_{\boldsymbol{V}}\boldsymbol{z}_i = \boldsymbol{x}_i[\boldsymbol{a}\odot\sigma'(\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{x}_i)]^\top\boldsymbol{U}. \tag{70}$$

Lemma C.3. *The bound of $\ell'(\boldsymbol{z}_i) = (\boldsymbol{z}_i - \boldsymbol{y}_i)$ is $\sup_{i\in\{1,\dots,n\}}|\boldsymbol{z}_i - \boldsymbol{y}_i| = B_0 = \|\boldsymbol{a}\|_2\sqrt{p}B_\sigma + B_y < \infty$.*

Lemma C.4. *The upper bound of $\|\nabla_{\boldsymbol{T}}\boldsymbol{z}_i\|$ is given by*

$$\|\nabla_{\boldsymbol{T}}\boldsymbol{z}_i\| \le B_{\sigma'}B_x\|\boldsymbol{a}\|_2\|\boldsymbol{T}\| \quad \forall i = 1,2,\dots,n, \tag{71}$$

and hence the upper bound of $\|\nabla_{\boldsymbol{T}}\mathcal{L}\|$ is given by

$$\|\nabla_{\boldsymbol{T}}\mathcal{L}\| \le B_0B_{\sigma'}B_x\|\boldsymbol{a}\|_2\|\boldsymbol{T}\|. \tag{72}$$

Lemma C.5. *The upper bound of* $|\Delta_{\boldsymbol{T}} \boldsymbol{z}_i|$ *is given by*

$$|\Delta_{\boldsymbol{T}} \boldsymbol{z}_i| \leq B_{\sigma''} \|\boldsymbol{a}\|_1 B_x^2 \|\boldsymbol{T}\|^2 \quad \forall i = 1, 2, \ldots, n, \tag{73}$$

and hence the upper bound of the absolute Laplacian $|\Delta_{\boldsymbol{T}} \mathcal{L}|$ *is given by*

$$|\Delta_{\boldsymbol{T}} \mathcal{L}| = \left| \frac{1}{n} \sum_{i=1}^{n} \left(\|\nabla_{\boldsymbol{T}} \boldsymbol{z}_i\|^2 + (\boldsymbol{z}_i - \boldsymbol{y}_i) \Delta_{\boldsymbol{T}} \boldsymbol{z}_i \right) \right| \leq \left((B_{\sigma'} B_x \|\boldsymbol{a}\|_2)^2 + B_0 B_{\sigma''} \|\boldsymbol{a}\|_1 B_x^2 \right) \|\boldsymbol{T}\|^2. \tag{74}$$

The above lemmas are proved in Appendix D.

C.1. Proof of Theorem 3.2 for Loss in Definition 2.6

Proof. We analyze the potential defined in Definition 2.6:

$$\tilde{V}(\boldsymbol{T}) = \mathcal{L}(\boldsymbol{T}) + \frac{\lambda}{2} \|\boldsymbol{T}\|^2 \log\left(1 + \|\boldsymbol{T}\|^2\right), \tag{75}$$

where $\|\boldsymbol{T}\|$ refers to the 2-norm of $\boldsymbol{T}$ read as a vector.

Let the regularization term be $R(\boldsymbol{T}) = \frac{\lambda}{2} \|\boldsymbol{T}\|^2 \log\left(1 + \|\boldsymbol{T}\|^2\right)$.

• **Analysis of the Gradient Term:** The gradient of the regularization term is:

$$\nabla R(\boldsymbol{T}) = \lambda \boldsymbol{T} \log\left(1 + \|\boldsymbol{T}\|^2\right) + \lambda \boldsymbol{T} \frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2}. \tag{76}$$

Taking the norm and lower-bounding it, we get:

$$\|\nabla R(\boldsymbol{T})\| \geq \lambda \|\boldsymbol{T}\| \log\left(1 + \|\boldsymbol{T}\|^2\right). \tag{77}$$

By the gradient bound in Lemma C.4 (Equation 72), we have:

$$\begin{aligned}
\|\nabla \tilde{V}(\boldsymbol{T})\|^2 &= \left(\|\nabla R(\boldsymbol{T})\| + \|\nabla \mathcal{L}(\boldsymbol{T})\|\right)^2 \\
&\geq \|\nabla R(\boldsymbol{T})\|^2 - 2\|\nabla R(\boldsymbol{T})\| \|\nabla \mathcal{L}(\boldsymbol{T})\| \\
&\geq \left(\|\nabla R(\boldsymbol{T})\|\right)^2 - 2\left(\|\nabla R(\boldsymbol{T})\|\right) \sup\left(\|\nabla \mathcal{L}(\boldsymbol{T})\|\right) \\
&= \lambda^2 \|\boldsymbol{T}\|^2 \log\left(1 + \|\boldsymbol{T}\|^2\right)^2 - 2B_0 B_{\sigma'} B_x \|\boldsymbol{a}\|_2 \|\boldsymbol{T}\|^2 \lambda \left(\log\left(1 + \|\boldsymbol{T}\|^2\right) + \frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2} \right).
\end{aligned} \tag{78}$$

