BEYOND LINEAR APPROXIMATIONS: A NOVEL PRUNING APPROACH FOR ATTENTION MATRIX

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have shown immense potential in enhancing various aspects of our daily lives, from conversational AI to search and AI assistants. However, their growing capabilities come at the cost of extremely large model sizes, making deployment on edge devices challenging due to memory and computational constraints. This paper introduces a novel approach to LLM weight pruning that directly optimizes for approximating the attention matrix, a core component of transformer architectures. Unlike existing methods that focus on linear approximations, our approach accounts for the non-linear nature of the Softmax attention mechanism. We provide theoretical guarantees for the convergence of our Gradient Descent-based optimization method to a near-optimal pruning mask solution. Our preliminary empirical results demonstrate the effectiveness of this approach in maintaining model performance while significantly reducing computational costs. This work establishes a new theoretical foundation for pruning algorithm design in LLMs, potentially paving the way for more efficient LLM inference on resource-constrained devices.

1 INTRODUCTION

Large Language Models (LLMs) based on the transformer architecture (Vaswani et al., 2017), including GPT-4o (OpenAI, 2024a), Claude (Anthropic, 2024), and OpenAI's recent o1 (OpenAI, 2024b), have shown immense potential to enhance our daily lives. They revolutionize fields like conversational AI (Liu et al., 2024), AI agents (Xi et al., 2023; Chen et al., 2024b), search AI (OpenAI, 2024b), and AI assistants (Mahmood et al., 2023; Zhang et al., 2023; Kuo et al., 2024; Feng et al., 2024). With their growing capabilities, LLMs are powerful tools shaping the future of technology. However, the current state-of-the-art LLM weights number is extremely large. For instance, the smallest version of Llama 3.1 (Llama Team, 2024) needs 8 billion parameters, which takes more than 16GB GPU memory with half float precision and requires significant inference time. Due to large memory and high computational cost, deploying such models on edge devices such as smartphones becomes challenging.

To reduce the LLM model size, many studies work on pruning the LLMs model weights to relax the device memory constraint and minimize response latency. The classical pruning problem in LLMs can be formulated as follows. Given a weight matrix $W \in \mathbb{R}^{d \times d}$ and some calibration data $X \in \mathbb{R}^{n \times d}$, where n is input token length and d is feature dimension, the goal is to find a matrix $\widetilde{W}$ under some sparse constraint such that $\|XW - X\widetilde{W}\|$ being small under some norm function. The above formulation has been widely used in many state-of-the-art pruning methods, such as SparseGPT (Frantar & Alistarh, 2023) and Wanda (Sun et al., 2024).

However, the current object functions only focus on the approximation of a linear function XW. Their optimal solutions do not have a good approximation to the attention matrix (see Figure 2 for details). Note that the attention mechanism is the kernel module of the transformer architecture. The high-level insight of their bad performance is that the $\mathsf{Softmax}$ function is very sensitive to the large positive values of the input due to its $\exp$ scaling effect while pruning mask based on linear approximation cannot capture this sensitivity. Thus, in this work, we directly compute the pruning mask on weights to approximate the attention matrix, which is a highly non-linear function, $\mathsf{Softmax}(XWX^\top) \in \mathbb{R}^{n \times n}$. To the best of our knowledge, this paper is the first work studying attention weight pruning to directly approximate the attention matrix. We provide a theoretical

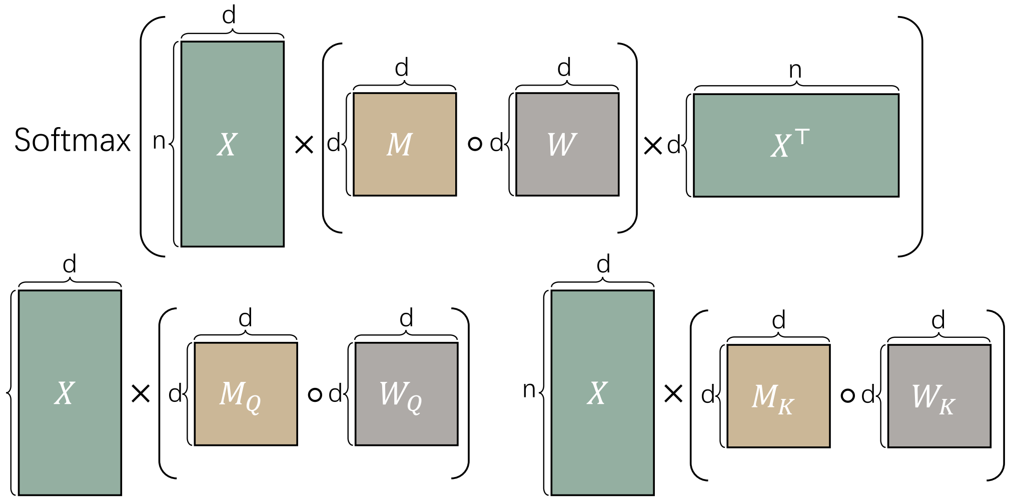


Figure 1: Comparison of our Attention Weights Pruning method and Linear Pruning method such as Wanda and SparseGPT. The top figure illustrates our proposed method of the attention matrix approximation, where pruning is applied directly to the fused attention weight matrix W, using only one pruning mask M. The bottom figure describes the Linear Pruning method of the linear function approximation, where pruning is applied separately to the query weight matrix W_Q and key weight matrix W_K, using two different pruning masks M_Q and M_K, respectively.

guarantee that optimization based on Gradient Descent (GD) on our loss function can converge to a good pruning mask solution (Theorem 1.3). Furthermore, we preliminarily verified the effectiveness of our method with empirical support (Section 6). Our theoretical foundation may pave the way for more efficient LLM inference on resource-constrained devices.

In the following, we introduce some key backgrounds and our contributions in detail.

1.1 KEY BACKGROUND

We define the attention matrix in self-attention mechanism as below:

Definition 1.1 (Attention Matrix)**.** *Let $X \in \mathbb{R}^{n \times d}$ be the input. Given query and key weights matrix $W_Q, W_K \in \mathbb{R}^{d \times d}$, we define $W := W_Q W_K^\top$. Then, we have the Softmax attention matrix being*

$$\mathsf{Softmax}(XWX^\top) = D^{-1}\exp(XWX^\top),$$

where (1) $D := \mathrm{diag}(\exp(XWX^\top) \cdot \mathbf{1}_n)$, (2) $\exp$ denotes the exponential function and is applied entry-wisely, (3) $\mathrm{diag}()$ operation takes a vector and outputs a diagonal matrix with the entries of that vector, and (4) $\mathbf{1}_n$ denotes the length-n all ones vector.

Further, we introduce the problem setup of our Attention Weights Pruning. By selectively reducing the number of non-zero elements in the attention weight matrix W in Definition 1.1, we can preserve model performance while lowering computational cost and GPU memory usage. Below, we formally define the Attention Weights Pruning problem and the corresponding loss function:

Definition 1.2 (Attention Weights Pruning)**.** *Let $M \in [0,1]^{d \times d}$ be the pruning mask. Let X, W be defined in Definition 1.1. Let $A := \exp(XWX^\top)$ and $\widetilde{A} := \exp(X(M \circ W)X^\top)$, where $\circ$ is the Hadamard product. Let $D := \mathrm{diag}(A \cdot \mathbf{1}_n)$ and $\widetilde{D} := \mathrm{diag}(\widetilde{A} \cdot \mathbf{1}_n)$. Let $\lambda \in \mathbb{R}_+$ be the regularization parameter. We define Attention Weights Pruning loss function to be*

$$\mathcal{L}(M) := \frac{1}{2}\|D^{-1}A - \widetilde{D}^{-1}\widetilde{A}\|_F^2 + \frac{1}{2}\lambda\|M\|_F^2.$$

Thus, the Attention Weights Pruning optimization problem is $\min_M \ \mathcal{L}(M)$.

1.2 OUR CONTRIBUTIONS

This is the first work studying the Attention Weights Pruning problem, which is an approximation problem to a non-linear function. We provide an algorithm to obtain the near-optimal pruning mask based on Gradient Descent (GD) with convergence guarantee.

Theorem 1.3 (Main result, informal version of Theorem 4.1)**.** *For any $\epsilon > 0$, our Algorithm 1 can converge to the near-optimal pruning mask for the Attention Weights Pruning problem (Definition 1.2) in $O(d \operatorname{poly}(n)/\epsilon)$ time with $O(\xi + \epsilon)$ error, where ξ is a small term depending on intrinsic property of the data and weights.*

In the above theorem, ξ can be arbitrarily small as $\xi \to 0$ when the regularization coefficient $\lambda \to 0$. So our analysis shows that although the objective function is highly non-linear, the GD training can converge to a near-optimal pruning mask solution, supported by our experiments in Section 6.

Our contributions are as follows:

- This is the first work that analyzes the weights pruning problem based on $\mathsf{Softmax}$ attention, which is a non-linear function.
- We provide the closed form of the gradient of Attention Weights Pruning loss function (Theorem 5.3), and Lipschitz of that gradient (Theorem 5.4),
- We provide Gradient Descent based Algorithm 1 to obtain the near-optimal pruning mask and its convergence guarantee (Theorem 4.1).
- We conduct preliminary experiments to verify the effectiveness of our method (Section 6).

Roadmap. Our paper is organized as follows. In Section 2, we review the related work. Section 3 introduces key concepts and definitions essential for the subsequent sections. In Section 4, we present our main result. Section 5 offers a technical overview of the methods employed. Experimental results are discussed in Section 6. Finally, Section 8 summarizes our findings and offers concluding remarks.

2 RELATED WORK

2.1 PRUNING AND COMPRESSION FOR LLMS

Model compression plays a critical role in improving the efficiency and deployment of large language models (LLMs) (Zhu et al., 2023) for its effectiveness in reducing computational overhead while preserving performance. Common compression techniques include quantization (Park et al., 2024; Xiao et al., 2023; Hooper et al., 2024), pruning (Chen et al., 2021; Hoefler et al., 2021; Hubara et al., 2021; Jin et al., 2022; Frantar & Alistarh, 2022; 2023; Sun et al., 2024; Li et al., 2024a; Zandieh et al., 2024; Zhang et al., 2024b; Xia et al., 2023; Ashkboos et al., 2024; Chen et al., 2024a), and knowledge distillation (Hsieh et al., 2023; Shridhar et al., 2023; Jiang et al., 2023; Wang et al., 2023). Specifically, pruning techniques have been developed extensively, such as unstructured pruning, which removes individual weights (Li et al., 2024a; Sun et al., 2024), and structured pruning, which eliminates entire components like neurons or attention heads (Michel et al., 2019; Ashkboos et al., 2024; Xia et al., 2024). Sun et al. (2024) proposed Wanda, a novel unstructured pruning technique that uses weight-activation products to induce up to 50% sparsity in LLMs without retraining, achieving competitive results with significantly lower computational cost. SparseGPT (Frantar & Alistarh, 2023) introduced a one-shot pruning method that achieves up to 60% sparsity in large GPT-family models with minimal impact on performance. A follow-up work (Li et al., 2024a) improved the complexity analysis of SparseGPT, reducing the running time from $O(d^3)$ to $O(d^{2.53})$, enabling faster pruning on LLMs. These techniques together contribute to more scalable and resource-efficient LLMs, maintaining competitive performance while having substantial reductions in computational resources.

2.2 ATTENTION ACCELERATION

Attention mechanism has faced criticism due to its quadratic time complexity with respect to context length (Vaswani et al., 2017). Addressing this criticism, a variety of approaches are employed, including sparse attention (Hubara et al., 2021; Kurtic et al., 2023; Frantar & Alistarh, 2023; Li et al., 2024a), low-rank approximations (Razenshteyn et al., 2016; Li et al., 2016; Hu et al., 2022; Zeng & Lee, 2024; Hu et al., 2024d), and kernel-based methods (Charikar et al., 2020; Liu & Zenke, 2020; Deng et al., 2023a; Zandieh et al., 2023; Liang et al., 2024a), to reduce computational overhead and improve scalability. Aggarwal & Alman (2022) enable the derivation of a low-rank representation of the attention matrix, which accelerates both the training and inference processes of single attention layer, tensor attention, and multi-layer transformer, achieving almost linear time complexity (Alman & Song, 2023; 2024a;b; Liang et al., 2024e;b). Other approaches like Mamba (Gu & Dao, 2023; Dao & Gu, 2024), Linearizing Transformers (Zhang et al., 2024a; Mercat et al., 2024), Hopfield Models (Hu et al., 2023; Wu et al., 2024b; Hu et al., 2024c; Xu et al., 2024a; Wu et al., 2024a; Hu et al., 2024a;b;e), and PolySketchFormer (Kacham et al., 2023) focus on architectural modifications and implementation optimizations to enhance performance. System-level optimizations such as FlashAttention (Dao et al., 2022; Dao, 2023; Shah et al., 2024) and block-wise parallel decoding (Stern et al., 2018) further improve efficiency. Collectively, these innovations have significantly augmented transformer models' ability to handle longer input sequences, unlocking broader applications across multiple sectors (Chen et al., 2023; Peng et al., 2023; Ding et al., 2024; Ma et al., 2024; Xu et al., 2024b; An et al., 2024; Jin et al., 2024; Li et al., 2024b; Liang et al., 2024c; Shi et al., 2024a).

3 PRELIMINARY

In this section, we introduce some basic concepts and key definitions. In Section 3.1, we introduce some basic notations we use in this paper. In Section 3.2, we provide the definition of attention weights pruning.

3.1 NOTATIONS

For any positive integer n, we use $[n]$ to denote set $\{1, 2, \cdots, n\}$. For each $a, b \in \mathbb{R}^n$, we use $a \circ b \in \mathbb{R}^n$ to denote the Hadamard product, i.e., the i-th entry of $(a \circ b)$ is $a_i b_i$ for all $i \in [n]$. For $A \in \mathbb{R}^{m \times n}$, let $A_i \in \mathbb{R}^n$ denote the i-th row and $A_{*,j} \in \mathbb{R}^m$ denote the j-th column of A, where $i \in [m]$ and $j \in [n]$. We use $\exp(A)$ to denote a matrix where $\exp(A)_{i,j} := \exp(A_{i,j})$ for a matrix $A \in \mathbb{R}^{n \times d}$. We use $\|A\|_F$ to denote the Frobenius norm of a matrix $A \in \mathbb{R}^{n \times d}$, i.e., $\|A\|_F := \sqrt{\sum_{i \in [n]} \sum_{j \in [d]} |A_{i,j}|^2}$. For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, $A \succeq 0$ means that A is positive semidefinite (PSD), i.e., for all $x \in \mathbb{R}^n$, we have $x^\top A x \geq 0$.

3.2 ATTENTION WEIGHTS PRUNING

We aim to determine a near-optimal pruning mask M for the attention weights in a self-attention mechanism. Furthermore, we incorporate causal attention masking[1] into our method to be more aligned with the current decoder-only LLM architecture, while our analysis can be applied to any general attention mask, e.g., block-wise attention mask. To formalize this, we provide the formal definition of causal attention mask and attention weights pruning in this section.

The causal attention mask ensures that each token in the sequence can attend only to itself and preceding tokens. Here, we provide the formal definition for the causal attention mask:

Definition 3.1 (Causal attention mask, Liang et al. (2024b)). *We define the causal attention mask as* $M_c \in \{0,1\}^{n \times n}$, *where* $(M_c)_{i,j} = 1$ *if* $i \geq j$ *and* $(M_c)_{i,j} = 0$ *otherwise.*

Now, we incorporate Attention Weights Pruning (see Definition 1.2) with causal attention mask M_c.

Definition 3.2 (Attention Weights Pruning with Causal Attention Mask). *Let* $M_c \in \{0,1\}^{n \times n}$ *be the causal attention mask defined in Definition 3.1. Let* $A := \exp(XWX^\top) \circ M_c$ *and* $\widetilde{A} := \exp(X(M \circ W)X^\top) \circ M_c$. *Let* $D := \mathrm{diag}(A \cdot \mathbf{1}_n)$ *and* $\widetilde{D} := \mathrm{diag}(\widetilde{A} \cdot \mathbf{1}_n)$. *We define Attention Weights Pruning with Causal Attention Mask loss function to be*

$$\mathcal{L}(M) := \mathcal{L}_{\mathrm{attn}}(M) + \mathcal{L}_{\mathrm{reg}}(M)$$

---

[1] In this paper, we always use *pruning mask* to refer $M \in \mathbb{R}^{d \times d}$, which is our target, and use *causal attention mask* to refer $M_c \in \mathbb{R}^{n \times n}$, which is a fixed mask for standard self-attention.

where

$$\mathcal{L}_{\rm attn}(M) := \frac{1}{2}\|D^{-1}A - \widetilde{D}^{-1}\widetilde{A}\|_F^2 \text{ and } \mathcal{L}_{\rm reg}(M) := \frac{1}{2}\lambda\|M\|_F^2.$$

Algorithm 1 Gradient Descent for Pruning Mask (Theorem 4.1). Let $X_1, X_2, \dots, X_k \in \mathbb{R}^{n \times d}$ be our calibration dataset of size k. We iteratively run our GD method on this dataset.

```
procedure PruneMaskGD( $X_1, X_2, \dots, X_k \in \mathbb{R}^{n\times d}$, $W \in \mathbb{R}^{d\times d}$, $M_c \in \{0,1\}^{d\times d}$, $\rho \in [0,1], \lambda \in [0,1]$, $\epsilon \in (0, 0.1)$)    ▷ Theorem 4.1
    Initialize $M \in \{1\}^{d\times d}$
    Initialize $\eta, T$ by Theorem 4.1
    for $j = 1 \to k$ do    ▷ Iterate over the dataset
        $u_j \leftarrow \exp(X_j W X_j^\top)$
        $f_j \leftarrow \mathrm{diag}((u_j \circ M_c) \cdot \mathbf{1}_n)^{-1}(u_j \circ M_c)$
    end for
    for $i = 1 \to T$ do    ▷ Iterate over $T$ steps
        for $j = 1 \to k$ do    ▷ Iterate over the dataset
            $\widetilde{u}_j \leftarrow \exp(X_j(M \circ W)X_j^\top)$
            $\widetilde{f}_j \leftarrow \mathrm{diag}((\widetilde{u}_j \circ M_c) \cdot \mathbf{1}_n)^{-1}(\widetilde{u}_j \circ M_c)$
            $c_j \leftarrow \widetilde{f}_j - f_j$
            $p_{1,j} \leftarrow c_j \circ \widetilde{f}_j$
            $p_{2,j} \leftarrow \mathrm{diag}(p_{1,j} \cdot \mathbf{1}_n)\widetilde{f}_j$
            $p_j \leftarrow p_{1,j} - p_{2,j}$
        end for
        $M \leftarrow M - (\eta/k) \cdot (W \circ (\sum_{j=1}^k X_j^\top p_j X_j) + \lambda M)$    ▷ Gradient Descent
    end for
    $m \leftarrow \mathrm{vec}(M)$    ▷ Flatten $M$ into a vector
    $m_{\rm sorted} \leftarrow \mathrm{sort}(m)$    ▷ Sort the elements of $M$
    $\tau \leftarrow m_{\rm sorted}[\lfloor \rho \cdot d^2 \rfloor]$    ▷ Get the $\rho$-th largest element
    $M_{ij} \leftarrow \begin{cases} 1 & \text{if } M_{ij} > \tau \\ 0 & \text{otherwise} \end{cases}$    ▷ Set the top $\rho$ entries to 1, others to 0
    return $M$
end procedure
```

4 MAIN RESULTS

In this section, we provide our main results. We provide an Algorithm 1 for Attention Weights Pruning problem based on Gradient Descent (GD). We also prove the convergence for our GD algorithm in Theorem 4.1.

