ATTENTION SCHEME INSPIRED SOFTMAX REGRESSION

Zhihang Li * **Zhizhou Sha**† **Zhao Song**‡ **Mingda Wan**§

ABSTRACT

In this work, we introduce ATTREG (Attention-Inspired Softmax Regression), a novel theoretical framework designed to advance the understanding of attention mechanisms within large language models (LLMs). In the area of convex optimization such as using the central path method to solve linear programming, the softmax function has been used as a crucial tool for controlling the progress and stability of potential functions [Cohen, Lee and Song STOC 2019, Brand SODA 2020]. By redefining the softmax regression problem through an attention-inspired approach, we establish a regularized variant, RATTREG (Regularized Attention-Inspired Softmax Regression), which incorporates an exponential activation function tailored for enhanced convergence and efficiency. Our comprehensive analysis encompasses the formulation of new problem definitions, the derivation of first and second-order derivatives to understand gradient dynamics, and a theoretical investigation into the convergence properties of the proposed models. We also develop an efficient computational approach using an adapted Newton method, supported by a sparsification technique, to address the challenges of high dimensionality and data sparsity inherent in LLMs. The implications of this study are significant, offering deeper insights into the operational dynamics of attention mechanisms and opening new avenues for optimizing the training processes of advanced neural network architectures. In a certain sense, our provable convergence result provides theoretical support for why we can use greedy algorithms to train the softmax function in practice.

1 INTRODUCTION

In the past few years, Large Language Models (LLMs) have experienced explosive development. There is a series of results of LLMs, like Transformer Vaswani et al. (2017), GPT-1 Radford et al. (2018), BERT Devlin et al. (2018), GPT-2 Radford et al. (2019), GPT-3 Brown et al. (2020), PaLM Chowdhery et al. (2022), OPT Zhang et al. (2022). The success of a recent chatbot named ChatGPT ChatGPT (2022) by OpenAI has exemplified the use of LLMs in human-interaction tasks. Very recently, OpenAI released their new version of LLM, named GPT-4 OpenAI (2023), which has been tested to perform much better even than previous ChatGPT Bubeck et al. (2023). These LLMs are trained on massive amounts of textual data to generate natural language text. They have already shown their power on various real-work tasks, including natural language translation He et al. (2021), sentiment analysis Usama et al. (2020), language modeling Martin et al. (2019), and even creative writing ChatGPT (2022); OpenAI (2023).

In the development of LLMs, the computation of *attention* plays a crucial role by significantly improving the model's capability to concentrate on pertinent sections of the input text, as highlighted in multiple foundational studies Vaswani et al. (2017); Radford et al. (2018); Devlin et al. (2018); Radford et al. (2019); Brown et al. (2020). Typically, the attention computation is defined to be $\mathsf{Att}(Q, K, V) := D^{-1}AV$, where $A := \exp(QK^\top) \in \mathbb{R}^{n \times n}$ is a square matrix and $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n \times n}$ is a diagonal matrix. Matrix $Q \in \mathbb{R}^{n \times d}$ represents the query tokens, which are typically derived from the previous hidden state of the decoder. And we use matrix

* lizhihangdll@gmail.com. Huazhong Agricultural University.
† shazz20@mails.tsinghua.edu.cn. Tsinghua University.
‡ magic.linuxkde@gmail.com. The Simons Institute for the Theory of Computing at UC Berkeley.
§ dylan.r.mathison@gmail.com. Anhui University.

$K \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{n \times d}$ to denote the key tokens and values. We also compute A, each entry of A is computed as a dot product between the query vector q and the key vector k_i, and the softmax function is applied to obtain the attention weights $A_{i,j}$. The attention mechanism operates by leveraging the correlations between words or tokens within the text, thereby allowing the model to dynamically adjust its focus during the processing of information. This process is not only vital for understanding context but also for making nuanced interpretations of the text. Despite its widespread use and clear benefits, the underlying theoretical principles of how attention works remain somewhat elusive and not fully understood.