• **Analysis of the Laplacian Term:** To compute the Laplacian of the regularization term, we take the divergence of $\nabla R(\boldsymbol{T})$:

$$\Delta R(\boldsymbol{T}) = \sum_{k=1}^{D} \frac{\partial}{\partial T_k} \left(\lambda T_k \left[\log\left(1 + \|\boldsymbol{T}\|^2\right) + \frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2} \right] \right). \tag{79}$$

Applying the product rule yields:

$$\begin{aligned}
\Delta R(\boldsymbol{T}) &= D\lambda \left[\log\left(1 + \|\boldsymbol{T}\|^2\right) + \frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2} \right] + \sum_{k=1}^{D} \lambda T_k \left[\frac{2T_k}{1 + \|\boldsymbol{T}\|^2} + \frac{2T_k}{(1 + \|\boldsymbol{T}\|^2)^2} \right] \\
&= D\lambda \left[\log\left(1 + \|\boldsymbol{T}\|^2\right) + \frac{\|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2} \right] + \frac{2\lambda \|\boldsymbol{T}\|^2}{1 + \|\boldsymbol{T}\|^2} + \frac{2\lambda \|\boldsymbol{T}\|^2}{(1 + \|\boldsymbol{T}\|^2)^2}.
\end{aligned} \tag{80}$$

Observe that as $\|\boldsymbol{T}\| \to \infty$, $\Delta R(\boldsymbol{T}) = \mathcal{O}(\log \|\boldsymbol{T}\|^2)$. Applying Lemma C.5 (Equation 109), the Laplacian of the data term is bounded by a quadratic. Therefore, the total Laplacian is bounded by:

$$\Delta \tilde{V}(\boldsymbol{T}) \leq \left((B_{\sigma'} B_x \|\boldsymbol{a}\|_2)^2 + B_0 B_{\sigma''} \|\boldsymbol{a}\|_1 B_x^2 \right) \|\boldsymbol{T}\|^2 + \Delta R(\boldsymbol{T}). \tag{81}$$

• **Verifying the Villani Condition:** Substituting Equations 78 and 81 into the Villani limit expression for a given $s > 0$:

$$\begin{aligned}\lim_{\|\boldsymbol{T}\|\to\infty}\left(\frac{1}{s}\|\nabla\tilde{V}(\boldsymbol{T})\|^2-\Delta\tilde{V}(\boldsymbol{T})\right) \geq \lim_{\|\boldsymbol{T}\|\to\infty}\Bigg[&\frac{1}{s}\Bigg(\lambda^2\|\boldsymbol{T}\|^2\log\big(1+\|\boldsymbol{T}\|^2\big)^2\\ &-2B_0B_{\sigma'}B_x\|\boldsymbol{a}\|_2\lambda\|\boldsymbol{T}\|^2\left(\log\big(1+\|\boldsymbol{T}\|^2\big)+\frac{\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}\right)\Bigg)\\ &-\left(\Big((B_{\sigma'}B_x\|\boldsymbol{a}\|_2)^2+B_0B_{\sigma''}\|\boldsymbol{a}\|_1B_x^2\Big)\|\boldsymbol{T}\|^2+\Delta R(\boldsymbol{T})\right)\Bigg]\\ =\lim_{\|\boldsymbol{T}\|\to\infty}\|\boldsymbol{T}\|^2\Bigg[&\underbrace{\frac{\lambda^2}{s}(\log\big(1+\|\boldsymbol{T}\|^2\big))^2}_{\text{Eq. 78}\to+\infty\text{ (Dominant)}}\\ &\underbrace{-\frac{2\lambda B_0B_{\sigma'}B_x\|\boldsymbol{a}\|_2}{s}\left(\log\big(1+\|\boldsymbol{T}\|^2\big)+\frac{\|\boldsymbol{T}\|^2}{1+\|\boldsymbol{T}\|^2}\right)}_{\text{Eq. 78}\to+\infty\text{ (slower)}}\\ &\underbrace{-\Big((B_{\sigma'}B_x\|\boldsymbol{a}\|_2)^2+B_0B_{\sigma''}\|\boldsymbol{a}\|_1B_x^2\Big)}_{\text{Eq. 81 Constant}}\underbrace{-\frac{\Delta R(\boldsymbol{T})}{\|\boldsymbol{T}\|^2}}_{\text{Eq. 80}\to 0}\Bigg].\end{aligned} \tag{82}$$

The limit does diverge to $+\infty$:

$$\lim_{\|\boldsymbol{T}\|\to\infty}\left(\frac{1}{s}\|\nabla\tilde{V}(\boldsymbol{T})\|^2-\Delta\tilde{V}(\boldsymbol{T})\right)=\infty. \tag{83}$$

This demonstrates that the Villani condition is satisfied for all $\lambda > 0$ and $s > 0$. □

C.2. Proof of Theorem 3.2 for Loss in Definition 2.7

Proof. We recall from Definition 2.7 that,

$$V_\epsilon(\boldsymbol{T})=\mathcal{L}(\boldsymbol{T})+\frac{\lambda}{2}(\|\boldsymbol{U}\|_F^{2+\epsilon}+\|\boldsymbol{V}\|_F^{2+\epsilon}).$$

By Lemma C.1 we know that V_ϵ is a confining function. Now, to show that $V_\epsilon(\boldsymbol{T})$ is Villani, we have to verify if the following is satisfied:

$$\frac{\|\nabla V_\epsilon(\boldsymbol{T})\|^2}{s}-\Delta V_\epsilon(\boldsymbol{T})\to\infty \quad \text{as} \quad \|\boldsymbol{T}\|\to\infty. \tag{84}$$

• **Analysis of the Gradient Term:** Let's analyze the asymptotic behavior of the quantity $\mathcal{L}(\boldsymbol{T})+\frac{\lambda}{2}(\|\boldsymbol{U}\|_F^{2+\epsilon}+\|\boldsymbol{V}\|_F^{2+\epsilon})$.