Theorem 4.1 (Main result, formal version of Theorem 1.3)**.** *Let* $M,\ X,\ W, \widetilde{D}, \widetilde{A}\ \lambda,\ \mathcal{L},\ \mathcal{L}_{\rm attn}$ *be defined in Definition 3.2. Assume* $XX^\top \succeq \beta I$ *and* $\min_{i,j\in[n]}(\widetilde{D}^{-1}\widetilde{A})_{i,j} \geq \delta > 0$*. Furthermore,*

- *Let* $\mu = 2\min_{i,j\in[d]}\{|W_{i,j}|\} \cdot \beta \cdot \delta$*.*
- *Let* $\xi = 12n^{-1.5}\max_{i,j\in[d]}\{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot \lambda d/\mu$*.*

Then, for any $\epsilon > 0$*, provided* $\eta < 1/L$ *where* L *is the Lipschitz constant for* $\nabla_M \mathcal{L}(M)$ *(see Theorem 5.4), GD (Algorithm 1) with fixed step size* η *and run for* $T = 4\mathcal{L}(M^{(0)})/(\eta\mu\epsilon n^2)$ *iterations results in the following guarantee,*

$$\frac{1}{n^2}\min_{t<T}\mathcal{L}_{\rm attn}(M^{(t)}) + \frac{\lambda^2}{\mu n^2}\|M^{(t)}\|_F^2 \leq (\xi + \epsilon)/2.$$

Proof. Let $g(M) = 2\mathcal{L}_{\rm attn}(M) + \frac{2\lambda^2}{\mu}\|M\|_F^2$. Note that $\mathcal{L}(M)$ satisfies the $(g(M), n^2\xi, 2, \mu)$-proxy PL inequality (Lemma 5.5). Also, we have $\mathcal{L}(M)$ is non-negative and has L-Lipschitz gradients Theorem 5.4. Thus, we finish the proof using Theorem 5.2. $\square$

Remark 4.2. *The two assumptions in Theorem 4.1 are practical. The first assumption of the positive definite matrix is widely used in theoretical deep learning analysis, e.g., Li & Liang (2018); Du et al. (2019); Allen-Zhu et al. (2019b); Arora et al. (2019). The second assumption is natural, as* $\widetilde{D}^{-1}\widetilde{A}$ *is the pruned attention matrix, where each entry is*

$$\frac{\exp(X(M \circ W)X^\top)_{i,j}}{\sum_{j=1}^n \exp(X(M \circ W)X^\top)_{i,j}} > 0,$$

which has a natural lower bound.

Our error upper bound in Theorem 4.1 is $O(\xi + \epsilon)$, where ϵ can be arbitrarily small. For ξ, we can let it be small by choosing a proper λ, i.e., the ξ error term can be made arbitrarily small by choosing small λ. However, if we choose a very small λ, the algorithm's run time gets larger as $T \propto 1/\eta \propto L \propto (\lambda + \text{other terms})$. Thus, although the objective function is highly non-linear, we can show that the Gradient Descent of our Algorithm 1 can converge to a good solution of the Attention Weights Pruning problem.

After solving the optimization problem, we obtain a pruning mask with real-valued entries. In practice, however, this pruning mask must be converted into a binary form, specifically $M \in \{0,1\}^{d \times d}$. We define the pruning ratio ρ as the percentage of weights to be pruned. We apply this ratio by setting the pruning mask entries to zero for weights that fall below the ρ-th percentile and to one for those above. This ensures that only the specified proportion of weights are pruned.

5 TECHNIQUE OVERVIEW

In Section 5.1, we introduce some useful tools from previous work. In Section 5.2, we derive the close form of the gradient of Attention Weights Pruning. In Section 5.3, we calculate the Lipschitz constant of that gradient. In Section 5.4, we prove the PL inequality for our loss function.

5.1 PREVIOUS TOOLS ON CONVERGENCE OF GD

To analyze the convergence behavior of GD for our optimization problem (Definition 3.2), we first introduce the concept of g-proxy, ξ-optimal Polyak–Łojasiewicz(PL) inequality (Polyak, 1963; Lojasiewicz, 1963; Karimi et al., 2016), under which GD will converge:

Definition 5.1 (g-proxy, ξ-optimal PL inequality, Definition 1.2 in Frei & Gu (2021)). *We say that a function* $f : \mathbb{R}^p \to \mathbb{R}$ *satisfies a g-proxy, ξ-optimal Polyak–Łojasiewicz inequality with parameters* $\alpha > 0$ *and* $\mu > 0$ *(in short, fsatisfies the* (g, ξ, α, μ)*-PL inequality) if there exists a function* $g : \mathbb{R}^p \to \mathbb{R}$ *and scalars* $\xi \in \mathbb{R}$, $\mu > 0$ *such that for all* $w \in \mathbb{R}^p$,

$$\|\nabla f(w)\|^\alpha \geq \frac{1}{2}\mu(g(w) - \xi).$$

PL inequality is a powerful tool for studying non-convex optimization, and it has been used in recent studies on provable guarantees for neural networks trained by gradient descent (Li & Liang, 2018; Allen-Zhu et al., 2019a;b;c; Frei et al., 2019; Cao & Gu, 2020; Ji & Telgarsky, 2019; Frei et al., 2021; Shi et al., 2021; 2024b). It provides a proxy convexity property, although the objective function is non-convex. In detail, for a function with good smoothness property, we can find some proxy functions and show the convergence by utilizing these proxy functions.

Leveraging this PL inequality, Frei & Gu (2021) derives the following GD convergence guarantees.

Theorem 5.2 (Theorem 3.1 in Frei & Gu (2021)). *Suppose* $f(w)$ *satisfies the* $(g(\cdot), \xi, \alpha, \mu)$*-proxy PL inequality for some function* $g(\cdot) : \mathbb{R}^p \to \mathbb{R}$. *Assume that* f *is non-negative and has L-Lipschitz gradients. Then for any* $\epsilon > 0$, *provided* $\eta < 1/L$, *GD with fixed step size* η *and run for* $T = 2\eta^{-1}(\mu\epsilon/2)^{-2/\alpha} f(w^{(0)})$ *iterations results in the following guarantee,*

$$\min_{t<T} g(w^{(t)}) \leq \xi + \epsilon.$$

The above theorem establishes that under the (g, ξ, α, μ)-PL inequality and Lipschitz continuity of the gradient, GD converges to a point where the proxy function $g(w)$ is within ϵ of ξ. To apply this result to our specific problem, we need to verify these conditions for our loss function $\mathcal{L}(M)$.

5.2 CLOSED FORM OF GRADIENT

As a first step, we compute the close form of the gradient $\nabla_M \mathcal{L}(M)$. The pruning mask M is inside a non-linear function Softmax, which complicates our calculation. We defer the proof to Section B.

Theorem 5.3 (Closed form of gradient, informal version of Theorem C.5)**.** *Let $\mathcal{L}(M)$ be defined in Definition 3.2. Let p be defined in Definition C.1. Let $X \in \mathbb{R}^{n \times d}$, $M \in [0,1]^{d \times d}$, $W \in \mathbb{R}^{d \times d}$. Then, we have*

$$\frac{\mathrm{d}\mathcal{L}(M)}{\mathrm{d}M} = W \circ (X^\top p X) + \lambda M.$$

Based on Theorem 5.3, we calculate the gradient of the pruning mask from Line 10 to Line 15 in our Algorithm 1.

5.3 LIPSCHITZ OF GRADIENT

Having obtained the close form of gradient, we proceed to investigate its Lipschitz continuity. We aim to show that the gradient $\nabla_M \mathcal{L}(M)$ is Lipschitz continuous with respect to M.

Theorem 5.4 (Lipschitz of the gradient, informal version of Theorem E.8)**.** *Let R be some fixed constant satisfies $R > 1$. Let $X \in \mathbb{R}^{n \times d}, W \in \mathbb{R}^{d \times d}$. We have $\|X\|_F \leq R$ and $\|W\|_F \leq R$. Let $\mathcal{L}(M)$ be defined in Definition 3.2. For $M, \widetilde{M} \in \mathbb{R}^{d \times d}$, we have*

$$\|\nabla_M \mathcal{L}(M) - \nabla_M \mathcal{L}(\widetilde{M})\|_F \leq (\lambda + 30 d n^{7/2} R^6) \cdot \|M - \widetilde{M}\|_F.$$

We defer the proof to Section E. Establishing the Lipschitz continuity of the gradient satisfies one of the necessary conditions for applying Theorem 5.2. The above theorem implicates that the gradient for M is upper bounded, providing a way to choose step size.

5.4 PL INEQUALITY OF GRADIENT

Next, we need to verify that our loss function satisfies the PL inequality with appropriate parameters. To complete the verification of the conditions required for convergence, we demonstrate that $\mathcal{L}(M)$ satisfies the PL inequality. We show that $\nabla_M \mathcal{L}(M)$ satisfies the PL inequality in this lemma:

Lemma 5.5 (PL inequality, informal version of Lemma F.10)**.** *Let $M, X, W, \widetilde{D}, \widetilde{A}$, λ, $\mathcal{L}$, $\mathcal{L}_{\mathrm{attn}}$ be defined in Definition 3.2. Assume that $XX^\top \succeq \beta I$ and $\min_{i,j \in [n]} (\widetilde{D}^{-1}\widetilde{A})_{i,j} \geq \delta > 0$. Also,*

- *Let $\mu = 2 \min_{i,j \in [d]} \{|W_{i,j}|\} \cdot \beta \cdot \delta$.*
- *Let $\xi = 12\sqrt{n} \max_{i,j \in [d]} \{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot \lambda d / \mu$.*

We have

$$\|\nabla_M \mathcal{L}(M)\|_F^2 \geq \frac{1}{2}\mu(2\mathcal{L}_{\mathrm{attn}}(M) + \frac{2\lambda^2}{\mu}\|M\|_F^2 - \xi).$$

We defer the proof to Section F. By confirming that $\mathcal{L}(M)$ satisfies the PL inequality and that its gradient is Lipschitz continuous, we then apply Theorem 5.2 to conclude that GD will converge to a solution within our desired error tolerance, and further prove Theorem 4.1.

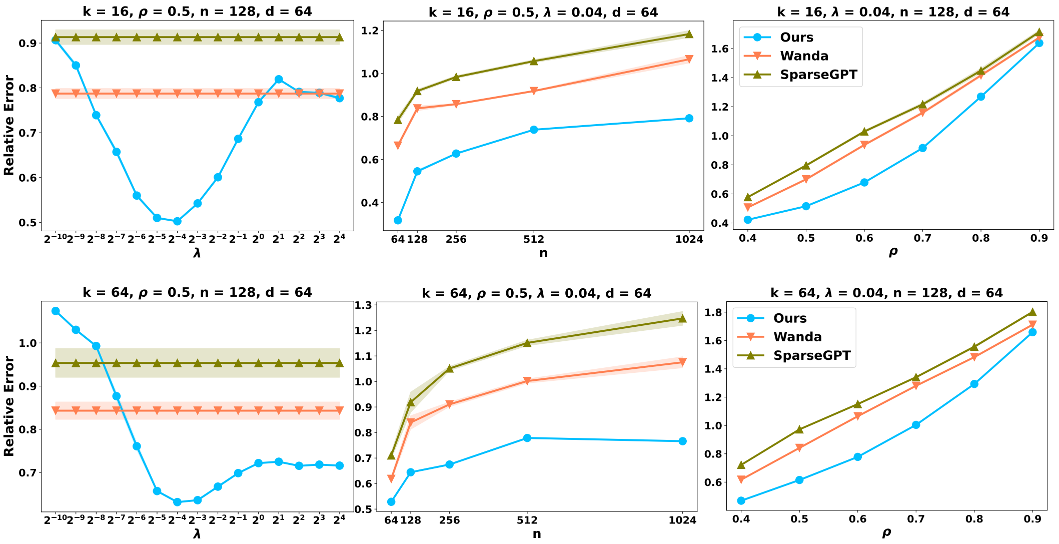
To prove the PL inequality, we also need the following two key Lemmas, which introduce our two assumptions in our Theorem 4.1, $XX^\top \succeq \beta I$ and $\min_{i,j \in [n]} (\widetilde{D}^{-1}\widetilde{A})_{i,j} \geq \delta > 0$.

Lemma 5.6 (Informal version of Lemma F.4)**.** *Let $B \in \mathbb{R}^{n \times n}$ and $X \in \mathbb{R}^{n \times d}$. Assume that $XX^\top \succeq \beta I$. Then, we have*

$$\|X^\top B X\|_F \geq \beta \|B\|_F.$$

Lemma 5.7 (Informal version of Lemma F.7)**.** *Let $B \in \mathbb{R}^{n \times n}$ and each row summation is zero, i.e., $B \cdot \mathbf{1}_n = \mathbf{0}_n$. Let $\widetilde{B} \in [0,1]^{n \times n}$ and each row summation is 1, i.e., $\widetilde{B} \cdot \mathbf{1}_n = \mathbf{1}_n$. Assume that $\min_{i,j \in [n]} \widetilde{B}_{i,j} \geq \delta > 0$. Then, we can show*

$$\|B \circ \widetilde{B} - \mathrm{diag}((B \circ \widetilde{B}) \cdot \mathbf{1}_n)\widetilde{B}\|_F \geq \delta \cdot \|B\|_F.$$



(a) Varying λ. (b) Varying n. (c) Varying ρ.

Figure 2: The comparison among our Algorithm 1, Wanda, and SparseGPT. The y-axis is a relative error, which is defined as $\frac{\|\widetilde{D}^{-1}\widetilde{A} - D^{-1}A\|_F^2}{\|D^{-1}A\|_F^2}$, where $D^{-1}A$ is original attention matrix and $\widetilde{D}^{-1}\widetilde{A}$ is approximated attention matrix based on three methods. We always use $d = 64$. We use $k = 16$ for the first row and $k = 64$ for the second row. The x-axis is (a) regularization coefficient λ for the left column; (b) input sequence length n for the middle column; (c) pruning ratio ρ for the right column.

6 EXPERIMENT

In this section, we discuss the experiments conducted to illustrate the effectiveness of our Algorithm 1. We first introduce our settings in Section 6.1. Then, we present our results in Section 6.2.

6.1 SETTINGS

Method and evaluation. We implement our method following the pseudocode in Algorithm 1, using NumPy and JAX for acceleration. We evaluate our method on unstructured sparsity, meaning that zeros can occur anywhere within the attention weight matrix W. Specifically, we use Definition 3.2 as our loss function, optimizing over the pruning mask M using gradient descent based on the closed-form expression derived in Theorem 5.3. To accelerate convergence, we leverage momentum into the optimization process and fix the momentum parameter at 0.9. After obtaining the optimal pruning mask, we convert M to a binary pruning mask to prune W, maintaining sparsity at the desired pruning ratio ρ. We use the relative error as our evaluation metric, which is defined as

$$\frac{\|\widetilde{D}^{-1}\widetilde{A} - D^{-1}A\|_F^2}{\|D^{-1}A\|_F^2}, \tag{1}$$

where $\widetilde{D}$, $\widetilde{A}$, D, A are defined in Definition 3.2.

Baselines. We compare our method with two linear pruning approaches, namely Wanda (Sun et al., 2024) and SparseGPT (Frantar & Alistarh, 2023). Wanda is a pruning method that removes weights with the smallest magnitudes multiplied by the corresponding input activations, achieving sparsity without requiring retraining or weight updates. SparseGPT is a second-order pruning method that utilizes the Hessian matrix to prune a portion of the weight matrix while simultaneously updating the remaining parameters. We implement Wanda and SparseGPT as described in their respective papers. Notably, since the settings of SparseGPT and Wanda are linear, we do not prune the fused weight matrix W directly; instead, we prune W_Q and W_K separately (see Figure 1).

Data. In order to assess the efficacy of different methods in approximating the attention matrix, we construct the data via a carefully defined generating process. Specifically, we create multiple independent random Gaussian matrices $G \in \mathbb{R}^{d \times d}$, where each entry of G drawn from a normal distribution i.e., $G_{i,j} \sim \mathcal{N}(0,1)$ for $i, j \in [d]$. Then, we perform singular value decomposition (SVD) on matrix G, i.e., $U, S, V^\top = \mathrm{SVD}(G)$. We retain the first four singular values in S and set others to zero, constraining the rank to four. Our W_Q and W_K are then constructed as $U \operatorname{diag}(S) V^\top$. The weight matrix W used in our setting is formed by taking the product $W = W_Q W_K^\top$. For $X \in \mathbb{R}^{n \times d}$, we generate it as a full-rank Gaussian random matrix.

Setup. In our experiments, the weight matrix dimension $d = 64$ is kept as constant across all figures, and we simulate two datasets of size $k = 16$ and $k = 64$. We set the input sequence length $n = 128$ for experiments (a) and (c) in Figure 2. The pruning ratio $\rho = 0.5$ is set for experiments (a) and (b) in Figure 2. For our method, the regularization coefficient $\lambda := \widetilde{\lambda}/n$ where we abuse the notation to denote $\widetilde{\lambda}$ as the same used in Definition 3.2 and λ here is the parameter we really control in experiments. λ is set as 0.04 for experiments (b) and (c) in Figure 2 (intuition drawn from experiment (a)). The total number of epochs is set as $T = 100$. The step size is set as $\eta = 0.1/\lambda$ because Theorem 4.1 indicates that η is inversely proportional to λ with some constant, i.e., $\eta \propto 1/L \propto 1/(\lambda + \text{other terms})$.