There remains a substantial gap in the literature concerning a comprehensive theoretical framework that explains the inner workings and efficacy of attention mechanisms in LLMs. Given the current state of research, a natural question arises: *How in the theory is the Attention module being trained?* To address this question, we delve into the inner workings of attention modules and, inspired by their mechanisms, we propose a new regression problem, ATTREG (Attention Regression). This problem is designed to provide a theoretical understanding of the attention mechanism's convergence capabilities. Additionally, we incorporate a regularization term into the model formulation to further refine our approach. To facilitate practical implementation, we also develop and outline an algorithm specifically tailored to solve this enhanced regression problem, ensuring it effectively captures the dynamics of attention in LLMs.

We summarize our contributions as follows: (1) We introduce ATTREG (Attention-Inspired Softmax Regression Problem, Definition 2.6), a novel concept aimed at exploring the convergence capabilities of attention mechanisms. This theoretical framework is designed to enhance our understanding of the operational dynamics in modern language models. Additionally, we integrate a regularization component to formulate a regularized version, RATTREG (Attention-Inspired Regularized Softmax Regression Problem, Definition 2.7). Compared to previous work Li et al. (2023b), we take a step forward to understanding and explanation of the Attention theory by considering the softmax operation. (2) In our thorough analysis of the proposed models, we provide extensive details on the mathematical underpinnings of RATTREG. This includes a complete derivation of the first-order derivatives, which help in understanding the gradient dynamics, and the second-order derivatives, which are crucial for assessing the curvature of the optimization landscape. These calculations are elaborated in Section 5. Our analytical approach not only clarifies the theoretical structure of the models but also lays the groundwork for more efficient computational strategies. (3) Leveraging these derivatives and the Hessian matrices, and incorporating existing optimization techniques from the literature Deng et al. (2022); Song et al. (2022), we develop an adapted Newton method enhanced with a sparsification tool to efficiently solve RATTREG (Theorem 3.1). The efficiency of our method is significantly influenced by the sparsity of our input matrices, which aligns well with the inherently sparse nature of attention mechanism's weight matrices.

Roadmap. In Section 2 we state the setup of the problem we study. In Section 3 we provide our main result. In Section 4, we present a technical overview of our work. In Section 5, we provide the main results for the Hessian analysis. In Section 6, we introduce the approximate Newton method we use. In Section 7, we restate the formal version of our results. We give a conclusion in Section 8.

2 PRELIMINARY

This section introduces the foundational definitions and optimization problems that serve as the backbone for our theoretical and algorithmic developments. Section 2.1 provides a detailed exposition of key mathematical constructs, including the softmax function, its associated loss function, and auxiliary normalized quantities. These elements are central to the problem formulations and analytical framework. Section 2.2 formalizes the two primary optimization problems explored in this work: the attention-inspired softmax regression problem (ATTREG) and its regularized counterpart (RATTREG). Together, these definitions set the stage for the results and methods presented in the following sections.

2.1 KEY CONCEPTS

We define function softmax f as follows

Definition 2.1 (Function f)**.** *Given a matrix $A \in \mathbb{R}^{n \times d}$. Let $\mathbf{1}_n$ denote a length-n vector that all entries are ones. We define prediction function $f : \mathbb{R}^d \to \mathbb{R}^n$ as $f(x) := \langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \cdot \exp(Ax)$.*

Definition 2.2 (Loss function $L_{\exp}$)**.** *Given a matrix $A \in \mathbb{R}^{n \times d}$ and a vector $b \in \mathbb{R}^n$. We define loss function $L_{\exp} : \mathbb{R}^d \to \mathbb{R}$ as $L_{\exp}(x) := 0.5 \cdot \| \langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \exp(Ax) - b \|_2^2$.*

For convenient, we define two helpful notations α and c

Definition 2.3 (Normalized coefficients)**.** *We define $\alpha : \mathbb{R}^d \to \mathbb{R}$ as $\alpha(x) := \langle \exp(Ax), \mathbf{1}_n \rangle$. Then, we can rewrite $f(x)$ (see Definition 2.1) and $L_{\exp}(x)$ (see Definition 2.2) as follows*

- $f(x) = \alpha(x)^{-1} \cdot \exp(Ax)$.
- $L_{\exp}(x) = 0.5 \cdot \| \alpha(x)^{-1} \cdot \exp(Ax) - b \|_2^2$.
- $L_{\exp}(x) = 0.5 \cdot \| f(x) - b \|_2^2$.