The gradient of the data term is,

$$\nabla\mathcal{L}(\boldsymbol{T})=\frac{1}{n}\sum_{i=1}^{n}\ell_i'(\boldsymbol{z}_i)\nabla_{\boldsymbol{T}}(\boldsymbol{a}^\top\sigma(\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{x}_i)), \tag{85}$$

where $\ell'(\boldsymbol{z}_i)=\boldsymbol{z}_i-\boldsymbol{y}_i$. The total gradient norm squared is $\|\nabla_{\boldsymbol{T}}\boldsymbol{z}_i\|^2=\|\nabla_{\boldsymbol{U}}\boldsymbol{z}_i\|_F^2+\|\nabla_{\boldsymbol{V}}\boldsymbol{z}_i\|_F^2$.

Since we have

$$\begin{aligned}\nabla\left[\frac{\lambda}{2}(\|\boldsymbol{U}\|_F^{2+\epsilon}+\|\boldsymbol{V}\|_F^{2+\epsilon})\right]&=\left(\nabla_{\boldsymbol{U}}\left[\frac{\lambda}{2}(\|\boldsymbol{U}\|_F^2)^{\frac{2+\epsilon}{2}}\right],\nabla_{\boldsymbol{V}}\left[\frac{\lambda}{2}(\|\boldsymbol{V}\|_F^2)^{\frac{2+\epsilon}{2}}\right]\right)\\ &=\frac{\lambda}{2}\left(\frac{2+\epsilon}{2}(\|\boldsymbol{U}\|_F^2)^{\frac{\epsilon}{2}}\cdot\nabla_{\boldsymbol{U}}(\|\boldsymbol{U}\|_F^2),\frac{2+\epsilon}{2}(\|\boldsymbol{V}\|_F^2)^{\frac{\epsilon}{2}}\cdot\nabla_{\boldsymbol{V}}(\|\boldsymbol{V}\|_F^2)\right)\\ &=\frac{\lambda}{2}\left(\frac{2+\epsilon}{2}\|\boldsymbol{U}\|_F^{\epsilon}\cdot(2\boldsymbol{U}),\frac{2+\epsilon}{2}\|\boldsymbol{V}\|_F^{\epsilon}\cdot(2\boldsymbol{V})\right)\\ &=\frac{\lambda}{2}\left((2+\epsilon)\|\boldsymbol{U}\|_F^{\epsilon}\boldsymbol{U},(2+\epsilon)\|\boldsymbol{V}\|_F^{\epsilon}\boldsymbol{V}\right)\end{aligned}, \tag{86}$$

we get,

$$\|\nabla R_\varepsilon(\boldsymbol{T})\|^2 = \left(\frac{\lambda}{2}\right)^2 (2+\varepsilon)^2 \left(\|\boldsymbol{U}\|_F^{2+2\varepsilon} + \|\boldsymbol{V}\|_F^{2+2\varepsilon}\right) \geq \left(\frac{\lambda}{2}\right)^2 (2+\varepsilon)^2 2^{-\varepsilon} \|\boldsymbol{T}\|^{2+2\varepsilon}, \tag{87}$$

where we have used $\frac{(\|\boldsymbol{U}\|^2)^{1+\varepsilon}+(\|\boldsymbol{V}\|^2)^{1+\varepsilon}}{2} \geq \left(\frac{\|\boldsymbol{U}\|^2+\|\boldsymbol{V}\|^2}{2}\right)^{1+\varepsilon} = 2^{-1-\varepsilon}\|\boldsymbol{T}\|^{2+2\varepsilon}$ by Jensen's Inequality.

Hence, substituting the gradient bound from Lemma C.4 (Equation 72), the gradient of the total potential $\nabla V_\epsilon = \nabla \mathcal{L} + \lambda \boldsymbol{T}$ satisfies:

$$\begin{aligned}
\|\nabla V_\epsilon\|^2 &= \|\nabla \mathcal{L}\|^2 + \left\|\nabla \left[\frac{\lambda}{2}(\|\boldsymbol{U}\|_F^{2+\epsilon} + \|\boldsymbol{V}\|_F^{2+\epsilon})\right]\right\|^2 \\
&\quad + \lambda \langle \frac{1}{n}\sum_{i=1}^n (\boldsymbol{z}_i - \boldsymbol{y}_i) \nabla_{\boldsymbol{T}} \boldsymbol{z}_i, \left((2+\epsilon)\|\boldsymbol{U}\|_F^{\epsilon}\boldsymbol{U},\ (2+\epsilon)\|\boldsymbol{V}\|_F^{\epsilon}\boldsymbol{V}\right)\rangle \\
&\geq \left(\frac{\lambda}{2}\right)^2 (2+\epsilon)^2 \left(\|\boldsymbol{U}\|_F^{2+2\epsilon} + \|\boldsymbol{V}\|_F^{2+2\epsilon}\right) \\
&\quad - \lambda \langle \frac{1}{n}\sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{z}_i) \nabla_{\boldsymbol{T}} \boldsymbol{z}_i, \left((2+\epsilon)\|\boldsymbol{U}\|_F^{\epsilon}\boldsymbol{U},\ (2+\epsilon)\|\boldsymbol{V}\|_F^{\epsilon}\boldsymbol{V}\right)\rangle \\
&\geq \left(\frac{\lambda}{2}\right)^2 (2+\epsilon)^2 2^{-\varepsilon} \|\boldsymbol{T}\|^{2+2\varepsilon} - \lambda(2+\varepsilon) B_0 B_{\sigma'} B_x \|\boldsymbol{a}\|_2 \|\boldsymbol{T}\|^{2+\varepsilon}.
\end{aligned} \tag{88}$$

In the last inequality, we used $\|\boldsymbol{U}\|_F^{2+2\varepsilon} + \|\boldsymbol{V}\|_F^{2+2\varepsilon} \leq (\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2)^{1+\varepsilon} = \|\boldsymbol{T}\|^{2+2\varepsilon}$.

• **Analysis of the Laplacian Term:** Since $\frac{\partial}{\partial U_{ij}}\|\boldsymbol{U}\|_F^{2+\varepsilon} = (2+\varepsilon)\|\boldsymbol{U}\|_F^{\varepsilon} U_{ij}$ from the analysis above, we have

$$\frac{\partial^2}{\partial U_{ij}^2}\|\boldsymbol{U}\|_F^{2+\varepsilon} = (2+\varepsilon)\varepsilon \|\boldsymbol{U}\|_F^{\varepsilon-2} U_{ij}^2 + (2+\varepsilon)\|\boldsymbol{U}\|_F^{\varepsilon}. \tag{89}$$

So,

$$\Delta_{\boldsymbol{U}}\|\boldsymbol{U}\|_F^{2+\varepsilon} = (2+\varepsilon)\|\boldsymbol{U}\|_F^{\varepsilon} \cdot \varepsilon + (2+\varepsilon)\|\boldsymbol{U}\|_F^{\varepsilon} \cdot pr = (2+\varepsilon)(\varepsilon + pr)\|\boldsymbol{U}\|_F^{\varepsilon}. \tag{90}$$

By symmetry,

$$\Delta_{\boldsymbol{V}}\|\boldsymbol{V}\|_F^{2+\varepsilon} = (2+\varepsilon)(\varepsilon + dr)\|\boldsymbol{V}\|_F^{\varepsilon}. \tag{91}$$

Thus, we have:

$$\begin{aligned}
\Delta R_\varepsilon(\boldsymbol{T}) = \frac{\lambda}{2}\left[\Delta_{\boldsymbol{U}}\|\boldsymbol{U}\|_F^{2+\varepsilon} + \Delta_{\boldsymbol{V}}\|\boldsymbol{V}\|_F^{2+\varepsilon}\right] &= \frac{\lambda}{2}(2+\varepsilon)\left[(\varepsilon+pr)\|\boldsymbol{U}\|_F^{\varepsilon} + (\varepsilon+dr)\|\boldsymbol{V}\|_F^{\varepsilon}\right] \\
&\leq \frac{\lambda}{2}(2+\varepsilon)\left[(\varepsilon+pr)\|\boldsymbol{T}\|^{\varepsilon} + (\varepsilon+dr)\|\boldsymbol{T}\|^{\varepsilon}\right] \\
&= \frac{\lambda}{2}(2+\varepsilon)\left(2\varepsilon + D\right)\|\boldsymbol{T}\|^{\varepsilon},
\end{aligned} \tag{92}$$

where $D = (p+d)r$ is the total parameter dimension.

By Lemma C.5 (Equation 109), we have:

$$|\Delta_{\boldsymbol{T}}\mathcal{L}| \leq \left((B_{\sigma'}B_x\|\boldsymbol{a}\|_2)^2 + B_0 B_{\sigma''}\|\boldsymbol{a}\|_1 B_x^2\right)\|\boldsymbol{T}\|^2. \tag{93}$$