6.2 RESULTS

Overall, the results in Figure 2 show that our Algorithm 1 outperforms Wanda and SparseGPT with a large margin, which supports our theoretical analysis in Theorem 4.1. In the following, we will discuss each setting in detail.

Relation with regularization coefficient λ. The leftmost column of Figure 2 investigates the impact of the regularization coefficient λ on relative error. As λ increases from very small values, the relative error initially decreases sharply for our algorithm, reaching a minimum before gradually rising again, which forms a U shape curve. This behavior indicates that there is an optimal λ where our algorithm achieves its best performance around 2^{-4}. The U-shape curve phenomena are well-known in most hyper-parameter choosing, e.g., regularization coefficient.

Relation with input sequence length n. The center column of Figure 2 explores how the relative error changes with respect to the input sequence length n. As n increases, the relative error for all three methods grows, though at different rates. Our method demonstrates a slower increase, maintaining a significant margin over both Wanda and SparseGPT, particularly for larger values of n. Wanda, while showing better performance than SparseGPT for larger sequence lengths, becomes comparable to SparseGPT as n is relatively small.

Relation with pruning ratio ρ. The rightmost column of Figure 2 illustrates the relationship between the relative error and the pruning ratio ρ for the three methods under comparison: our algorithm, Wanda, and SparseGPT. As the pruning ratio ρ increases, all methods exhibit a rise in relative error, indicating a degradation in approximation accuracy. However, our algorithm consistently outperforms both Wanda and SparseGPT across the range of ρ, with a lower relative error. SparseGPT and Wanda follow a similar trend, closely tracking each other.

7 EXPERIMENT ON REAL DATASET AND LLMS

In this section, we discuss the new experiments conducted on the real dataset and LLMs to illustrate the effectiveness of our method.

7.1 SETTINGS

Method and evaluation. We implement our method using PyTorch. We evaluate our method on unstructured sparsity, meaning that zeros can occur anywhere within the attention weight matrices W_Q and W_K. Specifically, we use the loss function defined below:

$$\mathcal{L}(M_Q, M_K) := \frac{1}{2}\|D^{-1}A - \widetilde{D}^{-1}\widetilde{A}\|_F^2 + \frac{1}{2}(\|M_Q\|_F^2 + \|M_K\|_F^2) \quad (2)$$

where D and A are the original attention matrix, M_Q, M_K are pruning masks, and

$$\widetilde{A} := \exp(X(M_Q \circ W_Q)(M_K \circ W_K)^\top X^\top), \quad \widetilde{D} := \operatorname{diag}(\widetilde{A} \cdot \mathbf{1}_n).$$

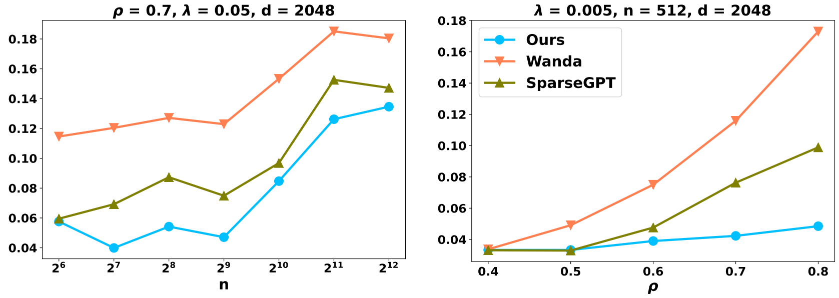


Figure 3: The comparison among our algorithm, Wanda, and SparseGPT on Llama 3.2-1B.

After obtaining the optimal pruning mask, we convert the pruning mask to a binary pruning mask to prune W_Q and W_K, maintaining sparsity at the desired pruning ratio ρ. We use the relative error as our evaluation metric, which is defined as Eq. (1).

Baselines. The pruning is performed on the last attention layer of the pretrained model Llama 3.2-1B Meta (2024). The baselines are Wanda and SparseGPT, the same as Section 6.1, and we reimplement them on PyTorch.

Data. To simulate real-world large language models, we utilize the Colossal Clean Crawled Corpus (C4 Dataset) Raffel et al. (2020), which is also used as the calibration dataset in our baselines Wanda and SparseGPT. Additionally, with a primary focus on pruning the attention matrix, we extract the input hidden states corresponding to the target attention matrix from the pretrained Llama 3.2 model using a customized hook function and use these as our input X.

Setup. The weight matrix dimension is $d = 2048$ in Llama 3.2-1B. For all the experiments, we set the regularization coefficient λ as 0.05 and the learning rate η as 0.005. For the varying n experiment, we set the pruning ratio ρ as 0.7. For the varying ρ experiment, we set the input sequence length n as 512.

7.2 RESULTS

Overall, the results in Figure 3 show that our algorithm outperforms Wanda and SparseGPT in real world LLMs, which supports our theoretical analysis in Theorem 4.1 and enhance our preliminary experiment in Section 6. In the following, we will discuss each setting in detail.

Relation with input sequence length n. The left column of Figure 3 shows that our algorithm outperforms the baselines in different sequence lengths continuously.

Relation with pruning ratio ρ. The right column of Figure 3 illustrates that as the pruning ratio ρ increases, all methods exhibit a rise in relative error. But our algorithm consistently outperforms both Wanda and SparseGPT across the range of ρ with a much lower increasing rate.

Assumptions verification. Notice that Theorem 4.1 relies on two assumptions: $XX^\top \succeq \beta I$ and $\min_{i,j \in [n]} (\widetilde{D}^{-1} \widetilde{A})_{i,j} \geq \delta > 0$. We verify these assumptions using the C4 dataset, obtaining $\beta \approx 0.034$ and $\delta \approx 0.0025$, thereby demonstrating the practicality of our theoretical framework.

8 CONCLUSION

This paper introduced a novel approach to LLM weight pruning that directly optimizes for approximating the attention matrix. We provided theoretical guarantees for the convergence of our Gradient Descent-based algorithm to a near-optimal pruning mask solution. Preliminary results demonstrated the method's effectiveness in maintaining model performance while reducing computational costs. This work establishes a new theoretical foundation for pruning algorithm design in LLMs, potentially enabling more efficient inference on resource-constrained devices.

REFERENCES

Amol Aggarwal and Josh Alman. Optimal-degree polynomial approximations for exponentials and gaussian kernel density estimation. In *Proceedings of the 37th Computational Complexity Conference*, pp. 1–23, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, 2019b.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *NeurIPS*, 2019c.

Josh Alman and Zhao Song. Fast attention requires bounded entries. *Advances in Neural Information Processing Systems*, 36, 2023.

Josh Alman and Zhao Song. The fine-grained complexity of gradient computation for training large language models. *arXiv preprint arXiv:2402.04497*, 2024a.

Josh Alman and Zhao Song. How to capture higher-order correlations? generalizing matrix softmax attention to kronecker computation. In *The Twelfth International Conference on Learning Representations*, 2024b.

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-free long-context scaling of large language models. *arXiv preprint arXiv:2402.17463*, 2024.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. `https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf`.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*, 2024.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3349–3356, 2020.

Moses Charikar, Michael Kapralov, Navid Nouri, and Paris Siminelakis. Kernel density estimation through density constrained near neighbor search. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 172–183. IEEE, 2020.

Bo Chen, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. Hsr-enhanced sparse attention acceleration. *arXiv preprint arXiv:2410.10165*, 2024a.

Tianyi Chen, Bo Ji, Tianyu Ding, Biyi Fang, Guanyi Wang, Zhihui Zhu, Luming Liang, Yixin Shi, Sheng Yi, and Xiao Tu. Only train once: A one-shot neural network training and pruning framework. *Advances in Neural Information Processing Systems*, 34:19637–19651, 2021.

Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. *arXiv preprint arXiv:2407.07061*, 2024b.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Yichuan Deng, Zhao Song, Zifan Wang, and Han Zhang. Streaming kernel pca algorithm with small space. *arXiv preprint arXiv:2303.04555*, 2023a.

Yichuan Deng, Zhao Song, Shenghao Xie, and Chiwun Yang. Unmasking transformers: A theoretical approach to data recovery via attention weights. *arXiv preprint arXiv:2310.12462*, 2023b.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *ICLR*. arXiv preprint arXiv:1810.02054, 2019.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. How far are we from agi. *arXiv preprint arXiv:2405.10313*, 2024.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pp. 10323–10337. PMLR, 2023.

Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34: 7937–7949, 2021.

Spencer Frei, Yuan Cao, and Quanquan Gu. Algorithm-dependent generalization bounds for overparameterized deep residual networks. *Advances in neural information processing systems*, 32, 2019.

Spencer Frei, Yuan Cao, and Quanquan Gu. Provable generalization of sgd-trained neural networks of any width in the presence of adversarial label noise. In *International Conference on Machine Learning*, pp. 3427–3438. PMLR, 2021.

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

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *arXiv preprint arXiv:2401.18079*, 2024.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

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

Jerry Yao-Chieh Hu, Donglin Yang, Dennis Wu, Chenwei Xu, Bo-Yu Chen, and Han Liu. On sparse modern hopfield model. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Jerry Yao-Chieh Hu, Pei-Hsuan Chang, Haozheng Luo, Hong-Yu Chen, Weijian Li, Wei-Po Wang, and Han Liu. Outlier-efficient hopfield layers for large transformer-based models. In *Forty-first International Conference on Machine Learning (ICML)*, 2024a.

Jerry Yao-Chieh Hu, Bo-Yu Chen, Dennis Wu, Feng Ruan, and Han Liu. Nonparametric modern hopfield models. *arXiv preprint arXiv:2404.03900*, 2024b.

Jerry Yao-Chieh Hu, Thomas Lin, Zhao Song, and Han Liu. On computational limits of modern hopfield models: A fine-grained complexity analysis. In *Forty-first International Conference on Machine Learning (ICML)*, 2024c.

Jerry Yao-Chieh Hu, Maojiang Su, En-Jui Kuo, Zhao Song, and Han Liu. Computational limits of low-rank adaptation (lora) for transformer-based models. *arXiv preprint arXiv:2406.03136*, 2024d.

Jerry Yao-Chieh Hu, Dennis Wu, and Han Liu. Provably optimal memory capacity for modern hopfield models: Tight analysis for transformer-compatible dense associative memories. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024e.

Itay Hubara, Brian Chmiel, Moshe Island, Ron Banner, Joseph Naor, and Daniel Soudry. Accelerated sparse neural training: A provable and efficient method to find n: m transposable masks. *Advances in neural information processing systems*, 34:21099–21111, 2021.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. Lion: Adversarial distillation of proprietary large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024.

Tian Jin, Michael Carbin, Dan Roy, Jonathan Frankle, and Gintare Karolina Dziugaite. Pruning's effect on generalization through the lens of training and regularization. *Advances in Neural Information Processing Systems*, 35:37947–37961, 2022.

Praneeth Kacham, Vahab Mirrokni, and Peilin Zhong. Polysketchformer: Fast transformers via sketches for polynomial kernels. *arXiv preprint arXiv:2310.01655*, 2023.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016.

Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. Wikibench: Community-driven data curation for ai evaluation on wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024.

Eldar Kurtic, Denis Kuznedelev, Elias Frantar, Michael Goin, and Dan Alistarh. Sparse finetuning for inference acceleration of large language models. *arXiv preprint arXiv:2310.06927*, 2023.

Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. A tighter complexity analysis of sparsegpt. *arXiv preprint arXiv:2408.12151*, 2024a.

Xiaoyu Li, Yingyu Liang, Zhenmei Shi, Zhao Song, and Yufa Zhou. Fine-grained attention i/o complexity: Comprehensive analysis for backward passes. *arXiv preprint arXiv:2410.09397*, 2024b.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *International Conference on Machine Learning*, pp. 2358–2367. PMLR, 2016.

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. Differential privacy mechanisms in neural tangent kernel regression. *arXiv preprint arXiv:2407.13621*, 2024a.

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Yufa Zhou. Multi-layer transformers gradient can be approximated in almost linear time. *arXiv preprint arXiv:2408.13233*, 2024b.

Yingyu Liang, Zhenmei Shi, Zhao Song, and Chiwun Yang. Toward infinite-long prefix in transformer. *arXiv preprint arXiv:2406.14036*, 2024c.

Yingyu Liang, Zhenmei Shi, Zhao Song, and Yufa Zhou. Unraveling the smoothness properties of diffusion models: A gaussian mixture perspective. *arXiv preprint arXiv:2405.16418*, 2024d.

Yingyu Liang, Zhenmei Shi, Zhao Song, and Yufa Zhou. Tensor attention training: Provably efficient learning of higher-order transformers. *arXiv preprint arXiv:2405.16411*, 2024e.

Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *arXiv preprint arXiv:2401.02777*, 2024.

Tianlin Liu and Friedemann Zenke. Finding trainable sparse networks through neural tangent transfer. In *International Conference on Machine Learning*, pp. 6336–6347. PMLR, 2020.

AI @ Meta Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke Zettlemoyer, Omer Levy, and Chunting Zhou. Megalodon: Efficient llm pretraining and inference with unlimited context length. *arXiv preprint arXiv:2404.08801*, 2024.

Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. Llm-powered conversational voice assistants: Interaction patterns, opportunities, challenges, and design guidelines. *arXiv preprint arXiv:2309.13879*, 2023.

Jean Mercat, Igor Vasiljevic, Sedrick Keh, Kushal Arora, Achal Dave, Adrien Gaidon, and Thomas Kollar. Linearizing large language models. *arXiv preprint arXiv:2405.06640*, 2024.

Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, 2024. Accessed: 2024-11-21.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024a. Accessed: May 14.

OpenAI. Introducing openai o1-preview. `https://openai.com/index/introducing-openai-o1-preview/`, 2024b. Accessed: September 12.

Gunho Park, Minsub Kim, Sungjae Lee, Jeonghoon Kim, Beomseok Kwon, Se Jung Kwon, Byeongwook Kim, Youngjoo Lee, Dongsoo Lee, et al. Lut-gemm: Quantized matrix multiplication based on luts for efficient inference in large-scale generative language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Ilya Razenshteyn, Zhao Song, and David P Woodruff. Weighted low rank approximations with provable guarantees. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 250–263, 2016.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *arXiv preprint arXiv:2407.08608*, 2024.

Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. In *International Conference on Learning Representations*, 2021.

Zhenmei Shi, Yifei Ming, Xuan-Phi Nguyen, Yingyu Liang, and Shafiq Joty. Discovering the gems in early layers: Accelerating long-context llms with 1000x input token reduction. *arXiv preprint arXiv:2409.17422*, 2024a.

Zhenmei Shi, Junyi Wei, and Yingyu Liang. Provable guarantees for neural networks via gradient feature learning. *Advances in Neural Information Processing Systems*, 36, 2024b.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Dennis Wu, Jerry Yao-Chieh Hu, Teng-Yun Hsiao, and Han Liu. Uniform memory retrieval with larger capacity for modern hopfield models. In *Forty-first International Conference on Machine Learning (ICML)*, 2024a.

Dennis Wu, Jerry Yao-Chieh Hu, Weijian Li, Bo-Yu Chen, and Han Liu. STanhop: Sparse tandem hopfield model for memory-enhanced time series prediction. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Haojun Xia, Zhen Zheng, Yuchao Li, Donglin Zhuang, Zhongzhu Zhou, Xiafei Qiu, Yong Li, Wei Lin, and Shuaiwen Leon Song. Flash-llm: Enabling cost-effective and highly-efficient large generative model inference with unstructured sparsity. *Proceedings of the VLDB Endowment*, 17(2): 211–224, 2023.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations*, 2024.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pp. 38087–38099. PMLR, 2023.

Chenwei Xu, Yu-Chao Huang, Jerry Yao-Chieh Hu, Weijian Li, Ammar Gilani, Hsi-Sheng Goan, and Han Liu. Bishop: Bi-directional cellular learning for tabular data with generalized sparse modern hopfield model. In *Forty-first International Conference on Machine Learning (ICML)*, 2024a.

Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, and Yingyu Liang. Towards few-shot adaptation of foundation models via multitask finetuning. *arXiv preprint arXiv:2402.15017*, 2024b.

Amir Zandieh, Insu Han, Majid Daliri, and Amin Karbasi. Kdeformer: Accelerating transformers via kernel density estimation. In *ICML*. arXiv preprint arXiv:2302.02451, 2023.

Amir Zandieh, Majid Daliri, and Insu Han. Qjl: 1-bit quantized jl transform for kv cache quantization with zero overhead. *arXiv preprint arXiv:2406.03482*, 2024.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.

Jieyu Zhang, Ranjay Krishna, Ahmed H Awadallah, and Chi Wang. Ecoassistant: Using llm assistant more affordably and accurately. *arXiv preprint arXiv:2310.03046*, 2023.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. *arXiv preprint arXiv:2402.04347*, 2024a.

Yuxin Zhang, Lirui Zhao, Mingbao Lin, Sun Yunyun, Yiwu Yao, Xingjia Han, Jared Tanner, Shiwei Liu, and Rongrong Ji. Dynamic sparse no training: Training-free fine-tuning for sparse llms. In *The Twelfth International Conference on Learning Representations*, 2024b.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*, 2023.