Definition 2.4. *We define function $c : \mathbb{R}^d \in \mathbb{R}^n$ as $c(x) := f(x) - b$.. Then we can rewrite $L_{\exp}(x)$ (see Definition 2.2) as*

$$L_{\exp}(x) = 0.5 \cdot \| c(x) \|_2^2.$$

Definition 2.5 (Informal version of Definition B.8)**.** *Given matrix $A \in \mathbb{R}^{n \times d}$. For a given vector $w \in \mathbb{R}^n$, let $W = \mathrm{diag}(w)$. We define $L_{\mathrm{reg}} : \mathbb{R}^d \to \mathbb{R}$ as*

$$L_{\mathrm{reg}}(x) := 0.5 \| WAx \|_2^2$$

2.2 Problem Definition

Here we provide the definition of AttReg and RAttReg.

Definition 2.6 (AttReg, Attention-Inspired Softmax Regression Problem)**.** *Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, the softmax regression problem is aiming for minimize the following objective function*

$$\min_{x \in \mathbb{R}^d} \| \langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \exp(Ax) - b \|_2^2.$$

It is natural in practice to consider regularization Li et al. (2023a), then we propose the regularized version of softmax regression.

Definition 2.7 (RAttReg, Attention-Inspired Regularized Softmax Regression Problem)**.** *Given $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and $w \in \mathbb{R}^n$, the goal of the regularized softmax regression is to solve the following minimization problem,*

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} \cdot \| \langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \exp(Ax) - b \|_2^2 + \frac{1}{2} \cdot \| WAx \|_2^2.$$

3 Main Result

We now present our main result. The following theorem proves that RAttReg can be solved in $\widetilde{O}(\mathrm{nnz}(A) + d^\omega)$ time with high probability, implying that the running time is very low when the matrix A is sparse. We note that since $\langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \exp(Ax)$ is always a probability distribution, it is natural to assume that each entry of b is nonnegative and its ℓ_1 norm is at most 1.

Theorem 3.1 (Main Result, informal version of Theorem 7.1)**.** *Under mild assumptions, the* RAttReg *(Definition 2.7) can be solved with high precision (the output solution is close the the optimal solution) by an algorithm (Algorithm 1) in time $\widetilde{O}(\mathrm{nnz}(A) + d^\omega)$, with high probability.*

We note that in previous work Li et al. (2023b), the assumption $\|b\|_2 \leq R$ was made. This is because their setting does not incorporate the normalization parameter. Assuming $\|b\|_1 \leq 1$ would be unjustified in their context, as they are not attempting to learn the distribution.

4 TECHNICAL OVERVIEW

This section provides a concise summary of our methods and theoretical analysis. Section 4.1 presents the decomposition of the Hessian matrix for softmax regression and introduces a structured approach to simplify its computation. Section 4.2 analyzes the positive definiteness of the Hessian by bounding its components through low-rank and diagonal approximations. Section 4.3 establishes the Lipschitz continuity of the Hessian, leveraging key decompositions and bounding techniques. Finally, Section 4.4 outlines an efficient implementation of the Newton method, utilizing sparsification techniques to approximate the Hessian in near input-sparsity time, which significantly accelerates the optimization process.

4.1 DECOMPOSITION OF HESSIAN FOR SOFTMAX REGRESSION

Recall the target function of our problem is in the form of

$$\min_{x \in \mathbb{R}^d} 0.5 \cdot \|f(x) - b\|_2^2 + 0.5 \cdot \|WAx\|_2^2,$$

We divide the loss function with respect to above target function to the following two terms $L(x) := L_{\exp}(x) + L_{\mathrm{reg}}(x)$, where $L_{\exp}(x) := 0.5 \cdot \|\langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \cdot \exp(Ax) - b\|_2^2$ and $L_{\mathrm{reg}}(x) := 0.5 \cdot \|WAx\|_2^2$. Calculating the Hessian of $L_{\exp}(x)$ directly is too complicated. To simplify this, we define two terms of $\alpha(x) := \langle \exp(Ax), \mathbf{1}_n \rangle$, $f(x) := \langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \cdot \exp(Ax)$. Then, in order to get the final Hessian to the loss functions, we calculate the Hessian step by step. To be specific, we divide the Hessian calculation into the following items: (1) Hessian of $\exp(Ax)$; (2) Hessian of $\alpha(x)$ and $\alpha^{-1}(x)$; (3) Hessian of $f(x)$. After that, we notice a structured decomposition of Hessian of $L(x)$. We show that