Substituting Equation 88, Equation 92, and the data Laplacian bound into the Villani limit:

$$\lim_{\|\boldsymbol{T}\|\to\infty}\left(\frac{1}{s}\|\nabla V_\epsilon\|^2-\Delta V_\epsilon\right)\geq\lim_{\|\boldsymbol{T}\|\to\infty}\left[\frac{1}{s}\left(\left(\frac{\lambda}{2}\right)^2(2+\epsilon)^2 2^{-\epsilon}\|\boldsymbol{T}\|^{2+2\epsilon}-\lambda(2+\epsilon)B_0B_{\sigma'}B_x\|\boldsymbol{a}\|_2 2^{-\epsilon/2}\|\boldsymbol{T}\|^{2+\epsilon}\right)\right.$$
$$\left.-\left(\left((B_{\sigma'}B_x\|\boldsymbol{a}\|_2)^2+B_0B_{\sigma''}\|\boldsymbol{a}\|_1B_x^2\right)\|\boldsymbol{T}\|^2+\frac{\lambda}{2}(2+\epsilon)(2\epsilon+D)\|\boldsymbol{T}\|^\epsilon\right)\right]$$
$$=\lim_{\|\boldsymbol{T}\|\to\infty}\|\boldsymbol{T}\|^2\left[\underbrace{\frac{1}{s}\left(\frac{\lambda}{2}\right)^2(2+\epsilon)^2 2^{-\epsilon}\|\boldsymbol{T}\|^{2\epsilon}}_{\text{Eq. 88}\to+\infty}-\underbrace{\frac{\lambda}{s}(2+\epsilon)B_0B_{\sigma'}B_x\|\boldsymbol{a}\|_2 2^{-\epsilon/2}\|\boldsymbol{T}\|^{\epsilon}}_{\text{Eq. 88}\to+\infty\text{ (slower)}}\right.$$
$$\left.-\underbrace{\left((B_{\sigma'}B_x\|\boldsymbol{a}\|_2)^2+B_0B_{\sigma''}\|\boldsymbol{a}\|_1B_x^2\right)}_{\text{Eq. 109 Constant}}-\underbrace{\frac{\lambda}{2}(2+\epsilon)(2\epsilon+D)\|\boldsymbol{T}\|^{\epsilon-2}}_{\text{Eq. 92}\to 0}\right] \tag{94}$$
$$=+\infty. \tag{95}$$

Since the quantity tends to positive infinity, the Villani condition is satisfied for all $\lambda>0$ and $\varepsilon>0$. This completes the verification that the potential (Definition 2.7) induces the isoperimetric properties necessary for the Poincaré Inequality to hold. □

D. Proofs of Intermediate Lemmas for Theorem 3.2

Proof of Lemma C.1. The Mean Square loss is non-negative, $\ell_i(\boldsymbol{W})\geq 0$, thus $\mathcal{L}(\boldsymbol{T})\geq 0$. The potential is bounded below by the factor regularization term:

$$V_\epsilon(\boldsymbol{T})\geq\frac{\lambda}{2}\|\boldsymbol{T}\|^2. \tag{96}$$

Since $\lambda>0$, the quadratic growth of $\frac{\lambda}{2}\|\boldsymbol{T}\|^2$ ensures $\lim_{\|\boldsymbol{T}\|\to\infty}V_\epsilon(\boldsymbol{T})=+\infty$. This guarantees integrability of $e^{-\beta V_\epsilon}$ and normalizability of μ_β. □

Proof of Lemma C.2. Since $\mathrm{d}\boldsymbol{s}_i=\mathrm{d}\boldsymbol{U}\,\boldsymbol{h}_i$, we have $\mathrm{d}\boldsymbol{z}_i=(a\odot\sigma'(\boldsymbol{s}_i))^\top\,\mathrm{d}\boldsymbol{s}_i=\boldsymbol{g}_i^\top(\mathrm{d}\boldsymbol{U}\,\boldsymbol{h}_i)$. This can be expressed in a form of inner product

$$\boldsymbol{g}_i^\top(\mathrm{d}\boldsymbol{U}\,\boldsymbol{h}_i)=\mathrm{tr}\left(\boldsymbol{h}_i\boldsymbol{g}_i^\top\,\mathrm{d}\boldsymbol{U}\right)=\langle\boldsymbol{g}_i\boldsymbol{h}_i^\top,\mathrm{d}\boldsymbol{U}\rangle,$$

so

$$\nabla_{\boldsymbol{U}}\boldsymbol{z}_i=\boldsymbol{g}_i\boldsymbol{h}_i^\top$$

Consider the derivative acting on $\boldsymbol{V}$:$\mathrm{d}\boldsymbol{s}_i=\boldsymbol{U}\,\mathrm{d}\boldsymbol{h}_i=\boldsymbol{U}(\mathrm{d}\boldsymbol{V})^\top\boldsymbol{x}_i$, similarly

$$\mathrm{d}\boldsymbol{z}_i=\boldsymbol{g}_i^\top\boldsymbol{U}(\mathrm{d}\boldsymbol{V})^\top\boldsymbol{x}_i=\mathrm{tr}((\boldsymbol{x}_i(\boldsymbol{U}^\top\boldsymbol{g}_i)^\top)^\top\,\mathrm{d}\boldsymbol{V})=\langle\boldsymbol{x}_i(\boldsymbol{U}^\top\boldsymbol{g}_i)^\top,\mathrm{d}\boldsymbol{V}\rangle,$$

so

$$\nabla_{\boldsymbol{V}}\boldsymbol{z}_i=\boldsymbol{x}_i\left(\boldsymbol{U}^\top\boldsymbol{g}_i\right)^\top.$$

So the partial gradients w.r.t. the factors are:

$$\nabla_{\boldsymbol{U}}\boldsymbol{z}_i=[\boldsymbol{a}\odot\sigma'(\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{x}_i)](\boldsymbol{V}^\top\boldsymbol{x}_i)^\top$$
$$\nabla_{\boldsymbol{V}}\boldsymbol{z}_i=\boldsymbol{x}_i[\boldsymbol{a}\odot\sigma'(\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{x}_i)]^\top\boldsymbol{U}$$