Appendix

CONTENTS

1 Introduction 1
- 1.1 Key Background 2
- 1.2 Our Contributions 3

2 Related Work 3
- 2.1 Pruning and Compression for LLMs 3
- 2.2 Attention Acceleration 3

3 Preliminary 4
- 3.1 Notations 4
- 3.2 Attention Weights Pruning 4

4 Main Results 5

5 Technique Overview 6
- 5.1 Previous Tools on Convergence of GD 6
- 5.2 Closed Form of Gradient 7
- 5.3 Lipschitz of Gradient 7
- 5.4 PL Inequality of Gradient 7

6 Experiment 8
- 6.1 Settings 8
- 6.2 Results 9

7 Experiment on Real Dataset and LLMs 9
- 7.1 Settings 9
- 7.2 Results 10

8 Conclusion 10

A Preliminary 19
- A.1 Notations 19
- A.2 Facts 19

B Gradient Calculation 22
- B.1 Definitions 22
- B.2 Gradient for Each Row of $X(M \circ W)X^\top$ 24
- B.3 Gradient for Each Row of $\widetilde{u}(M)$ 25
- B.4 Gradient for Each Entry of $\widetilde{\alpha}(M)$ 25

B.5 Gradient for Each Entry of $\widetilde{f}(M)$ 26
B.6 Gradient for Each Entry of c(M) 27
B.7 Gradient for $\mathcal{L}_{\rm attn}(M)$ 27
B.8 Gradient for $\mathcal{L}_{\rm reg}(M)$ 28
B.9 Gradient for $\mathcal{L}(M)$ 29

C Matrix Form **29**
C.1 Matrix Form of $B(M)$ 29
C.2 Matrix Form of $\frac{\mathrm{d}}{\mathrm{d}M}\mathcal{L}(M)$ 33

D Bounds for Basic Functions **33**
D.1 Basic Assumptions 33
D.2 Bounds for Basic Functions 34
D.3 Bounds for Gradient of $\widetilde{f}(M)$ 36

E Lipschitz of Gradient **36**
E.1 Useful Facts 36
E.2 Lipschitz of $\widetilde{f}(M)$ 37
E.3 Lipschitz of $c(M)$ 38
E.4 Lipschitz of $\widetilde{f}(M) \circ c(M)$ 38
E.5 Lipschitz of $\mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)$ 38
E.6 Lipschitz of $\mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)$ 39
E.7 Lipschitz of Gradient 40

F Convergence of Gradient Descent **41**
F.1 Helpful Statements 41
F.2 Lower Bound on Frobenius Norm 42
F.3 Sandwich Lower Bound on Frobenius Norm 42
F.4 Lower Bound on Hadamard Product Between Two Matrices 43
F.5 Final Bound 43
F.6 PL Inequality 44

Roadmap. The appendix is organized as follows. In Section A, we give the preliminary of our paper. In Section B, we provide detailed gradient analysis of loss function. In Section C, we provide details about how we integrate the gradient of loss function into matrix form. In Section D, we bound some basic functions to be used later. In Section E, we provide proof for the Lipschitz property of the gradient of the loss function. In Section F, we provide proof of convergence for GD.

A PRELIMINARY

In Section A.1, we introduce some notations we use in this paper. In Section A.2, we provide some basic facts.

A.1 NOTATIONS

For any positive integer n, we use $[n]$ to denote set $\{1, 2, \cdots, n\}$. For two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, we use $\langle x, y\rangle$ to denote the inner product between x, y, i.e., $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$. For each $a, b \in \mathbb{R}^n$, we use $a \circ b \in \mathbb{R}^n$ to denote the Hadamard product, i.e. the i-th entry of $(a \circ b)$ is $a_i b_i$ for all $i \in [n]$. We use e_i to denote a vector where only i-th coordinate is 1, and other entries are 0. We use $\mathbf{1}_n$ to denote a length-n vector where all the entries are ones. We use $\|x\|_p$ to denote the ℓ_p norm of a vector $x \in \mathbb{R}^n$, i.e. $\|x\|_1 := \sum_{i=1}^n |x_i|$, $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$, and $\|x\|_\infty := \max_{i\in[n]} |x_i|$. For $A \in \mathbb{R}^{m\times n}$, let $A_i \in \mathbb{R}^n$ denote the i-th row and $A_{*,j} \in \mathbb{R}^m$ denote the j-th column of A, where $i \in [m]$ and $j \in [n]$. For a square matrix A, we use $\mathrm{tr}[A]$ to denote the trace of A, i.e., $\mathrm{tr}[A] = \sum_{i=1}^n A_{i,i}$. For two matrices $X, Y \in \mathbb{R}^{m\times n}$, the standard inner product between matrices is defined by $\langle X, Y\rangle := \mathrm{tr}[X^\top Y]$. We use $\exp(A)$ to denote a matrix where $\exp(A)_{i,j} := \exp(A_{i,j})$ for a matrix $A \in \mathbb{R}^{n\times d}$. For $k > n$, for any matrix $A \in \mathbb{R}^{k\times n}$, we use $\|A\|$ to denote the spectral norm of A, i.e. $\|A\| := \sup_{x\in\mathbb{R}^n} \|Ax\|_2/\|x\|_2$. We use $\|A\|_\infty$ to denote the ℓ_∞ norm of a matrix $A \in \mathbb{R}^{n\times d}$, i.e. $\|A\|_\infty := \max_{i\in[n],j\in[d]} |A_{i,j}|$. We use $\|A\|_F$ to denote the Frobenius norm of a matrix $A \in \mathbb{R}^{n\times d}$, i.e. $\|A\|_F := \sqrt{\sum_{i\in[n]} \sum_{j\in[d]} |A_{i,j}|^2}$. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, we use $A \succeq 0$ (positive semidefinite (PSD)), if for all $x \in \mathbb{R}^n$, we have $x^\top A x \geq 0$. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and the maximum eigenvalue of the square matrix A, respectively. Let $A \in \mathbb{R}^{n\times d}$. We use $a := \mathrm{vec}(A)$ to denote a length nd vector. We stack rows of A into a column vector, i.e. $\mathrm{vec}(A) := [a_1^\top, a_2^\top, \dots, a_n^\top]^\top$ where $a_i^\top$ is the i-th row of A, or simply $\mathrm{vec}(A)_{j+(i-1)d} := A_{i,j}$ for any $i \in [n], j \in [d]$.

A.2 FACTS

Fact A.1 (Indexing)**.** *Suppose we have matrices $U \in \mathbb{R}^{n\times m}, V \in \mathbb{R}^{m\times d}$. We define*

$$\underbrace{X}_{n\times d} := \underbrace{U}_{n\times m}\underbrace{V}_{m\times d}.$$

Then, we have the following:

- *Indexing for one row: $X_i = V^\top U_i \in \mathbb{R}^d$, i.e. $X_i^\top = U_i^\top V$, for $i \in [n]$.*
- *Indexing for one column: $X_{*,j} = UV_{*,j} \in \mathbb{R}^n$ for $j \in [d]$.*

Fact A.2. *We have*

Part 1.*Suppose we have vectors $u \in \mathbb{R}^n, v \in \mathbb{R}^n$. For $i \in [n]$, we define*

$$x_i := u_i v_i.$$

Then we have the following:

- $\underbrace{x}_{n\times 1} = \underbrace{u \circ v}_{n\times 1} = \underbrace{\mathrm{diag}(u)}_{n\times n}\underbrace{v}_{n\times 1} = \underbrace{\mathrm{diag}(v)}_{n\times n}\underbrace{u}_{n\times 1}$

Part 2.*Suppose we have matrix $W \in \mathbb{R}^{n\times n}$, vector $u \in \mathbb{R}^n$. For $i \in [n]$, we define*

$$X_{*,j} = W_{*,j} u_j.$$

Then we have the following:

- $X = W \operatorname{diag}(u)$

Fact A.3 (Calculus). *We have*

Part 1. (Scalar calculus) *For any* $t \in \mathbb{R}$*, function* $f : \mathbb{R} \to \mathbb{R}$*, we have*

- $\frac{\mathrm{d} f^n(t)}{\mathrm{d} t} = n f^{n-1}(t) \frac{\mathrm{d} f(t)}{\mathrm{d} t}$.

Part 2. (Vector calculus) *For any* $x, y \in \mathbb{R}^n$*,* $t \in \mathbb{R}$*, we have*

- $\frac{\mathrm{d}(x \circ y)}{\mathrm{d} t} = \frac{\mathrm{d} x}{\mathrm{d} t} \circ y + \frac{\mathrm{d} y}{\mathrm{d} t} \circ x$. *(Product rule of vector Hadamard product)*
- $\frac{\mathrm{d}\langle x, y\rangle}{\mathrm{d} t} = \langle \frac{\mathrm{d} x}{\mathrm{d} t}, y \rangle + \langle x, \frac{\mathrm{d} y}{\mathrm{d} t} \rangle$. *(Product rule of inner product)*
- $\frac{\mathrm{d} x}{\mathrm{d} x_i} = e_i$.

Part 3. (Matrix calculus) *For any* $X, Y \in \mathbb{R}^{n \times m}$*,* $Z \in \mathbb{R}^{m \times d}$*,* $t \in \mathbb{R}$ *which is independent of* Z*, function* $f : \mathbb{R} \to \mathbb{R}^{n \times d}$*, functions* $f_1(t), f_2(t), \ldots, f_n(t) : \mathbb{R} \to \mathbb{R}^{n \times d}$*, we have*

- $\frac{\mathrm{d}(X \circ Y)}{\mathrm{d} t} = \frac{\mathrm{d} X}{\mathrm{d} t} \circ Y + \frac{\mathrm{d} Y}{\mathrm{d} t} \circ X$. *(Product rule of matrix Hadamard product)*
- $\frac{\mathrm{d} \exp(f(t))}{\mathrm{d} t} = \exp(f(t)) \circ \frac{\mathrm{d} f(t)}{\mathrm{d} t}$*, where* $\exp(\cdot)$ *is applied entry-wise.*
- $\frac{\mathrm{d}(X Z)}{\mathrm{d} t} = \frac{\mathrm{d} X}{\mathrm{d} t} Z$.
- $\frac{\mathrm{d}(Z X^\top)}{\mathrm{d} t} = Z \frac{\mathrm{d} X^\top}{\mathrm{d} t}$.
- $\frac{\mathrm{d}}{\mathrm{d} t} \sum_{i=1}^n f_i(t) = \sum_{i=1}^n \frac{\mathrm{d} f_i(t)}{\mathrm{d} t}$.
- $\underbrace{\frac{\mathrm{d} X}{\mathrm{d} X_{i,j}}}_{n \times m} = \underbrace{e_i}_{n \times 1} \underbrace{e_j^\top}_{1 \times m}$.

Fact A.4 (Basic algebra). *Let* $u \in \mathbb{R}^n$*,* $v \in \mathbb{R}^n$*,* $w \in \mathbb{R}^n$*,* $X \in \mathbb{R}^{n \times d}$*,* $Y \in \mathbb{R}^{n \times d}$*, and* $Z \in \mathbb{R}^{n \times n}$*. Then, we have*

- $\langle u, v \rangle = \langle v, u \rangle = u^\top v = v^\top u$
- $u \circ v = v \circ u = \operatorname{diag}(u) v = \operatorname{diag}(v) u$
- $\langle u, v \rangle = \langle u \circ v, \mathbf{1}_n \rangle$
- $\langle u \circ v, w \rangle = \langle u \circ w, v \rangle = \langle w \circ v, u \rangle$
- $u^\top (v \circ w) = u^\top \operatorname{diag}(v) w$
- $(X \circ Y)^\top = X^\top \circ Y^\top$
- $X \circ e_i e_j^\top = X_{i,j} e_i e_j^\top$
- $\operatorname{diag}(u) Z \operatorname{diag}(v) = (u v^\top) \circ Z$
- $X Y^\top = \sum_{i \in [d]} X_{*,i} Y_{*,i}^\top$
- $X_{i,j} Y_{i,j} = (X \circ Y)_{i,j}$
- $\sum_{j \in [n]} u \circ A_{*,j} = u \circ \sum_{j \in [n]} A_{*,j}$
- $\|X\|_F^2 = \operatorname{tr}[X X^\top]$
- $\operatorname{tr}[X Y^\top] = \operatorname{tr}[Y^\top X]$

- $\| \mathrm{diag}(u)\|_F = \|u\|_2$

Fact A.5 (Norm bounds)**.** *For $a \in \mathbb{R}$, $u \in \mathbb{R}^d$, $X, Y \in \mathbb{R}^{n \times d}$, $Z \in \mathbb{R}^{d \times m}$ we have*

- $\|aX\|_F = |a|\|X\|_F$ *(absolute homogeneity).*
- $\|X + Y\|_F \leq \|X\|_F + \|Y\|_F$ *(triangle inequality).*
- $|\langle X, Y\rangle| \leq \|X\|_F \cdot \|Y\|_F$ *(Cauchy–Schwarz inequality).*
- $\|X^\top\|_F = \|X\|_F$.
- $\|Xu\|_2 \leq \|X\| \cdot \|u\|_2$
- $\|X \circ Y\|_F \leq \|X\|_F \cdot \|Y\|_F$.
- *For any $i \in [n]$, $j \in [d]$, we have $|X_{i,j}| \leq \|X\|_F$.*
- $\|X\| \leq \|X\|_F \leq \sqrt{k}\|X\|$ *where k is the rank of X.*
- $\|Y \cdot Z\|_F \leq \|Y\|_F \cdot \|Z\|_F$.

Fact A.6. *For matrices $A, B \in \mathbb{R}^{m \times n}$, we have*

$$\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + 2\langle A, B\rangle.$$

Proof. We can show

$$\begin{aligned} \|A + B\|_F^2 = &~ \mathrm{tr}[(A + B)^\top (A + B)] \\ = &~ \mathrm{tr}[A^\top A + A^\top B + B^\top A + B^\top B] \\ = &~ \mathrm{tr}[A^\top A] + \mathrm{tr}[B^\top B] + 2\,\mathrm{tr}[A^\top B] \\ = &~ \|A\|_F^2 + \|B\|_F^2 + 2\,\mathrm{tr}[A^\top B] \\ = &~ \|A\|_F^2 + \|B\|_F^2 + 2\langle A, B\rangle \end{aligned}$$

where the first step follows from $\mathrm{tr}[A^\top A] = \|A\|_F^2$ for matrix $A \in \mathbb{R}^{m \times n}$, the second step follows from the basic algebra, the third follows from $\mathrm{tr}[X^\top Y] = \mathrm{tr}[XY^\top]$ for matrices $X, Y \in \mathbb{R}^{m \times n}$, the fourth step follows from $\mathrm{tr}[A^\top A] = \|A\|_F^2$ for matrix $A \in \mathbb{R}^{m \times n}$, and the last step follows from definition of inner product of matrices. ☐

Lemma A.7. *Let $M \in \mathbb{R}^{n \times n}$. Let $X \in \mathbb{R}^{n \times n}$ be independent of M. We have*

$$\frac{\mathrm{d}(M \circ X)}{\mathrm{d}M_{i,j}} = X_{i,j} e_i e_j^\top$$

Proof. We can show

$$\begin{aligned} \frac{\mathrm{d}(M \circ X)}{\mathrm{d}M_{i,j}} = &~ M \circ \frac{\mathrm{d}X}{\mathrm{d}M_{i,j}} + X \circ \frac{\mathrm{d}M}{\mathrm{d}M_{i,j}} \\ = &~ X \circ \frac{\mathrm{d}M}{\mathrm{d}M_{i,j}} \\ = &~ X \circ (e_i e_j^\top) \\ = &~ X_{i,j} e_i e_j^\top \end{aligned}$$

where the first step, the second step and the third step follow from Fact A.3, the fourth step follows from Fact A.4.

☐

B GRADIENT CALCULATION

B.1 DEFINITIONS

In this section, we introduce some definitions we used to compute $\frac{\mathrm{d}\mathcal{L}(M)}{\mathrm{d}M}$. First, we introduce the exponential function.

Definition B.1 (Exponential function u, $\widetilde{u}$)**.** *If the following condition hold*

- *Let* $X \in \mathbb{R}^{n \times d}$.
- *Let* $W \in \mathbb{R}^{d \times d}$.
- *Let* $M \in [0,1]^{d \times d}$.
- *Let* $i_0 \in [n]$.

We define $u \in \mathbb{R}^{n \times n}$ *as follows*

$$u := \exp(XWX^\top).$$

We define $\widetilde{u}(M) \in \mathbb{R}^{n \times n}$ *as follows*

$$\widetilde{u}(M) := \exp(X(M \circ W)X^\top).$$

We define i_0*-th row of* $\widetilde{u}(M)$ *as follows*

$$\widetilde{u}(M)_{i_0} := \exp(X(M \circ W)X^\top)_{i_0}.$$

Then, we introduce the sum function.

Definition B.2 (Sum function of softmax α, $\widetilde{\alpha}$)**.** *If the following condition hold*

- *Let* $M \in [0,1]^{d \times d}$.
- *Let* $M_c \in \{0,1\}^{n \times n}$ *be the causal attention mask defined in Definition 3.1.*
- *Let* u*,* $\widetilde{u}(M)$ *be defined as Definition B.1.*
- *Let* $i_0 \in [n]$.

We define $\alpha \in \mathbb{R}^n$ *as follows*

$$\alpha := (u \circ M_c) \cdot \mathbf{1}_n.$$

We define $\widetilde{\alpha}(M) \in \mathbb{R}^n$ *as follows*

$$\widetilde{\alpha}(M) := (\widetilde{u}(M) \circ M_c) \cdot \mathbf{1}_n.$$

We define i_0*-th entry of* $\widetilde{\alpha}(M)$ *as follows*

$$\widetilde{\alpha}(M)_{i_0} := \langle (\widetilde{u}(M) \circ M_c)_{i_0}, \mathbf{1}_n \rangle$$

Then, we introduce the Softmax probability function.

Definition B.3 (Softmax probability function f, $\widetilde{f}$)**.** *If the following conditions hold*

- *Let* $M \in [0,1]^{d \times d}$.
- *Let* $M_c \in \{0,1\}^{n \times n}$ *be the causal attention mask defined in Definition 3.1.*
- *Let* u*,* $\widetilde{u}(M)$ *be defined as Definition B.1.*
- *Let* α*,* $\widetilde{\alpha}(M)$ *be defined as Definition B.2.*
- *Let* $i_0, j_0 \in [n]$.

We define $f \in \mathbb{R}^{n \times n}$ *for each* $j \in [n]$ *as follows*

$$f := \mathrm{diag}(\alpha)^{-1}(u \circ M_c).$$

We define $\widetilde{f}(M) \in \mathbb{R}^{n \times n}$ *for each* $j \in [n]$ *as follows*

$$\widetilde{f}(M) := \mathrm{diag}(\widetilde{\alpha}(M))^{-1}(\widetilde{u}(M) \circ M_c).$$

We define i_0*-th row of* $\widetilde{f}(M)$ *as follows*

$$\widetilde{f}(M)_{i_0} := \widetilde{\alpha}(M)_{i_0}^{-1}(\widetilde{u}(M) \circ M_c)_{i_0}.$$

We define the entry in i_0*-th row,* j_0*-th column of* $\widetilde{f}(M)$ *as follows*

$$\widetilde{f}(M)_{i_0,j_0} := \widetilde{\alpha}(M)_{i_0}^{-1}(\widetilde{u}(M) \circ M_c)_{i_0,j_0}.$$

Then, we introduce the one unit loss function.