□

Proof of Lemma C.3. Apply Cauchy-Schwarz, $|\boldsymbol{z}_i|\leq\|a\|_2\sqrt{p}B_\sigma$, recall $\sigma(\boldsymbol{s}_i)\in\mathbb{R}^p$. Since $|\boldsymbol{y}_i|\leq B_y$. we have

$$|\boldsymbol{z}_i-\boldsymbol{y}_i|\leq B_0=\|a\|_2\sqrt{p}B_\sigma+B_y.$$

□

Proof of Lemma C.4. By Lemma C.2, $\|\nabla_{\boldsymbol{U}} \boldsymbol{z}_i\|_F = \|\boldsymbol{g}_i \boldsymbol{h}_i^\top\|_F = \|\boldsymbol{g}_i\|_2 \|\boldsymbol{h}_i\|_2$, where $\|\boldsymbol{g}_i\|_2 \leq B_{\sigma'} \|a\|_2$, $\|\boldsymbol{h}_i\|_2 = \|\boldsymbol{V}^\top \boldsymbol{x}_i\|_2 \leq \|\boldsymbol{V}^\top\|_2 \|\boldsymbol{x}_i\|_2 = \|\boldsymbol{V}\|_2 \|\boldsymbol{x}_i\|_2 \leq \|\boldsymbol{V}\|_F B_x$. We have

$$\|\nabla_{\boldsymbol{U}} \boldsymbol{z}_i\|_F \leq (B_{\sigma'} \|a\|_2) (B_x \|\boldsymbol{V}\|_F) = B_{\sigma'} B_x \|a\|_2 \|\boldsymbol{V}\|_F. \tag{97}$$

Similarly,

$$\|\nabla_{\boldsymbol{V}} \boldsymbol{z}_i\|_F \leq B_x \|\boldsymbol{U}\|_F (B_{\sigma'} \|a\|_2) = B_{\sigma'} B_x \|a\|_2 \|\boldsymbol{U}\|_F. \tag{98}$$

Since, $\mathrm{d}\boldsymbol{z}_i = \langle \nabla_{\boldsymbol{T}} \boldsymbol{z}_i,\ \mathrm{d}\boldsymbol{T} \rangle$, combine 97 and 98 together to get

$$\|\nabla_{\boldsymbol{T}} \boldsymbol{z}_i\|^2 = \|\nabla_{\boldsymbol{U}} \boldsymbol{z}_i\|_F^2 + \|\nabla_{\boldsymbol{V}} \boldsymbol{z}_i\|_F^2 \leq (B_{\sigma'} B_x \|a\|_2)^2 (\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2), \tag{99}$$

we have $\|\nabla_{\boldsymbol{T}} \boldsymbol{z}_i\| \leq B_{\sigma'} B_x \|a\|_2 \|\boldsymbol{T}\|$ for $\forall i = 1, 2, \ldots n$. By Lemma C.3, we have,

$$\|\nabla_{\boldsymbol{T}} \mathcal{L}\| = \left\| \frac{1}{n} \sum_{i=1}^n (\boldsymbol{z}_i - \boldsymbol{y}_i) \nabla_{\boldsymbol{T}} \boldsymbol{z}_i \right\| \leq \frac{1}{n} \sum_{i=1}^n |\boldsymbol{z}_i - \boldsymbol{y}_i| \, \|\nabla_T \boldsymbol{z}_i\| \leq \frac{1}{n} \sum_{i=1}^n B_0 B_{\sigma'} B_x \|a\|_2 \|\boldsymbol{T}\| = \frac{1}{n} \cdot n \cdot B_0 B_{\sigma'} B_x \|a\|_2 \|\boldsymbol{T}\| \tag{100}$$

□

Proof of Lemma C.5. We evaluate the data term Laplacian as follows:

$$\Delta \mathcal{L}(\boldsymbol{T}) = \frac{1}{n} \sum_{i=1}^n \left(\ell_i''(\boldsymbol{z}_i) \|\nabla_{\boldsymbol{T}} \boldsymbol{z}_i\|^2 + \ell_i'(\boldsymbol{z}_i) \Delta_{\boldsymbol{T}} \boldsymbol{z}_i \right) = \frac{1}{n} \sum_{i=1}^n \left(\|\nabla_{\boldsymbol{T}} \boldsymbol{z}_i\|^2 + (\boldsymbol{z}_i - \boldsymbol{y}_i) \Delta_{\boldsymbol{T}} \boldsymbol{z}_i \right). \tag{101}$$

Recall, $\boldsymbol{h} = \boldsymbol{V}^\top \boldsymbol{x} \in \mathbb{R}^r$, $\boldsymbol{s} = \boldsymbol{U}\boldsymbol{h} \in \mathbb{R}^p$, and $z = \boldsymbol{a}^\top \sigma(\boldsymbol{s})$.