Definition B.4 (One unit loss function c)**.** *If the following conditions hold*

- *Let* f*,* $\widetilde{f}$ *be defined in Definition B.3.*
- *Let* $M \in [0,1]^{d \times d}$*.*
- *Let* $i_0, j_0 \in [n]$*.*

We define $c(M) \in \mathbb{R}^{n \times n}$ *as follows*

$$c(M) := \widetilde{f}(M) - f$$

We define i_0*-th row of* $c(M)$ *as follows*

$$c(M)_{i_0} := \widetilde{f}(M)_{i_0} - f_{i_0}$$

We define j_0*-th column of* $c(M)$ *as follows*

$$c(M)_{*,j_0} := \widetilde{f}(M)_{*,j_0} - f_{*,j_0}$$

We define the entry in i_0*-th row,* j_0*-th column of* $c(M)$ *as follows*

$$c(M)_{i_0,j_0} := \widetilde{f}(M)_{i_0,j_0} - f_{i_0,j_0}$$

Then, we introduce the reconstruction error.

Definition B.5 (Reconstruction Error $\mathcal{L}_{\mathrm{attn}}$)**.** *If the following conditions hold*

- *Let* $M \in [0,1]^{d \times d}$*.*
- *Let* $c(M)$ *be defined in Definition B.4.*

We define $\mathcal{L}_{\mathrm{attn}}(M) \in \mathbb{R}$ *as follows*

$$\mathcal{L}_{\mathrm{attn}}(M) := \frac{1}{2}\|c(M)\|_F^2 = \frac{1}{2}\sum_{i_0=1}^{n}\sum_{j_0=1}^{n} c(M)_{i_0,j_0}^2.$$

Then, we introduce the regularization term.

Definition B.6 (Regularization Term $\mathcal{L}_{\mathrm{reg}}$)**.** *If the following conditions hold*

- $M \in [0,1]^{d \times d}$*.*

We define $\mathcal{L}_{\mathrm{reg}}(M) \in \mathbb{R}$ *as follows*

$$\mathcal{L}_{\mathrm{reg}}(M) := \frac{1}{2}\lambda\|M\|_F^2.$$

Finally, we introduce the overall loss function.

Definition B.7 (Overall loss function $\mathcal{L}$)**.** *If the following conditions hold*

- *Let* $M \in [0,1]^{d \times d}$.
- *Let* $\mathcal{L}_{\mathrm{attn}}(M)$ *be defined in Definition B.5.*
- *Let* $\mathcal{L}_{\mathrm{reg}}(M)$ *be defined in Definition B.6.*
- *Let* $\lambda \in \mathbb{R}_+$ *be the regularization parameter.*

We define $\mathcal{L}(M)$ *as follows*

$$\mathcal{L}(M) := \mathcal{L}_{\mathrm{attn}}(M) + \mathcal{L}_{\mathrm{reg}}(M)$$

B.2 GRADIENT FOR EACH ROW OF $X(M \circ W)X^\top$

We introduce the Lemma of gradient for each row of $X(M \circ W)X^\top$.

Lemma B.8. *Let* $i_1 \in [d]$*,* $j_1 \in [d]$*,* $i_0 \in [n]$*, we have*

$$\underbrace{\frac{\mathrm{d}(X(M \circ W)X^\top)_{i_0}}{\mathrm{d}M_{i_1,j_1}}}_{n \times 1} = \underbrace{W_{i_1,j_1}}_{\mathrm{scalar}} \underbrace{X_{i_0,i_1}}_{\mathrm{scalar}} \underbrace{X_{*,j_1}}_{n \times 1}$$

Proof. We can simplify the derivative expression

$$\begin{aligned} \frac{\mathrm{d}(X(M \circ W)X^\top)_{i_0}}{\mathrm{d}M_{i_1,j_1}} &= \frac{\mathrm{d}X(X(M \circ W))_{i_0}}{\mathrm{d}M_{i_1,j_1}} \\ &= \frac{\mathrm{d}X(M \circ W)^\top X_{i_0}}{\mathrm{d}M_{i_1,j_1}} \\ &= X\frac{\mathrm{d}(M \circ W)^\top}{\mathrm{d}M_{i_1,j_1}} X_{i_0} \end{aligned} \tag{3}$$

where the first and second step follows from Fact A.1, the third step follows from Fact A.3.

We further compute Eq. (3):

$$\begin{aligned} \frac{\mathrm{d}(M \circ W)^\top}{\mathrm{d}M_{i_1,j_1}} &= \frac{\mathrm{d}M^\top \circ W^\top}{\mathrm{d}M_{i_1,j_1}} \\ &= \frac{\mathrm{d}M^\top \circ W^\top}{\mathrm{d}(M^\top)_{j_1,i_1}} \\ &= (W^\top)_{j_1,i_1} e_{j_1} e_{i_1}^\top \\ &= W_{i_1,j_1} e_{j_1} e_{i_1}^\top \end{aligned} \tag{4}$$

where the first follows from Fact A.4, the second step follows from for any matrix X, $X_{i,j} = (X^\top)_{j,i}$, the third step follows from Fact A.7, and the fourth step follows from for any matrix X, $X_{i,j} = (X^\top)_{j,i}$.

Finally, we have

$$\begin{aligned} \frac{\mathrm{d}(X(M \circ W)X^\top)_{i_0}}{\mathrm{d}M_{i_1,j_1}} &= XW_{i_1,j_1} e_{j_1} e_{i_1}^\top X_{i_0} \\ &= W_{i_1,j_1}(Xe_{j_1})(e_{i_1}^\top X_{i_0}) \\ &= W_{i_1,j_1} X_{*,j_1} X_{i_0,i_1} \end{aligned}$$

where the first step follows from Eq. (3) and Eq. (4), and the second step and the third step follows from basic algebra. $\square$

We introduce the Lemma of gradient for each row of $\widetilde{u}(M)$.

B.3 GRADIENT FOR EACH ROW OF $\widetilde{u}(M)$

Lemma B.9. *If the following condition hold:*

- *Let $\widetilde{u}(M)$ be defined in Definition B.1.*

Let $i_1 \in [d]$, $j_1 \in [d]$, $i_0 \in [n]$, we have

$$\underbrace{\frac{\mathrm{d}\widetilde{u}(M)_{i_0}}{\mathrm{d}M_{i_1,j_1}}}_{n\times 1} = \underbrace{\widetilde{u}(M)_{i_0}}_{n\times 1} \circ (\underbrace{W_{i_1,j_1}}_{\text{scalar}} \underbrace{X_{i_0,i_1}}_{\text{scalar}} \underbrace{X_{*,j_1}}_{n\times 1})$$

Proof. We have

$$\begin{aligned}
\underbrace{\frac{\mathrm{d}\widetilde{u}(M)_{i_0}}{\mathrm{d}M_{i_1,j_1}}}_{n\times 1} &= \underbrace{\frac{\mathrm{d}\exp(X(M\circ W)X^\top)_{i_0}}{\mathrm{d}M_{i_1,j_1}}}_{n\times 1} \\
&= \exp(\underbrace{X}_{n\times d} \underbrace{(M\circ W)}_{d\times d} \underbrace{X^\top}_{d\times 1})_{i_0} \circ \underbrace{\frac{\mathrm{d}(X(M\circ W)X^\top)_{i_0}}{\mathrm{d}M_{i_1,j_1}}}_{n\times 1} \\
&= \underbrace{\widetilde{u}(M)_{i_0}}_{n\times 1} \circ \underbrace{\frac{\mathrm{d}(X(M\circ W)X^\top)_{i_0}}{\mathrm{d}M_{i_1,j_1}}}_{n\times 1} \\
&= \underbrace{\widetilde{u}(M)_{i_0}}_{n\times 1} \circ (\underbrace{W_{i_1,j_1}}_{\text{scalar}} \underbrace{X_{i_0,i_1}}_{\text{scalar}} \underbrace{X_{*,j_1}}_{n\times 1})
\end{aligned}$$

where the first step follows from Definition B.1, the second step follows from Fact A.3, the third step follows from Definition B.1, and the fourth step follows from Lemma B.8. ☐

B.4 GRADIENT FOR EACH ENTRY OF $\widetilde{\alpha}(M)$

We introduce the Lemma of gradient for each entry of $\widetilde{\alpha}(M)$.

Lemma B.10. *If the following conditions hold:*

- *Let $\widetilde{u}(M)$ be defined in Definition B.1.*
- *Let $\widetilde{\alpha}(M)$ be defined in Definition B.2.*

Let $i_1 \in [d]$, $j_1 \in [d]$, $i_0 \in [n]$, we have

$$\underbrace{\frac{\mathrm{d}\widetilde{\alpha}(M)_{i_0}}{\mathrm{d}M_{i_1,j_1}}}_{\text{scalar}} = \langle \widetilde{u}(M)_{i_0} \circ (M_c)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle$$

Proof. We have

$$\begin{aligned}
\underbrace{\frac{\mathrm{d}\widetilde{\alpha}(M)_{i_0}}{\mathrm{d}M_{i_1,j_1}}}_{\text{scalar}} &= \underbrace{\frac{\mathrm{d}\langle(\widetilde{u}(M)\circ M_c)_{i_0}, \mathbf{1}_n\rangle}{\mathrm{d}M_{i_1,j_1}}}_{\text{scalar}} \\
&= \langle \frac{\mathrm{d}(\widetilde{u}(M)\circ M_c)_{i_0}}{\mathrm{d}M_{i_1,j_1}}, \mathbf{1}_n \rangle \\
&= \langle \frac{\mathrm{d}\widetilde{u}(M)_{i_0}}{\mathrm{d}M_{i_1,j_1}} \circ (M_c)_{i_0}, \mathbf{1}_n \rangle \\
&= \langle \widetilde{u}(M)_{i_0} \circ (W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1}) \circ (M_c)_{i_0}, \mathbf{1}_n \rangle
\end{aligned}$$

$$= \langle \widetilde{u}(M)_{i_0} \circ (M_c)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle$$

where the first step follows from Definition B.2, the second step follows from product rule of inner product in Fact A.3, the third step follows from product rule of Hadamard product in Fact A.3, the fourth step follows from Lemma B.9, and the last step follows from Fact A.4. ☐

B.5 GRADIENT FOR EACH ENTRY OF $\widetilde{f}(M)$

We introduce the Lemma of gradient for each entry of $\widetilde{f}(M)$.

Lemma B.11. *If the following conditions hold:*

- *Let $\widetilde{u}(M)$ be defined in Definition B.1.*
- *Let $\widetilde{\alpha}(M)$ be defined in Definition B.2.*
- *Let $\widetilde{f}(M)$ be defined in Definition B.3.*

Let $i_1 \in [d]$, $j_1 \in [d]$, $i_0 \in [n]$, $j_0 \in [n]$, we have

$$\frac{\mathrm{d}\widetilde{f}(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} = \widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1} - \widetilde{f}(M)_{i_0,j_0} \langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle$$

Proof. We have

$$\begin{aligned}\frac{\mathrm{d}\widetilde{f}(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} &= \frac{\mathrm{d}\widetilde{\alpha}(M)_{i_0}^{-1}(\widetilde{u}(M) \circ M_c)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} \\ &= \frac{\mathrm{d}\widetilde{\alpha}(M)_{i_0}^{-1}}{\mathrm{d}M_{i_1,j_1}}(\widetilde{u}(M) \circ M_c)_{i_0,j_0} + \frac{\mathrm{d}(\widetilde{u}(M) \circ M_c)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} \widetilde{\alpha}(M)_{i_0}^{-1}\end{aligned} \tag{5}$$

where the first step follows from Definition B.3, and the second step follows from Fact A.3.

In the following part, we compute the two terms separately.

For the first term above, we have

$$\begin{aligned}&\frac{\mathrm{d}\widetilde{\alpha}(M)_{i_0}^{-1}}{\mathrm{d}M_{i_1,j_1}}(\widetilde{u}(M) \circ M_c)_{i_0,j_0} \\ &= (\widetilde{u}(M) \circ M_c)_{i_0,j_0}(-1)\widetilde{\alpha}(M)_{i_0}^{-2} \frac{\mathrm{d}\widetilde{\alpha}(M)_{i_0}}{\mathrm{d}M_{i_1,j_1}} \\ &= -(\widetilde{u}(M) \circ M_c)_{i_0,j_0} \langle \widetilde{u}(M)_{i_0} \circ (M_c)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle / \widetilde{\alpha}(M)_{i_0}^2 \\ &= -(\widetilde{\alpha}(M)_{i_0}^{-1} (M_c)_{i_0,j_0} \widetilde{u}(M)_{i_0,j_0}) \langle \widetilde{\alpha}(M)_{i_0}^{-1} \widetilde{u}(M)_{i_0} \circ (M_c)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle \\ &= -\widetilde{f}(M)_{i_0,j_0} \langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle\end{aligned} \tag{6}$$

where the first step follows from Fact A.3, the second step follows from Lemma B.10, the third step follows from basic algebra, and the fourth step follows from Definition B.3.

For the second term above, we have

$$\begin{aligned}&\frac{\mathrm{d}(\widetilde{u}(M) \circ M_c)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} \widetilde{\alpha}(M)_{i_0}^{-1} \\ &= \frac{\mathrm{d}\widetilde{u}(M)_{i_0,j_0}(M_c)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} \widetilde{\alpha}(M)_{i_0}^{-1} \\ &= (M_c)_{i_0,j_0} (\frac{\mathrm{d}\widetilde{u}(M)_{i_0}}{\mathrm{d}M_{i_1,j_1}})_{j_0} \widetilde{\alpha}(M)_{i_0}^{-1} \\ &= ((M_c)_{i_0,j_0} \widetilde{u}(M)_{i_0,j_0} \widetilde{\alpha}(M)_{i_0}^{-1}) W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1}\end{aligned}$$

$$= \widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1} \tag{7}$$

where the first step and the second step follow from basic algebra, the third step follows from Lemma B.9, and the fourth step follows from Definition B.3.

So, we have

$$\begin{aligned}\frac{\mathrm{d}\widetilde{f}(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} &= \frac{\mathrm{d}\widetilde{\alpha}(M)_{i_0}^{-1}}{\mathrm{d}M_{i_1,j_1}}(\widetilde{u}(M)\circ M_c)_{i_0,j_0} + \frac{\mathrm{d}(\widetilde{u}(M)\circ M_c)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}}\widetilde{\alpha}(M)_{i_0}^{-1} \\ &= \widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1} - \widetilde{f}(M)_{i_0,j_0}\langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1}\rangle\end{aligned}$$

where the first step follows from Eq. (5), and the second step follows from Eq. (6) and Eq. (7). ☐

B.6 GRADIENT FOR EACH ENTRY OF $c(M)$

We introduce the Lemma of gradient for each entry of $c(M)$.

Lemma B.12. *If the following conditions hold:*

- *Let $\widetilde{f}(M)$, f be defined in Definition B.3.*
- *Let $c(M)$ be defined in Definition B.4.*

Let $i_1 \in [d]$, $j_1 \in [d]$, $i_0 \in [n]$, $j_0 \in [n]$, we have

$$\frac{\mathrm{d}c(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} = \widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1} - \widetilde{f}(M)_{i_0,j_0}\langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1}\rangle$$

Proof. We have

$$\begin{aligned}\frac{\mathrm{d}c(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} &= \frac{\mathrm{d}(\widetilde{f}(M)_{i_0,j_0} - f_{i_0,j_0})}{\mathrm{d}M_{i_1,j_1}} \\ &= \frac{\mathrm{d}\widetilde{f}(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} \\ &= \widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1} - \widetilde{f}(M)_{i_0,j_0}\langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1}\rangle\end{aligned}$$

where the first step follows from Definition B.4, the second step follows from Fact A.3, and the third step follows from Lemma B.11. ☐

B.7 GRADIENT FOR $\mathcal{L}_{\mathrm{attn}}(M)$

We introduce the Lemma of gradient for $\mathcal{L}_{\mathrm{attn}}(M)$.

Lemma B.13. *If the following conditions hold:*

- *Let $\widetilde{f}(M)$ be defined in Definition B.3.*
- *Let $c(M)$ be defined in Definition B.4.*
- *Let $\mathcal{L}_{\mathrm{attn}}(M)$ be defined in Definition B.5.*

Let $i_1 \in [d]$, $j_1 \in [d]$, $i_0 \in [n]$, $j_0 \in [n]$, we have

$$\frac{\mathrm{d}\mathcal{L}_{\mathrm{attn}}(M)}{\mathrm{d}M_{i_1,j_1}} = \sum_{i_0=1}^{n}\sum_{j_0=1}^{n} B_1(M) + B_2(M)$$

where we have definitions:

- $B_1(M) := c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1}$
- $B_2(M) := -c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1}\rangle$

Proof. We have

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}_{\mathrm{attn}}(M)}{\mathrm{d}M_{i_1,j_1}} &= \frac{1}{2}\frac{\mathrm{d}\|c(M)\|_F^2}{\mathrm{d}M_{i_1,j_1}} \\
&= \frac{1}{2}\frac{\mathrm{d}\sum_{i_0=1}^n \sum_{j_0=1}^n (c(M)_{i_0,j_0})^2}{\mathrm{d}M_{i_1,j_1}} \\
&= \frac{1}{2}\sum_{i_0=1}^n \sum_{j_0=1}^n \frac{\mathrm{d}(c(M)_{i_0,j_0})^2}{\mathrm{d}M_{i_1,j_1}} \\
&= \sum_{i_0=1}^n \sum_{j_0=1}^n c(M)_{i_0,j_0}\frac{\mathrm{d}c(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}}
\end{aligned}$$

where the first step follows from Definition B.5, the second step follows from the definition of Frobenius norm of matrix, the third step follows from Fact A.3, and the fourth step follows from Fact A.3.

Following Lemma B.12, we have

$$\begin{aligned}
&\sum_{i_0=1}^n \sum_{j_0=1}^n c(M)_{i_0,j_0}\frac{\mathrm{d}c(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}} \\
&= \sum_{i_0=1}^n \sum_{j_0=1}^n c(M)_{i_0,j_0}(\widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1} - \widetilde{f}(M)_{i_0,j_0}\langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1}\rangle) \\
&= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1} - c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1}\rangle \\
&:= \sum_{i_0=1}^n \sum_{j_0=1}^n B_1(M) + B_2(M)
\end{aligned}$$

where the second step follows from basic algebra. ☐

B.8 GRADIENT FOR $\mathcal{L}_{\mathrm{reg}}(M)$

We introduce the Lemma of gradient for $\mathcal{L}_{\mathrm{reg}}(M)$.

Lemma B.14. *If the following condition hold:*

- *Let $\mathcal{L}_{\mathrm{reg}}(M)$ be defined in Definition B.6.*

Let $i_1 \in [d]$, $j_1 \in [d]$, we have

$$\frac{\mathrm{d}\mathcal{L}_{\mathrm{reg}}(M)}{\mathrm{d}M_{i_1,j_1}} = B_3(M)$$

where we have definition:

- $B_3(M) := \lambda M_{i_1,j_1}$

Proof. We have

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}_{\mathrm{reg}}(M)}{\mathrm{d}M_{i_1,j_1}} &= \frac{1}{2}\lambda\frac{\mathrm{d}\|M\|_F^2}{\mathrm{d}M_{i_1,j_1}} \\
&= \frac{1}{2}\lambda(\frac{\mathrm{d}}{\mathrm{d}M_{i_1,j_1}}\sum_{i_0=1}^d \sum_{j_0=1}^d M_{i_0,j_0}^2) \\
&= \lambda M_{i_1,j_1} \\
&:= B_3(M)
\end{aligned}$$

where the first step follows from Definition B.6, the second step follows from the definition of Frobenius norm of matrix, and the third step follows from Fact A.3. ☐

B.9 GRADIENT FOR $\mathcal{L}(M)$

We introduce the Lemma of gradient for $\mathcal{L}(M)$.

Lemma B.15. *If the following conditions hold:*

- *Let $\widetilde{u}(M)$ be defined in Definition B.1.*
- *Let $\widetilde{\alpha}(M)$ be defined in Definition B.2.*
- *Let $\widetilde{f}(M)$ be defined in Definition B.3.*
- *Let $\mathcal{L}_{\mathrm{attn}}(M)$ be defined in Definition B.5.*
- *Let $\mathcal{L}_{\mathrm{reg}}(M)$ be defined in Definition B.6.*
- *Let $\mathcal{L}(M)$ be defined in Definition B.7.*

Let $i_1 \in [d]$, $j_1 \in [d]$, $i_0 \in [n]$, $j_0 \in [n]$, we have

$$\frac{\mathrm{d}\mathcal{L}(M)}{\mathrm{d}M_{i_1,j_1}} = \sum_{i_0=1}^{n} \sum_{j_0=1}^{n} (B_1(M) + B_2(M)) + B_3(M)$$

where we have definitions:

- $B_1(M) := c(M)_{i_0,j_0} \widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1}$
- $B_2(M) := -c(M)_{i_0,j_0} \widetilde{f}(M)_{i_0,j_0} \langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle$
- $B_3(M) := \lambda M_{i_1,j_1}$

Proof.

$$\begin{aligned} \frac{\mathrm{d}\mathcal{L}(M)}{\mathrm{d}M_{i_1,j_1}} &= \frac{\mathrm{d}\mathcal{L}_{\mathrm{attn}}(M) + \mathcal{L}_{\mathrm{reg}}(M)}{\mathrm{d}M_{i_1,j_1}} \\ &= \frac{\mathrm{d}\mathcal{L}_{\mathrm{attn}}(M)}{\mathrm{d}M_{i_1,j_1}} + \frac{\mathrm{d}\mathcal{L}_{\mathrm{reg}}(M)}{\mathrm{d}M_{i_1,j_1}} \\ &= \sum_{i_0=1}^{n} \sum_{j_0=1}^{n} (B_1(M) + B_2(M)) + B_3(M) \end{aligned}$$

where the first step follows from Definition B.7, the second step follows from Fact A.3, and the third step follows from Lemma B.13 and Lemma B.14. ☐

C MATRIX FORM

C.1 MATRIX FORM OF $B(M)$

Given the matrix form, we define p to simplify the calculation.

Definition C.1. *If the following conditions hold*

- *Let $X \in \mathbb{R}^{n \times d}$.*
- *Let $M \in [0,1]^{d \times d}$.*
- *Let $W \in \mathbb{R}^{d \times d}$.*
- *Let $c(M)$ be defined in Definition B.4.*
- *Let $\widetilde{f}(M)$ be defined in Definition B.3.*

We define p_1 as follows

$$p_1 := c(M) \circ \widetilde{f}(M)$$

We define the j_0-th column of p_1 as follows

$$(p_1)_{*,j_0} := (c(M) \circ \widetilde{f}(M))_{*,j_0}$$

We define p_2 as follows

$$p_2 := \mathrm{diag}(p_1 \cdot \mathbf{1}_n)\widetilde{f}(M)$$

We define the i_0-th row of p_2 as follows

$$(p_2)_{i_0} := \mathbf{1}_n^\top (p_1)_{i_0} \widetilde{f}(M)_{i_0} = \widetilde{f}(M)_{i_0} c(M)_{i_0}^\top \widetilde{f}(M)_{i_0}$$

We define p as follows

$$p := p_1 - p_2 = c(M) \circ \widetilde{f}(M) - \mathrm{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)$$

We introduce the matrix view of $B_1(M)$ and its summation.

Lemma C.2 (Matrix view of $B_1(M)$)**.** *If we have the below conditions,*

- *Let $B_1(M, i_1, j_1) := c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}W_{i_1,j_1}X_{i_0,i_1}X_{j_0,j_1}$, which is defined in Lemma B.15*
- *We define $C_1(M) \in \mathbb{R}^{d\times d}$. For all $i_1, j_1 \in [d]$, let $C_1(i_1, j_1)$ denote the (i_1, j_1)-th entry of $C_1(M)$. We define $C_1(i_1, j_1) = B_1(M, i_1, j_1)$.*

Then, we can show that

- *Part 1. For $i_0, j_0 \in [n]$*

$$C_1(M) = \underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}}_{\text{scalar}}(W \circ (X_{i_0}X_{j_0}^\top))$$

- *Part 2.*

$$\sum_{i_0=1}^{n}\sum_{j_0=1}^{n} C_1(M) = W \circ (X^\top p_1 X)$$

Proof. **Part 1.** We have

$$\begin{aligned} C_1(i_1, j_1) &= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}W_{i_1,j_1}X_{i_0,i_1}X_{j_0,j_1} \\ &= \underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}}_{\text{scalar}}(\underbrace{X_{i_0}}_{d\times 1})_{i_1}(\underbrace{W_{*,j_1}}_{d\times 1})_{i_1}\underbrace{(X_{j_0})_{j_1}}_{\text{scalar}} \\ &= \underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}}_{\text{scalar}}(\underbrace{\mathrm{diag}(X_{i_0})W_{*,j_1}}_{d\times 1})_{i_1}\underbrace{(X_{j_0})_{j_1}}_{\text{scalar}} \end{aligned}$$

where the first step follows from the definition of C_1, the second step follows from Fact A.1, and the third step follows from Fact A.2.

Following from Fact A.1, we can get j_1-th column of C_1

$$\begin{aligned} C_1(*, j_1) &= \underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}}_{\text{scalar}}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}\underbrace{W_{*,j_1}}_{d\times 1}\underbrace{(X_{j_0})_{j_1}}_{\text{scalar}} \\ &= \underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}}_{\text{scalar}}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}(\underbrace{W}_{d\times d}\underbrace{\mathrm{diag}(X_{j_0})}_{d\times d})_{*,j_1} \end{aligned}$$

where the second step follows from Fact A.2.

Following from Fact A.1, we can get $C_1(M)$

$$\begin{aligned} C_1(M) = &\ \underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}}_{\text{scalar}}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}\underbrace{W}_{d\times d}\underbrace{\mathrm{diag}(X_{j_0})}_{d\times d} \\ = &\ \underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}}_{\text{scalar}}(W \circ (X_{i_0}X_{j_0}^\top)) \end{aligned} \tag{8}$$

where the second step follows from Fact A.4.

Part 2. We further compute the summation of $C_1(M)$.

$$\begin{aligned} \sum_{i_0=1}^n \sum_{j_0=1}^n C_1(M) = &\ \sum_{i_0=1}^n \sum_{j_0=1}^n \underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}}_{\text{scalar}}(W \circ X_{i_0}X_{j_0}^\top) \\ = &\ \sum_{i_0=1}^n \sum_{j_0=1}^n (W \circ (c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}X_{i_0}X_{j_0}^\top)) \\ = &\ W \circ \sum_{i_0=1}^n \sum_{j_0=1}^n c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}X_{i_0}X_{j_0}^\top \\ = &\ W \circ \sum_{j_0=1}^n \sum_{i_0=1}^n ((p_1)_{*,j_0})_{i_0}X_{i_0}X_{j_0}^\top \end{aligned}$$

where the first step follows from Eq. (8), the second step follows from basic algebra, the third step follows from Fact A.4, and the fourth step follows from Definition C.1.

Then following from Fact A.2, we have

$$\begin{aligned} &W \circ \sum_{j_0=1}^n \sum_{i_0=1}^n ((p_1)_{*,j_0})_{i_0}X_{i_0}X_{j_0}^\top \\ = &\ W \circ \sum_{j_0=1}^n X^\top (p_1)_{*,j_0}X_{j_0}^\top \\ = &\ W \circ (X^\top p_1 X) \end{aligned}$$

□

We introduce the matrix view of $B_2(M)$ and its summation.

Lemma C.3 (Matrix view of $B_2(M)$)**.** *If we have the below conditions,*

- *Let* $B_2(M, i_1, j_1) := -c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1}X_{i_0,i_1}X_{*,j_1}\rangle$ *be defined in Lemma B.15.*
- *We define* $C_2(M) \in \mathbb{R}^{d\times d}$*. For all* $i_1, j_1 \in [d]$*, let* $C_2(i_1, j_1)$ *denote the* (i_1, j_1)*-th entry of* $C_2(M)$*. We define* $C_2(i_1, j_1) = B_2(M, i_1, j_1)$*.*

Then, we can show that

- *Part 1. For* $i_0, j_0 \in [n]$

$$C_2(M) = -\, c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}\underbrace{W}_{d\times d}\underbrace{\mathrm{diag}(X^\top \widetilde{f}(M)_{i_0})}_{d\times d}$$

- *Part 2.*

$$\sum_{i_0=1}^n \sum_{j_0=1}^n C_2(M) = -W \circ (X^\top p_2 X)$$

Proof. **Part 1.** We have

$$\begin{aligned} -C_2(i_1,j_1) &= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1}X_{i_0,i_1}X_{*,j_1}\rangle \\ &= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\widetilde{f}(M)_{i_0}^\top}_{1\times n}\underbrace{X_{*,j_1}W_{i_1,j_1}X_{i_0,i_1}}_{n\times 1} \\ &= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\widetilde{f}(M)_{i_0}^\top}_{1\times n}\underbrace{X_{*,j_1}(X_{i_0})_{i_1}(W_{*,j_1})_{i_1}}_{n\times 1} \\ &= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\widetilde{f}(M)_{i_0}^\top}_{1\times n}\underbrace{X_{*,j_1}(\mathrm{diag}(X_{i_0})W_{*,j_1})_{i_1}}_{n\times 1} \end{aligned}$$

where the first step follows from the definition of C_2, the second step, the third step and the fourth step follows from Fact A.4.

Following from Fact A.1, we can get j_1-th column of C_2

$$\begin{aligned} -C_2(*,j_1) &= \underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}\underbrace{W_{*,j_1}}_{d\times 1}\underbrace{c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\widetilde{f}(M)_{i_0}^\top}_{1\times n}\underbrace{X_{*,j_1}}_{n\times 1} \\ &= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}\underbrace{W_{*,j_1}}_{d\times 1}\underbrace{\widetilde{f}(M)_{i_0}^\top}_{1\times n}\underbrace{X_{*,j_1}}_{n\times 1} \\ &= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}\underbrace{W_{*,j_1}}_{d\times 1}\underbrace{(X^\top\widetilde{f}(M)_{i_0})_{j_1}}_{\text{scalar}} \\ &= c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}(\underbrace{W\,\mathrm{diag}(X^\top\widetilde{f}(M)_{i_0})}_{d\times d})_{*,j_1} \end{aligned}$$

where the second step and the fourth step follows from Fact A.4, and the third step follows from Fact A.1.

Following from Fact A.1, we can get C_2.

$$-C_2(M) = c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}\underbrace{W}_{d\times d}\underbrace{\mathrm{diag}(X^\top\widetilde{f}(M)_{i_0})}_{d\times d} \tag{9}$$

Part 2. We further compute the summation of C_2

$$\begin{aligned} -\sum_{i_0=1}^n\sum_{j_0=1}^n C_2(M) &= \sum_{i_0=1}^n\sum_{j_0=1}^n c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}\underbrace{\mathrm{diag}(X_{i_0})}_{d\times d}\underbrace{W}_{d\times d}\underbrace{\mathrm{diag}(X^\top\widetilde{f}(M)_{i_0})}_{d\times d} \\ &= \sum_{i_0=1}^n\sum_{j_0=1}^n c(M)_{i_0,j_0}\widetilde{f}(M)_{i_0,j_0}((X_{i_0}\widetilde{f}(M)_{i_0}^\top X)\circ W) \\ &= W\circ\sum_{i_0=1}^n(X_{i_0}\widetilde{f}(M)_{i_0}^\top X)\sum_{j_0=1}^n((p_1)_{i_0})_{j_0} \end{aligned}$$

where the first step follows from Eq. (9), the second step and the third step follows from Fact A.4.

Following from Fact A.2, we have

$$\begin{aligned} & W\circ\sum_{i_0=1}^n(X_{i_0}\widetilde{f}(M)_{i_0}^\top X)\sum_{j_0=1}^n((p_1)_{i_0})_{j_0} \\ = {} & W\circ\sum_{i_0=1}^n(X_{i_0}\widetilde{f}(M)_{i_0}^\top X)\mathbf{1}_n^\top(p_1)_{i_0} \\ = {} & W\circ(X^\top\,\mathrm{diag}(p_1\cdot\mathbf{1}_n)\widetilde{f}(M)X) \\ = {} & W\circ(X^\top p_2 X) \end{aligned}$$

where the third step follows from Definition C.1. $\square$

We introduce the matrix view of $B_3(M)$.

Lemma C.4 (Matrix view of $B_3(M)$)**.** *If the following conditions hold*

- *Let $B_3(M, i_1, j_1) := \lambda M_{i_1,j_1}$ be defined in Lemma B.15.*
- *We define $C_3(M) \in \mathbb{R}^{d\times d}$. For all $i_1, j_1 \in [d]$, let $C_3(i_1, j_1)$ denote the (i_1, j_1)-th entry of $C_3(M)$. We define $C_3(i_1, j_1) = B_3(M, i_1, j_1)$.*

We can show that

$$C_3(M) = \lambda M.$$

Proof. The proof is straightforward. By the definition of $C_3(M)$, for all $i_1, j_1 \in [d]$, the (i_1, j_1)-th entry of $C_3(M)$ is given by $C_3(i_1, j_1) = B_3(M, i_1, j_1) = \lambda M_{i_1,j_1}$. Thus, the entire matrix $C_3(M)$ has entries that correspond to those of λM. Therefore, we can conclude that $C_3(M) = \lambda M$ as required. ☐

C.2 MATRIX FORM OF $\frac{\mathrm{d}}{\mathrm{d}M}\mathcal{L}(M)$

We introduce the matrix form of overall loss function.

Theorem C.5 (Close form of gradient, formal version of Theorem 5.3)**.** *If the following conditions hold*

- *Let $\mathcal{L}(M)$ be defined in Definition B.7.*
- *Let p be defined in Definition C.1.*
- *Let $X \in \mathbb{R}^{n\times d}$.*
- *Let $M \in [0, 1]^{d\times d}$.*
- *Let $W \in \mathbb{R}^{d\times d}$.*

We can show that

$$\frac{\mathrm{d}\mathcal{L}(M)}{\mathrm{d}M} = W \circ (X^\top pX) + \lambda M.$$

Proof. We have

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}(M)}{\mathrm{d}M} &= \sum_{i_0=1}^{n}\sum_{j_0=1}^{n}(C_1(M) + C_2(M)) + C_3(M) \\
&= W \circ (X^\top p_1 X) - W \circ (X^\top p_2 X) + \lambda M \\
&= W \circ (X^\top (p_1 - p_2)X) + \lambda M \\
&= W \circ (X^\top pX) + \lambda M
\end{aligned}$$

where the first step follows from Lemma B.15, the second step follows from Lemma C.2, Lemma C.3, and Lemma C.4, the third step follows from basic algebra, and the fourth step follows from Definition C.1. ☐

D BOUNDS FOR BASIC FUNCTIONS

D.1 BASIC ASSUMPTIONS

Here we introduce our bounded parameters assumption.

Assumption D.1 (Bounded parameters)**.** *We assume the following conditions*

- *Let R be some fixed constant satisfies $R > 1$.*

- *Let $X \in \mathbb{R}^{n \times d}, W \in \mathbb{R}^{d \times d}$. We have $\|X\|_F \le R$ and $\|W\|_F \le R$.*

Here we present the lemma of bounds for M and M_c.

Lemma D.2 (Bounds for M and M_c)**.** *Let $M \in [0,1]^{d \times d}$ and $M_c \in \{0,1\}^{n \times n}$ be the causal attention mask defined in Definition 3.1. For M, we have*

$$\|M\|_F \le d$$

For M_c, we have

$$\|M_c\|_F \le n$$

Proof. This Lemma simply follows from the definition of Frobenius norm, given that the max value of each entry in M and M_c is 1. ☐

D.2 BOUNDS FOR BASIC FUNCTIONS

We first introduce the lemma of bounds for basic function.

Lemma D.3. *Under Assumption D.1, for all $i_0 \in [n]$, $j_0 \in [n]$, $i_1 \in [d]$, $j_1 \in [d]$, we have the following bounds*

- *Part 1.*
$$\|\widetilde{f}(M)\|_F \le \sqrt{n}$$
- *Part 2.*
$$\|c(M)\|_F \le 2\sqrt{n}$$
- *Part 3.*
$$\|(c(M) \circ \widetilde{f}(M))\|_F \le 2\sqrt{n}$$
- *Part 4.*
$$|\widetilde{f}(M)_{i_0,j_0}| \le 1$$
- *Part 5.*
$$|W_{i_1,j_1}| \le R$$
- *Part 6.*
$$|X_{i_0,i_1}| \le R$$
- *Part 7.*
$$\|\widetilde{f}(M)_{i_0}\|_2 \le 1$$
- *Part 8.*
$$|\widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1}| \le R^3$$
- *Part 9.*
$$|\widetilde{f}(M)_{i_0,j_0} \langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle| \le R^3$$
- *Part 10.*
$$\| \operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\|_F \le 2n$$

Proof. **Proof of Part 1.** Each entry in $\widetilde{f}(M)$ present a probability, thus for $i_0 \in [n]$, $j_0 \in [n]$, we have

$$0 \leq \widetilde{f}(M)_{i_0,j_0} \leq 1.$$

For any i_0-th row of $\widetilde{f}(M)$, following from the definition of Softmax function, we know

$$\sum_{j_0=1}^{n} \widetilde{f}(M)_{i_0,j_0} = 1.$$

So we have

$$\sum_{j_0=1}^{n} \widetilde{f}(M)_{i_0,j_0}^2 \leq 1$$

which follows from $\widetilde{f}(M)_{i_0,j_0} \leq (\widetilde{f}(M)_{i_0,j_0})^2$. Then, we can show

$$\|\widetilde{f}(M)\|_F = \sqrt{\sum_{i_0=1}^{n}\sum_{j_0=1}^{n} \widetilde{f}(M)_{i_0,j_0}^2} \leq \sqrt{n}$$

Proof of Part 2. Following from **Part 1.**, we can show

$$\|\widetilde{f}(M)\|_F \leq \sqrt{n}$$

and

$$\|f\|_F \leq \sqrt{n}.$$

Then we have

$$\begin{aligned}\|c(M)\|_F &= \|\widetilde{f}(M) - f\|_F \\ &\leq \|\widetilde{f}(M)\|_F + \|f\|_F \\ &\leq 2\sqrt{n}\end{aligned}$$

where the first step follows from Definition B.4, the second step follows from triangle inequality.