For $\Delta_T \boldsymbol{z}_i = \sum_{j=1}^p \sum_{k=1}^r \frac{\partial^2 \boldsymbol{z}_i}{\partial \boldsymbol{U}_{jk}^2} + \sum_{\ell=1}^d \sum_{k=1}^r \frac{\partial^2 \boldsymbol{z}_i}{\partial \boldsymbol{V}_{\ell k}^2}$, we can analyze term by term.

Since, $\boldsymbol{s}_j = \sum_{k'} \boldsymbol{U}_{jk'} \boldsymbol{h}_{k'}$, we have $\frac{\partial \boldsymbol{s}_j}{\partial \boldsymbol{U}_{jk}} = \boldsymbol{h}_k$, so $\frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{U}_{jk}} = a_j \sigma'(\boldsymbol{s}_j) \frac{\partial \boldsymbol{s}_j}{\partial \boldsymbol{U}_{jk}} = a_j \sigma'(\boldsymbol{s}_j) \boldsymbol{h}_k$,

and hence, $\frac{\partial^2 \boldsymbol{z}_i}{\partial \boldsymbol{U}_{jk}^2} = a_j \sigma''(\boldsymbol{s}_j) \left(\frac{\partial \boldsymbol{s}_j}{\partial \boldsymbol{U}_{jk}} \right)^2 = a_j \sigma''(\boldsymbol{s}_j) \boldsymbol{h}_k^2$.

So,

$$\Delta_{\boldsymbol{U}} \boldsymbol{z}_i = \sum_{j=1}^p \sum_{k=1}^r a_j \sigma''(\boldsymbol{s}_j) \boldsymbol{h}_k^2 = \left(\sum_{k=1}^r \boldsymbol{h}_k^2 \right) \left(\sum_{j=1}^p a_j \sigma''(\boldsymbol{s}_j) \right) = \|\boldsymbol{h}\|_2^2 \sum_{j=1}^p a_j \sigma''(\boldsymbol{s}_j). \tag{102}$$

Similarly, for $\boldsymbol{h}_{k'} = \sum_{\ell'=1}^d \boldsymbol{V}_{\ell' k'} \boldsymbol{x}_{\ell'}$, we can express s with

$$\boldsymbol{s}_j = \sum_{k'=1}^r \boldsymbol{U}_{jk'} \boldsymbol{h}_{k'} = \sum_{k'=1}^r \boldsymbol{U}_{jk'} \left(\sum_{\ell'=1}^d \boldsymbol{V}_{\ell' k'} \boldsymbol{x}_{\ell'} \right),$$

so $\frac{\partial \boldsymbol{s}_j}{\partial \boldsymbol{V}_{\ell k}} = \boldsymbol{U}_{jk} \boldsymbol{x}_\ell$, hence $\frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{V}_{\ell k}} = \sum_{j=1}^p a_j \sigma'(\boldsymbol{s}_j) \frac{\partial \boldsymbol{s}_j}{\partial \boldsymbol{V}_{\ell k}} = \sum_{j=1}^p a_j \sigma'(\boldsymbol{s}_j) \boldsymbol{U}_{jk} \boldsymbol{x}_\ell$ and $\frac{\partial^2 \boldsymbol{z}_i}{\partial \boldsymbol{V}_{\ell k}^2} = \sum_{j=1}^p a_j \left[\sigma''(\boldsymbol{s}_j) \frac{\partial \boldsymbol{s}_j}{\partial \boldsymbol{V}_{\ell k}} \right] \boldsymbol{U}_{jk} \boldsymbol{x}_\ell$.
So

$$\Delta_{\boldsymbol{V}} z = \sum_{\ell=1}^d \sum_{k=1}^r \frac{\partial^2 z}{\partial \boldsymbol{V}_{\ell k}^2} = \sum_{\ell=1}^d \sum_{k=1}^r \left(\sum_{j=1}^p a_j \sigma''(\boldsymbol{s}_j) \boldsymbol{U}_{jk}^2 \boldsymbol{x}_\ell^2 \right) = \sum_{j=1}^p a_j \sigma''(\boldsymbol{s}_j) \left(\sum_{\ell=1}^d \boldsymbol{x}_\ell^2 \right) \left(\sum_{k=1}^r \boldsymbol{U}_{jk}^2 \right), \tag{103}$$

where $\boldsymbol{U}_j = \sum_{k=1}^r \boldsymbol{U}_{jk}^2$.
So $\Delta_T \boldsymbol{z}_i = \Delta_{\boldsymbol{U}} \boldsymbol{z}_i + \Delta_{\boldsymbol{V}} \boldsymbol{z}_i = \|\boldsymbol{h}\|_2^2 \sum_{j=1}^p a_j \sigma''(\boldsymbol{s}_j) + \|\boldsymbol{x}\|_2^2 \sum_{j=1}^p a_j \sigma''(\boldsymbol{s}_j) \|\boldsymbol{U}_j\|_2^2$.
Determine the bound,

$$|\Delta_{\boldsymbol{U}} \boldsymbol{z}_i| = \left| \|\boldsymbol{h}\|_2^2 \sum_j a_j \sigma''(\boldsymbol{s}_j) \right| \leq \|\boldsymbol{h}\|_2^2 \sum_j |a_j| \, |\sigma''(\boldsymbol{s}_j)| \leq \|\boldsymbol{h}\|_2^2 \|a\|_1 B_{\sigma''}. \tag{104}$$