Proof of Part 3. We have $0 \leq \widetilde{f}(M)_{i_0,j_0} \leq 1$, so we have

$$\begin{aligned}\|(c(M) \circ \widetilde{f}(M))\|_F &\leq \|c(M)\|_F \\ &\leq 2\sqrt{n}\end{aligned}$$

where the second step follows from **Part 2.**.

Proof of Part 4. See **Proof of Part 1.**.

Proof of Part 5. The proof simply follows from Assumption D.1 and Fact A.5.

Proof of Part 6. The proof simply follows from Assumption D.1 and Fact A.5.

Proof of Part 7. See **Proof of Part 1.**.

Proof of Part 8. The proof simply follows from **Part 4.**, **Part 5.**, **Part 6.** and **Part 7.**.

Proof of Part 9. The proof simply follows from **Part 4.**, **Part 5.**, **Part 6.** and **Part 7.**.

Proof of Part 10. We have

$$\begin{aligned}\|\operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\|_F &= \|(c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n\|_2 \\ &\leq \|\mathbf{1}_n\|_2 \|(c(M) \circ \widetilde{f}(M))\|_F \\ &= \sqrt{n} \cdot 2\sqrt{n} \\ &= 2n\end{aligned}$$

where the first step follows from Fact A.4 the second step follows from Fact A.5, the third step follows from **Part 3.**, and the last step follows from simple algebra. ☐

D.3 BOUNDS FOR GRADIENT OF $\widetilde{f}(M)$

We introduce the lemma of bounds for gradient of $\widetilde{f}(M)$.

Lemma D.4. *If the following conditions hold*

- *Let $\widetilde{f}(M)$ be defined in Definition B.3.*
- *Assumption D.1 holds.*

Then we have

$$\|\frac{\mathrm{d}\,\mathrm{vec}(\widetilde{f}(M))}{\mathrm{d}\,\mathrm{vec}(M)}\|_F \leq 2dnR^3$$

Proof. We have

$$\begin{aligned}
& |\frac{\mathrm{d}f(M)_{i_0,j_0}}{\mathrm{d}M_{i_1,j_1}}| \\
= & |\widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1} - \widetilde{f}(M)_{i_0,j_0} \langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle| \\
\leq & |\widetilde{f}(M)_{i_0,j_0} W_{i_1,j_1} X_{i_0,i_1} X_{j_0,j_1}| + |\widetilde{f}(M)_{i_0,j_0} \langle \widetilde{f}(M)_{i_0}, W_{i_1,j_1} X_{i_0,i_1} X_{*,j_1} \rangle| \\
\leq & 2R^3
\end{aligned}$$

For $\frac{\mathrm{d}\,\mathrm{vec}(\widetilde{f}(M))}{\mathrm{d}\,\mathrm{vec}(M)}$, we can show

$$\begin{aligned}
\|\frac{\mathrm{d}\,\mathrm{vec}(\widetilde{f}(M))}{\mathrm{d}\,\mathrm{vec}(M)}\|_F = & \sqrt{\sum_{i_2=1}^{n^2} \sum_{j_2=1}^{d^2} |\frac{\mathrm{d}\,\mathrm{vec}(\widetilde{f}(M))_{i_0}}{\mathrm{d}\,\mathrm{vec}(M)_{j_0}}|} \\
\leq & 2ndR^3
\end{aligned}$$

□

E LIPSCHITZ OF GRADIENT

E.1 USEFUL FACTS

Here we introduce the fact of mean value theorem for matrix function.

Fact E.1 (Mean value theorem for matrix function, Fact C.6 in Liang et al. (2024d))**.** *If the following conditions hold*

- *Let $X, Y \in C \subset \mathbb{R}^{d \times d}$ where C is an open convex domain.*
- *Let $g(X) : C \to \mathbb{R}^{n \times n}$ be a differentiable matrix function on C.*
- *Let $\|\frac{\mathrm{d}\,\mathrm{vec}(g(X))}{\mathrm{d}\,\mathrm{vec}(X)}\|_F \leq R$ for all $x \in C$.*

We have

$$\|g(Y) - g(X)\|_F \leq R\|Y - X\|_F.$$

Proof. For the convenience of proof, we define x and y as follows:

- $x := \mathrm{vec}(X)$ and $y := \mathrm{vec}(Y)$.
- $h(x) := \mathrm{vec}(g(X))$ and $h(y) := \mathrm{vec}(g(Y))$.
- $h'(a)$ denotes a matrix which its (i, j)-th term is $\frac{\mathrm{d}h(a)_j}{\mathrm{d}a_i}$.

Assume we have 1-variable function $\gamma(c) = f(x + c(y - x))$, we can apply Mean Value Theorem:

$$f(y) - f(x) = \gamma(1) - \gamma(0) = \gamma'(t)(1 - 0) = \nabla f(x + t(y - x))^\top (y - x) \tag{10}$$

where $t \in [0, 1]$. Let $G(c) := (h(y) - h(x))^\top h(c)$, we have

$$\begin{aligned} \|g(Y) - g(X)\|_F^2 &= G(y) - G(x) \\ &= \nabla G(x + t(y - x))^\top (y - x) \\ &= (\underbrace{h'(x + t(y - x))}_{d^2 \times n^2} \cdot \underbrace{h(y) - h(x)}_{n^2 \times 1})^\top \cdot (y - x) \\ &\leq \|h'(x + t(y - x))\| \cdot \|h(y) - h(x)\|_2 \cdot \|y - x\|_2 \end{aligned}$$

where the second step follows from Eq. (10), the third step follows from chain rule, the fourth step follows from Cauchy-Schwartz inequality.

By definition of matrix Frobenius norm and vector ℓ_2 norm, we have

$$\|g(Y) - g(X)\|_F = \|h(y) - h(x)\|_2$$

and

$$\|Y - X\|_F = \|y - x\|_2$$

so, we can show

$$\|g(Y) - g(X)\|_F \leq R\|Y - X\|_F$$

which follows from $\|\frac{\mathrm{d}\,\mathrm{vec}(g(X))}{\mathrm{d}\,\mathrm{vec}(X)}\|_F \leq R$ for all $x \in C$. □

Here we introduce the fact of Lipschitz for product of functions.

Fact E.2 (Lipschitz for product of functions, Fact H.2 in Deng et al. (2023b))**.** *Under following conditions*

- *Let $\{f_i(x)\}_{i=1}^n$ be a sequence of function with same domain and range.*
- *For each $i \in [n]$, we have*
  - *$f_i(x)$ is bounded: $\forall x$, $\|f_i(x)\|_F \leq R_i$ with $R_i \geq 1$.*
  - *$f_i(x)$ is Lipschitz continuous: $\forall x, y$, $\|f_i(x) - f_i(y)\|_F \leq L_i\|x - y\|_F$.*

Then we have

$$\|\prod_{i=1}^n f_i(x) - \prod_{i=1}^n f_i(y)\|_F \leq 2^{n-1} \cdot \max_{i \in [n]}\{L_i\} \cdot (\prod_{i=1}^n R_i) \cdot \|x - y\|_F$$

E.2 LIPSCHITZ OF $\widetilde{f}(M)$

We introduce the lemma about Lipschitz of $\widetilde{f}(M)$.

Lemma E.3 (Lipschitz of $\widetilde{f}(M)$)**.** *Under the following conditions*

- *Assumption D.1 holds.*
- *Let $\widetilde{f}(M)$ be defined as Definition B.3.*

For $M, \widetilde{M} \in \mathbb{R}^{d \times d}$, we have

$$\|\widetilde{f}(M) - \widetilde{f}(\widetilde{M})\|_F \leq 2dnR^3\|M - \widetilde{M}\|_F$$

Proof. We have

$$\begin{aligned} \|\widetilde{f}(M) - \widetilde{f}(\widetilde{M})\|_F &\leq \|\nabla \widetilde{f}(M)\|_F \cdot \|M - \widetilde{M}\|_F \\ &\leq 2dnR^3 \cdot \|M - \widetilde{M}\|_F \end{aligned}$$

where the first step follows from Fact E.1, the second step follows from Lemma D.4. □

E.3 LIPSCHITZ OF $c(M)$

We introduce the lemma about Lipschitz of $c(M)$.

Lemma E.4 (Lipschitz of $c(M)$)**.** *Under the following conditions*

- *Assumption D.1 holds.*
- *Let $c(M)$ be defined as Definition B.4.*

For $M, \widetilde{M} \in \mathbb{R}^{d \times d}$, we have

$$\|c(M) - c(\widetilde{M})\|_F \leq 2dnR^3 \|M - \widetilde{M}\|_F$$

Proof. We have

$$\begin{aligned}\|c(M) - c(\widetilde{M})\|_F &\leq \|\nabla c(M)\|_F \cdot \|M - \widetilde{M}\|_F \\ &= \|\nabla \widetilde{f}(M)\|_F \cdot \|M - \widetilde{M}\|_F \\ &\leq 2dnR^3 \cdot \|M - \widetilde{M}\|_F\end{aligned}$$

where the first step follows from Fact E.1, the second step follows from Lemma B.12, the third step follows from Lemma D.4. □

E.4 LIPSCHITZ OF $\widetilde{f}(M) \circ c(M)$

We introduce the lemma about Lipschitz of $\widetilde{f}(M) \circ c(M)$.

Lemma E.5 (Lipschitz of $\widetilde{f}(M) \circ c(M)$)**.** *Under the following conditions*

- *Assumption D.1 holds.*
- *Let $c(M)$ be defined as Definition B.4.*
- *Let $\widetilde{f}(M)$ be defined as Definition B.3.*

For $M, \widetilde{M} \in \mathbb{R}^{d \times d}$, we have

$$\|\widetilde{f}(M) \circ c(M) - \widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})\|_F \leq 6dn^{3/2}R^3 \|M - \widetilde{M}\|_F$$

Proof. We have

$$\begin{aligned}\mathrm{LHS} &\leq \|\widetilde{f}(M) \circ c(M) - \widetilde{f}(M) \circ c(\widetilde{M})\|_F + \|\widetilde{f}(M) \circ c(\widetilde{M}) - \widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})\|_F \\ &\leq \|\widetilde{f}(M)\|_F \cdot \|c(M) - c(\widetilde{M})\|_F + \|c(\widetilde{M})\|_F \cdot \|\widetilde{f}(M) - \widetilde{f}(\widetilde{M})\|_F \\ &\leq \sqrt{n} \cdot \|c(M) - c(\widetilde{M})\|_F + 2\sqrt{n} \cdot \|\widetilde{f}(M) - \widetilde{f}(\widetilde{M})\|_F \\ &\leq \sqrt{n} \cdot 2dnR^3 \|M - \widetilde{M}\|_F + 2\sqrt{n} \cdot 2dnR^3 \|M - \widetilde{M}\|_F\end{aligned}$$

where the first step follows from triangle inequality, the second step follows from Fact A.5, the third step follows from Lemma D.3, the fourth step follows from Lemma E.4 and Lemma E.3.

So we have

$$\|\widetilde{f}(M) \circ c(M) - \widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})\|_F \leq 6dn^{3/2}R^3 \|M - \widetilde{M}\|_F$$

□

E.5 LIPSCHITZ OF $\mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)$

We introduce the lemma about Lipschitz of $\mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)$.

Lemma E.6 (Lipschitz of $\mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)$)**.** *If the following conditions hold*

- *Assumption D.1 holds.*
- *Let $c(M)$ be defined as Definition B.4.*
- *Let $\widetilde{f}(M)$ be defined as Definition B.3.*

For $M, \widetilde{M} \in \mathbb{R}^{d \times d}$, we have

$$\| \mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n) - \mathrm{diag}((\widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})) \cdot \mathbf{1}_n)\|_F \leq 6dn^2R^3\|M - \widetilde{M}\|_F$$

Proof. We have

$$\begin{aligned} \mathrm{LHS} &= \|(\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n - (\widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})) \cdot \mathbf{1}_n\|_2 \\ &= \|((\widetilde{f}(M) \circ c(M)) - (\widetilde{f}(\widetilde{M}) \circ c(\widetilde{M}))) \cdot \mathbf{1}_n\|_2 \\ &\leq \|\widetilde{f}(M) \circ c(M) - \widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})\| \cdot \|\mathbf{1}_n\|_2 \\ &= \sqrt{n}\|\widetilde{f}(M) \circ c(M) - \widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})\| \end{aligned} \tag{11}$$

where the first step follows from Fact A.4, the second step follows from basic algebra, the third step follows from Fact A.5, and the fourth step follows from $\|\mathbf{1}_n\|_2 = \sqrt{n}$.

Then we have

$$\|\widetilde{f}(M) \circ c(M) - \widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})\| \leq \|\widetilde{f}(M) \circ c(M) - \widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})\|_F \tag{12}$$

which follows from Fact A.5.

Following Eq. (11), Eq. (12) and Lemma E.5, we have

$$\mathrm{LHS} \leq \sqrt{n} \cdot 6dn^{3/2}R^3\|M - \widetilde{M}\|_F = 6dn^2R^3\|M - \widetilde{M}\|_F$$

□

E.6 LIPSCHITZ OF $\mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)$

We introduce the lemma about Lipschitz of $\mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)$.

Lemma E.7 (Lipschitz of $\mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)$)**.** *If the following conditions hold*

- *Assumption D.1 holds.*
- *Let $c(M)$ be defined as Definition B.4.*
- *Let $\widetilde{f}(M)$ be defined as Definition B.3.*

For $M, \widetilde{M} \in \mathbb{R}^{d \times d}$, we have

$$\| \mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)\widetilde{f}(M) - \mathrm{diag}((\widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M})\|_F \leq 24dn^{7/2}R^3\|M - \widetilde{M}\|_F$$

Proof. Following Fact E.2, we have

$$\begin{aligned} &\| \mathrm{diag}((\widetilde{f}(M) \circ c(M)) \cdot \mathbf{1}_n)\widetilde{f}(M) - \mathrm{diag}((\widetilde{f}(\widetilde{M}) \circ c(\widetilde{M})) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M})\|_F \\ \leq{}& 2^1 \cdot \max\{6dn^2R^3, 6dn^{3/2}R^3\} \cdot (\sqrt{n} \cdot 2n)\|M - \widetilde{M}\|_F \\ ={}& 24dn^{7/2}R^3\|M - \widetilde{M}\|_F \end{aligned}$$

where we have the upper bound in Lemma D.3, the Lipschitz of $\mathrm{diag}((\widetilde{f}(M) \circ c(M))$ and $\widetilde{f}(M)$ in Lemma E.3 and Lemma E.6. □

E.7 LIPSCHITZ OF GRADIENT

We introduce the lemma about Lipschitz of the gradient.

Theorem E.8 (Lipschitz of the gradient, formal version of Theorem 5.4)**.** *We can show $\nabla_M \mathcal{L}(M)$ is L-Lipschitz.*

If the following conditions hold

- *Assumption D.1 holds.*
- *Let $c(M)$ be defined as Definition B.4.*
- *Let $\widetilde{f}(M)$ be defined as Definition B.3.*

For $M, \widetilde{M} \in \mathbb{R}^{d \times d}$, we have

$$\|\nabla_M \mathcal{L}(M) - \nabla_M \mathcal{L}(\widetilde{M})\|_F \leq (\lambda + 30dn^{7/2}R^6) \cdot \|M - \widetilde{M}\|_F$$

Proof. We have

$$\begin{aligned}
&\|\nabla_M \mathcal{L}(M) - \nabla_M \mathcal{L}(\widetilde{M})\|_F \\
=& \|W \circ (X^\top (c(M) \circ \widetilde{f}(M) - \mathrm{diag}(c(M) \circ \widetilde{f}(M) \cdot \mathbf{1}_n)\widetilde{f}(M) \\
&- c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) + \mathrm{diag}(c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M}))X) + \lambda M - \lambda \widetilde{M}\|_F \\
\leq& \|W \circ (X^\top (c(M) \circ \widetilde{f}(M) - \mathrm{diag}(c(M) \circ \widetilde{f}(M) \cdot \mathbf{1}_n)\widetilde{f}(M) \\
&- c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) + \mathrm{diag}(c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M}))X)\|_F + \|\lambda(M - \widetilde{M})\|_F
\end{aligned} \tag{13}$$

where the first step follows from Theorem C.5, and the second step follows from triangle inequality. Now we proof these two terms separately.

For the first term, we have

$$\begin{aligned}
&\|W \circ (X^\top (c(M) \circ \widetilde{f}(M) - \mathrm{diag}(c(M) \circ \widetilde{f}(M) \cdot \mathbf{1}_n)\widetilde{f}(M) \\
&- c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) + \mathrm{diag}(c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M}))X)\|_F \\
\leq& \|W\|_F \cdot \|X^\top (c(M) \circ \widetilde{f}(M) - \mathrm{diag}(c(M) \circ \widetilde{f}(M) \cdot \mathbf{1}_n)\widetilde{f}(M) \\
&- c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) + \mathrm{diag}(c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M}))X\|_F \\
\leq& \|W\|_F \cdot \|X\|_F^2 \cdot \|c(M) \circ \widetilde{f}(M) - \mathrm{diag}(c(M) \circ \widetilde{f}(M) \cdot \mathbf{1}_n)\widetilde{f}(M) \\
&- c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) + \mathrm{diag}(c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M})\|_F \\
\leq& \|W\|_F \cdot \|X\|_F^2 \cdot (\|c(M) \circ \widetilde{f}(M) - c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M})\|_F \\
&+ \|\,\mathrm{diag}(c(M) \circ \widetilde{f}(M) \cdot \mathbf{1}_n)\widetilde{f}(M) - \mathrm{diag}(c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M})\|_F) \\
=& R^3 \cdot (\|c(M) \circ \widetilde{f}(M) - c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M})\|_F \\
&+ \|\,\mathrm{diag}(c(M) \circ \widetilde{f}(M) \cdot \mathbf{1}_n)\widetilde{f}(M) - \mathrm{diag}(c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M})\|_F)
\end{aligned}$$

where the first step and the second step follows from Fact A.5, the third step follows from triangle inequality, and the fourth step follows from Assumption D.1.

Then we have

$$\begin{aligned}
&R^3 \cdot (\|c(M) \circ \widetilde{f}(M) - c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M})\|_F \\
&+ \|\,\mathrm{diag}(c(M) \circ \widetilde{f}(M) \cdot \mathbf{1}_n)\widetilde{f}(M) - \mathrm{diag}(c(\widetilde{M}) \circ \widetilde{f}(\widetilde{M}) \cdot \mathbf{1}_n)\widetilde{f}(\widetilde{M})\|_F) \\
\leq& R^3 \cdot (24dn^{7/2}R^3\|M - \widetilde{M}\|_F + 6dn^{3/2}R^3\|M - \widetilde{M}\|_F) \\
\leq& R^3 \cdot (30dn^{7/2}R^3\|M - \widetilde{M}\|_F) \\
=& 30dn^{7/2}R^6\|M - \widetilde{M}\|_F
\end{aligned} \tag{14}$$

where the first step follows from Lemma E.5 and Lemma E.7, the second step follows from $n \geq 1$.

For the second term, we have

$$\|\lambda(M - \widetilde{M})\|_F = \lambda \|M - \widetilde{M}\|_F \tag{15}$$

which follows from Fact A.5.

Finally, we have

$$\|\nabla_M \mathcal{L}(M) - \nabla_M \mathcal{L}(\widetilde{M})\|_F \leq (\lambda + 30dn^{7/2}R^6) \cdot \|M - \widetilde{M}\|_F$$

which follows from Eq. (13), Eq. (14), and Eq. (15). ☐

F CONVERGENCE OF GRADIENT DESCENT

F.1 HELPFUL STATEMENTS

Here, we present useful fact that we use to prove our convergence result.

Fact F.1. *We can show that for $a, b \in \mathbb{R}$*

- *Part 1.*

$$\sqrt{a^2 + b^2} \geq \frac{|a| + |b|}{\sqrt{2}}$$

- *Part 2. Suppose $|a| > |b|$*

$$\sqrt{|a| - |b|} \geq \sqrt{|a|} - \sqrt{|b|}$$

Proof. **Proof of Part 1.** Square both side of the inequality in **Part 1.**, we have

$$\mathrm{LHS} = a^2 + b^2$$

and

$$\mathrm{RHS} = \frac{a^2 + 2|a| \cdot |b| + b^2}{2}.$$

So we just need to prove

$$\begin{aligned}
\mathrm{LHS} - \mathrm{RHS} &= a^2 + b^2 - \frac{a^2 + 2|a| \cdot |b| + b^2}{2} \\
&= \frac{a^2 + b^2 - 2|a| \cdot |b|}{2} \\
&= \frac{(|a| - |b|)^2}{2} \\
&\geq 0
\end{aligned}$$

which is hold because for any $x \in \mathbb{R}$, $x^2 \geq 0$.

Proof of Part 2. Square both side of the inequality in **Part 2.**, we have

$$\mathrm{LHS} = |a| - |b|$$

and

$$\mathrm{RHS} = |a| + |b| - 2\sqrt{|a||b|}$$

So we just need to prove

$$\begin{aligned}
\mathrm{LHS} - \mathrm{RHS} &= |a| - |b| - |a| - |b| + 2\sqrt{|a||b|} \\
&= 2\sqrt{|a||b|} - 2|b| \\
&= 2\sqrt{|b|}(\sqrt{|a|} - \sqrt{|b|}) \\
&\geq 0
\end{aligned}$$

which is hold because $|a| > |b|$ and $|b| \geq 0$. ☐

F.2 LOWER BOUND ON FROBENIUS NORM

We present the lemma for the lower bound on the Frobenius norm in this section.

Lemma F.2. *If the following conditions hold*

- *Let* $B \in \mathbb{R}^{d \times d}$.
- *Let* $M \in [0,1]^{d \times d}$.
- *Let* $\lambda \in [0,1]$ *be some constant.*
- *Suppose that* $\|B\|_F \leq R$.

Then, we can show

- *Part 1.*
$$\|B + \lambda M\|_F^2 \geq \|B\|_F^2 + \lambda^2 \|M\|_F^2 - 2R\lambda d$$
- *Part 2.*
$$\|B + \lambda M\|_F \geq \frac{1}{\sqrt{2}}(\|B\|_F + \lambda \|M\|_F) - \sqrt{2R\lambda d}$$

Proof. **Proof of Part 1.** We can show that

$$\begin{aligned} \|B + \lambda M\|_F^2 &= \|B\|_F^2 + \lambda^2 \|M\|_F^2 + 2\langle B, \lambda M \rangle \\ &\geq \|B\|_F^2 + \lambda^2 \|M\|_F^2 - 2\|B\|_F \cdot \|\lambda M\|_F \\ &\geq \|B\|_F^2 + \lambda^2 \|M\|_F^2 - 2R\lambda d \end{aligned} \tag{16}$$

where the first step follows from Fact A.6, the second step follows from Fact A.5, the third step follows from the upper bound of $\|B\|_F$ and $\|M\|_F$.

Proof of Part 2. Taking the square root on both sides, we get

$$\begin{aligned} \|B + \lambda M\|_F &\geq \sqrt{\|B\|_F^2 + \lambda^2 \|M\|_F^2 - 2R\lambda d} \\ &\geq \sqrt{\|B\|_F^2 + \lambda^2 \|M\|_F^2} - \sqrt{2R\lambda d} \\ &\geq \frac{1}{\sqrt{2}}(\|B\|_F + \lambda \|M\|_F) - \sqrt{2R\lambda d} \end{aligned}$$

where the first step follows from Eq. (16), the second step follows from **Part 2.** of Fact F.1, and the third step follows from **Part 1.** of Fact F.1.

□

F.3 SANDWICH LOWER BOUND ON FROBENIUS NORM

Here, we introduce a sandwich trace fact.

Fact F.3. *If* $A \succeq \beta I$, *then* $\mathrm{tr}[B^\top A B] \geq \beta \, \mathrm{tr}[B^\top B]$.

Proof. As $A \succeq \beta I$, we have $A - \beta I \succeq 0$. Multiplying both sides by $B^\top$ on the left and B on the right (noting that these operations preserve the positive semidefiniteness), we have

$$B^\top (A - \beta I) B \succeq 0.$$

Taking the trace and utilizing the property that the trace of a positive semidefinite matrix is non-negative, we have

$$\mathrm{tr}[B^\top A B - \beta B^\top B] \geq 0,$$

which simplifies to

$$\mathrm{tr}[B^\top A B] - \beta \, \mathrm{tr}[B^\top B] \geq 0.$$

This concludes the proof. □

We establish a sandwich lower bound on the Frobenius norm.

Lemma F.4 (Formal version of Lemma 5.6)**.** *If the following conditions hold*

- *Let* $B \in \mathbb{R}^{n \times n}$ *and* $X \in \mathbb{R}^{n \times d}$.
- *Assume that* $XX^\top \succeq \beta I$.

Then, we have

$$\|X^\top B X\|_F \geq \beta \|B\|_F$$

Proof. We can show that

$$\begin{aligned} \|X^\top B X\|_F^2 = &~ \mathrm{tr}[X^\top B X X^\top B^\top X] \\ \geq &~ \beta \cdot \mathrm{tr}[X^\top B B^\top X] \\ = &~ \beta \cdot \mathrm{tr}[B^\top X X^\top B] \\ \geq &~ \beta^2 \cdot \mathrm{tr}[B^\top B] \\ = &~ \beta^2 \cdot \|B\|_F^2 \end{aligned}$$

where the first step, the third step and the fifth step follows from Fact A.4, the second step and the fourth step follows from Fact F.3 and $XX^\top \succeq \beta I$.

Taking the square root of both side, we finish the proof. ☐

F.4 LOWER BOUND ON HADAMARD PRODUCT BETWEEN TWO MATRICES

We present the lemma for lower bound on Hadamard product between two matrices in this section.

Lemma F.5. *If the following conditions hold*

- *Let* $B, W \in \mathbb{R}^{d \times d}$.

Then, we have

$$\max_{i,j \in [d]} \{|W_{i,j}|\} \cdot \|B\|_F \geq \|W \circ B\|_F \geq \min_{i,j \in [d]} \{|W_{i,j}|\} \cdot \|B\|_F.$$

Proof. The proof directly follows from the definition of the Frobenius norm. ☐

F.5 FINAL BOUND

We introduce some useful lemmas that we use to prove the final bound.

Lemma F.6. *If the following conditions hold*

- *Let* $b \in \mathbb{R}^n$ *and* $\langle b, \mathbf{1}_n \rangle = 0$.
- *Let* $f \in [\delta, 1]^n$ *and* $\langle f, \mathbf{1}_n \rangle = 1$.

Then we have

$$\|(b - \langle b, f \rangle \mathbf{1}_n) \circ f\|_2 \geq \delta \|b\|_2.$$

Proof. Note that $\langle b, \mathbf{1}_n \rangle = 0$ so that b and $\mathbf{1}_n$ are orthogonal with each other. Then, we have

$$\begin{aligned} \|(b - \langle b, f \rangle \mathbf{1}_n) \circ f\|_2 \geq &~ \delta \|b - \langle b, f \rangle \mathbf{1}_n\|_2 \\ = &~ \delta \sqrt{\|b\|_2^2 + \|\langle b, f \rangle \mathbf{1}_n\|_2^2} \\ \geq &~ \delta \|b\|_2, \end{aligned}$$

where the second step is from Pythagorean theorem. ☐

We present our final bound for proving the PL inequality.

Lemma F.7 (Formal version of Lemma 5.7)**.** *If the following conditions hold*

- *Let* $B \in \mathbb{R}^{n \times n}$ *and each row summation is zero, i.e.,* $B \cdot \mathbf{1}_n = \mathbf{0}_n$.
- *Let* $\widetilde{f}(M) \in [0,1]^{n \times n}$ *and each row summation is 1, i.e.,* $\widetilde{f}(M) \cdot \mathbf{1}_n = \mathbf{1}_n$.
- *Assume that* $\min_{i,j \in [n]} \widetilde{f}(M)_{i,j} \geq \delta > 0$.

Then, we can show

$$\|B \circ \widetilde{f}(M) - \mathrm{diag}((B \circ \widetilde{f}(M)) \cdot \mathbf{1}_n) \widetilde{f}(M)\|_F \geq \delta \cdot \|B\|_F$$

Proof. For any $i \in [n]$, let $B_i \in \mathbb{R}^n$ be the i-th row of B, and we have $\langle B_i, \mathbf{1}_n \rangle = 0$ by the first condition.

For any $i \in [n]$, let $\widetilde{f}(M)_i \in \mathbb{R}^n$ be the i-th row of $\widetilde{f}(M)$, and we have $\langle \widetilde{f}(M)_i, \mathbf{1}_n \rangle = 1$ by the second condition and $\widetilde{f}(M)_{i,j} \in [\delta, 1]$ by the third condition.

By Lemma F.6, for any $i \in [n]$, we have

$$\|(B_i - \langle B_i, \widetilde{f}(M)_i \rangle \mathbf{1}_n) \circ \widetilde{f}(M)_i\|_2 \geq \delta \|B_i\|_2.$$

Then, we have

$$\begin{aligned} & \|B \circ \widetilde{f}(M) - \mathrm{diag}((B \circ \widetilde{f}(M)) \cdot \mathbf{1}_n) \widetilde{f}(M)\|_F^2 \\ = & \sum_{i \in [n]} \|(B_i - \langle B_i, \widetilde{f}(M)_i \rangle \mathbf{1}_n) \circ \widetilde{f}(M)_i\|_2^2 \\ \geq & \sum_{i \in [n]} \delta^2 \|B_i\|_2^2 \\ = & \delta^2 \|B\|_F^2. \end{aligned}$$

□

F.6 PL INEQUALITY

Here we present the bound for one unit loss function.

Lemma F.8. *If the following conditions hold*

- *Let* $c(M)$ *be defined in Definition B.4.*

We have

$$\|c(M)\|_F \leq 2\sqrt{n}.$$

Proof. We have

$$\begin{aligned} \|c(M)\|_F & \leq \|\widetilde{f}(M)\|_F + \|f\|_F \\ & \leq 2\sqrt{n}, \end{aligned}$$

where the first step follows from Definition B.4 and triangle inequality, the second step follows $x_1^2 + \cdots + x_n^2 \leq (x_1 + \cdots + x_n)^2$ when $x_i \geq 0$ for any $i \in [n]$. □

We present the lemma for proving the PL inequality.

Lemma F.9. *If the following conditions hold*

- *Let* $\widetilde{f}(M)$ *be defined in Definition B.3.*
- *Let* $c(M)$ *be defined in Definition B.4.*

We have

$$\| \operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)\|_F \leq \sqrt{n}.$$

Proof. We have

$$\begin{aligned}
\| \operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)\|_F \leq & \max_{i \in n}\{|(c(M)_i \circ \widetilde{f}(M)_i) \cdot \mathbf{1}_n|\} \cdot \|\widetilde{f}(M)\|_F \\
\leq & \|\widetilde{f}(M)\|_F \\
\leq & \sqrt{n},
\end{aligned}$$

where the first step is by Frobenius norm definition and the second step follows from $\langle \widetilde{f}(M)_i, \mathbf{1}_n \rangle = 1$ and $c(M)_i \in [-1, 1]^n$ for any $i \in [n]$. ☐

Finally, we can show the lemma for PL inequality.

Lemma F.10 (PL inequality, formal version of 5.5)**.** *If the following conditions hold,*

- *Let* $M \in [0, 1]^{d \times d}$.
- *Let* $\lambda \in [0, 1]$ *be some constant.*
- *Assume that* $XX^\top \succeq \beta I$.
- *Assume that* $\min_{i,j \in [n]} \widetilde{f}(M)_{i,j} \geq \delta > 0$.
- *Let* $\mathcal{L}(M)$ *be defined in Definition B.7.*

Furthermore,

- *Let* $\alpha = 2$.
- *Let* $\mu = 2 \min_{i,j \in [d]}\{|W_{i,j}|\} \cdot \beta \cdot \delta$.
- *Let* $\xi = 12\sqrt{n} \max_{i,j \in [d]}\{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot \lambda d / \mu$.

We have

$$\|\nabla_M \mathcal{L}(M)\|_F^\alpha \geq \frac{1}{2}\mu(\|c(M)\|_F^2 + \frac{2\lambda^2}{\mu}\|M\|_F^2 - \xi).$$

Proof. We have $\widetilde{f}(M) \cdot \mathbf{1}_n = \mathbf{1}_n$ and $f \cdot \mathbf{1}_n = \mathbf{1}_n$ by Definition B.3. Note that $c(M) = \widetilde{f}(M) - f$ by Definition B.4. Thus, we have $c(M) \cdot \mathbf{1}_n = \mathbf{0}_n$.

On the other hand, we have

$$\begin{aligned}
& \|W \circ (X^\top(c(M) \circ \widetilde{f}(M) - \operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M))X)\|_F \\
\leq & \max_{i,j \in [d]}\{|W_{i,j}|\} \cdot \|X^\top(c(M) \circ \widetilde{f}(M) - \operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M))X\|_F \\
\leq & \max_{i,j \in [d]}\{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot \|c(M) \circ \widetilde{f}(M) - \operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)\|_F \\
\leq & \max_{i,j \in [d]}\{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot (\|c(M) \circ \widetilde{f}(M)\|_F + \| \operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)\|_F) \\
\leq & \max_{i,j \in [d]}\{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot (\|c(M)\|_F + \| \operatorname{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)\|_F) \\
\leq & \max_{i,j \in [d]}\{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot (2\sqrt{n} + \sqrt{n}) \\
= & \max_{i,j \in [d]}\{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot 3\sqrt{n}
\end{aligned}$$

where the first and forth steps follow Lemma F.5, the second step follows from Frobenius norm property, the third step follows from triangle inequality, the fifth step follows from Lemma F.8 and Lemma F.9.

Let $\alpha = 2$. We have the following

$$\begin{aligned}
&\|\nabla_M \mathcal{L}(M)\|_F^2 \\
= &\|W \circ (X^\top (c(M) \circ \widetilde{f}(M) - \mathrm{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M))X) + \lambda M\|_F^2 \\
\geq &\|W \circ (X^\top (c(M) \circ \widetilde{f}(M) - \mathrm{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M))X)\|_F^2 + \lambda^2 \|M\|_F^2 - \alpha_1 \\
\geq &\alpha_2 \cdot \|X^\top (c(M) \circ \widetilde{f}(M) - \mathrm{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M))X\|_F^2 + \lambda^2 \|M\|_F^2 - \alpha_1 \\
\geq &\alpha_2 \cdot \alpha_3 \cdot \|c(M) \circ \widetilde{f}(M) - \mathrm{diag}((c(M) \circ \widetilde{f}(M)) \cdot \mathbf{1}_n)\widetilde{f}(M)\|_F^2 + \lambda^2 \|M\|_F^2 - \alpha_1 \\
\geq &\alpha_2 \cdot \alpha_3 \cdot \alpha_4 \cdot \|c(M)\|_F^2 + \lambda^2 \|M\|_F^2 - \alpha_1 \\
= &\frac{1}{2}\mu(\|c(M)\|_F^2 + \frac{2\lambda^2}{\mu}\|M\|_F^2 - \xi),
\end{aligned}$$

where the second step follows from Lemma F.2 and $\alpha_1 = 6\sqrt{n}\max_{i,j \in [d]}\{|W_{i,j}|\} \cdot \|X\|_F^2 \cdot \lambda d$, the third step follows from Lemma F.5 and $\alpha_2 = \min_{i,j \in [d]}\{|W_{i,j}|\}$, the fourth step follows from Lemma F.4 and $\alpha_3 = \beta$, the fifth step follows from Lemma F.7 and $\alpha_4 = \delta$, and the last step follows from $\mu = 2\alpha_2 \cdot \alpha_3 \cdot \alpha_4$ and $\xi = 2\alpha_1/\mu$.

□