Since $\|\boldsymbol{h}\|_2 = \|\boldsymbol{V}^\top x\|_2 \le \|\boldsymbol{V}\|_2\|\boldsymbol{x}\|_2 \le \|\boldsymbol{V}\|_F B_x$, We have

$$|\Delta_{\boldsymbol{U}} z| \le B_{\sigma''}\|a\|_1 B_x^2 \|\boldsymbol{V}\|_F^2. \tag{105}$$

Similarly,

$$\begin{aligned}|\Delta_{\boldsymbol{V}} z| &= \left|\|\boldsymbol{x}\|_2^2 \sum_j a_j \sigma''(\boldsymbol{s}_j)\|\boldsymbol{U}_j\|_2^2\right| \le \|\boldsymbol{x}\|_2^2 \sum_j |a_j||\sigma''(\boldsymbol{s}_j)|\|\boldsymbol{U}_j\|_2^2 \le B_x^2 B_{\sigma''}\sum_j |a_j|\|\boldsymbol{U}_j\|_2^2 \\ &\le \|a\|_1 \sum_j \|\boldsymbol{U}_j\|_2^2 = \|a\|_1\|\boldsymbol{U}\|_F^2.\end{aligned} \tag{106}$$

Combine equation 105 and 106, we have,

$$|\Delta_T z| \le B_{\sigma''}\|a\|_1 B_x^2\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right) = B_{\sigma''}\|a\|_1 B_x^2\|\boldsymbol{T}\|^2. \tag{107}$$

So for $(\boldsymbol{z}_i - \boldsymbol{y}_i)\Delta_{\boldsymbol{T}}\boldsymbol{z}_i$, By Lemma C.3:

$$|(\boldsymbol{z}-\boldsymbol{y})\Delta_T \boldsymbol{z}| \le |\boldsymbol{z}-\boldsymbol{y}|\,|\Delta_T \boldsymbol{z}| \le B_0 B_{\sigma''}\|a\|_1 B_x^2\|\boldsymbol{T}\|^2. \tag{108}$$

Combine equation 99 and 108,

$$\Delta_T \mathcal{L} = \frac{1}{n}\sum_{i=1}^n \left(\|\nabla_{\boldsymbol{T}}\boldsymbol{z}_i\|^2 + (\boldsymbol{z}_i - \boldsymbol{y}_i)\Delta_T \boldsymbol{z}_i\right) \le \left((B_{\sigma'}B_x\|a\|_2)^2 + B_0 B_{\sigma''}\|a\|_1 B_x^2\right)\|\boldsymbol{T}\|^2. \tag{109}$$

□

E. Proofs of SDE convergence

Proof of Theorem 3.3. Let $\tilde{L}^{(k)}(\boldsymbol{T})$ denote any of the four regularized potentials defined in Definition 2.4 ($\tilde{V}_{\mathrm{ATT}}$), Definition 2.5 ($V_{\epsilon,\mathrm{ATT}}$), Definition 2.6 ($V$), and Definition 2.7 ($V_\epsilon$), where $k \in \{1,2,3,4\}$ indexes the specific model and regularization choice.

For any $k \in \{1,2,3,4\}$, perform the following decomposition of the excess risk:

$$\mathbb{E}[\tilde{L}^{(k)}(\boldsymbol{T}_t)] - \tilde{L}^{(k)\star} = \underbrace{\left(\mathbb{E}[\tilde{L}^{(k)}(X_s^{(k)}(\infty))] - \tilde{L}^{(k)\star}\right)}_{\varepsilon^{(k)}(s)} + \underbrace{\left(\mathbb{E}[\tilde{L}^{(k)}(\boldsymbol{T}_t)] - \mathbb{E}[\tilde{L}^{(k)}(X_s^{(k)}(\infty))]\right)}_{\le D^{(k)}(s,p_0)e^{-\lambda_s^{(k)}t}} \tag{110}$$

By Proposition 5 of (Shi et al., 2023),

$$\varepsilon^{(k)}(s) \le A^{(k)} s \tag{111}$$

and Proposition 4 of (Shi et al., 2023),

$$\mathbb{E}[\tilde{L}^{(k)}(\boldsymbol{T}_t)] - \mathbb{E}[\tilde{L}^{(k)}(X_s^{(k)}(\infty))] \le C^{(k)}(s)\|p_0 - \mu_s^{(k)}\|_{L^2((\mu_s^{(k)})^{-1})} e^{-\lambda_s^{(k)}t}, \tag{112}$$

where $p_0 \in L^2((\mu_s^{(k)})^{-1})$ is the initial probability density of the SDE (1) .

Since $s \le \min\left\{\frac{\epsilon}{2A^{(k)}}, S^{(k)}\right\}$, and the time horizon $t \ge \frac{1}{\lambda_s^{(k)}}\log\left(\frac{2D^{(k)}(s,p_0)}{\epsilon}\right)$, by Corollary 6 of (Shi et al., 2023) we have

$$\mathbb{E}[\tilde{L}^{(k)}(\boldsymbol{T}_t)] - \tilde{L}^{(k)\star} \le \epsilon. \tag{113}$$

□