$$\begin{aligned} \frac{\mathrm{d}^2 L_{\exp}}{\mathrm{d} x_i^2} &= A_{*,i}^\top B(x) A_{*,i} \\ \text{and} \quad \frac{\mathrm{d}^2 L_{\exp}}{\mathrm{d} x_i \mathrm{d} x_j} &= A_{*,i}^\top B(x) A_{*,j}, \end{aligned} \tag{1}$$

where $B(x)$ is only function of x and has no relation with respect to i and j. In order to apply existing sparsification tool to boost the Hessian calculation (which is one of our main motivations), we construct specific decomposition to the two terms of $B(x)$. We show that, B can be viewed as the sum of several rank-1 matrices and diagonal matrices.

4.2 HESSIAN IS POSITIVE DEFINITE

The key insight of this section lies in the analysis of volumetric barrier functions for solving semidefinite programming Anstreicher (2000); Huang et al. (2022). With the decomposition of the Hessian matrix for $L_{\exp}$, the next step is to bound it. To be specific, by dividing $B(x)$ in the way of low-rank parts and diagonal parts, we can lower and upper bound each segment of them. And by combining them, we can get the bound for $B(x)$,

$$-4I_n \preceq B(x) \preceq 8I_n.$$

Now combine the Hessian for $L_{\exp}$ and $L_{\mathrm{reg}}(x)$ (Hessian for $L_{\mathrm{reg}}(x)$ is trivial $A^\top W^2 A$) we get

$$\frac{\mathrm{d}^2 L}{\mathrm{d} x^2} = A^\top (B(x) + W^2) A.$$

We show that, by assuming all w_i^2's are lower bounded by $100 + l/\sigma_{\min}(A)^2$, the Hessian is positive definite $\frac{\mathrm{d}^2 L}{\mathrm{d} x^2} \succeq l \cdot I_d$. Further more, we show if all w_i^2's are lower bounded by $100 + l/\sigma_{\min}(A)^2$, then the matrix W^2 can approximate the sum of $B(x) + W^2$ with a constant guarantee, i.e.,

$$1 - 1/10) \cdot (B(x) + W^2) \preceq W^2 \preceq (1 + 1/10) \cdot (B(x) + W^2).$$

This allows us to apply sparsification tool on W to approximate the Hessian.

4.3 LIPSCHITZ PROPERTY FOR HESSIAN

The key insight of this section lies in the analysis of previous analysis for recurrent neural networks Allen-Zhu et al. (2019a;b). By the above calculation of Hessian, we divide the Hessian matrix to different segments. Now with the decomposition (to be specific, we divide the Hessian into low-rank parts and diagonal parts), we show Lipschitz property for each term. We first show Lipschitz property for the basic terms:

- $\|\exp(Ax)\|_2 \leq \sqrt{n} \cdot \exp(R^2)$
- $\|\exp(Ax) - \exp(Ay)\|_2 \leq R\exp(R^2) \cdot \|x - y\|_2$;
- $|\alpha(x) - \alpha(y)| \leq \sqrt{n} \cdot \|\exp(Ax) - \exp(Ay)\|_2$;
- $|\alpha(x)^{-1} - \alpha(y)^{-1}| \leq \beta^{-2} \cdot |\alpha(x) - \alpha(y)|$; (Later we will also prove an upper bound for β^{-1}, see Lemma F.9)
- $\|f(x) - f(y)\|_2 \leq R_f \cdot \|x - y\|_2$. (Here R_f is a function of $\beta^{-1}, \exp(R^2)$, see concrete definition in Lemma E.2)

Then, following the decomposition of the Hessian matrix, we show the Lipschitz property for each of the divided terms (we use G_i for $i \in 1, \ldots, 8$ to denote the terms) and combine them together to get the property of

$$\|G_1\| + \sum_{i=1}^{8} \|G_i\| \leq 100R \cdot \|f(x) - f(y)\|_2.$$

With this property and a fact that $\|\frac{\mathrm{d}^2 L}{\mathrm{d}x^2}(x) - \frac{\mathrm{d}^2 L}{\mathrm{d}x^2}(y)\| \leq \|A\| \cdot (\|G_1\| + \sum_{i=1}^{8} \|G_i\|) \cdot \|A\|$, by assuming any two points x, y satisfy $\|x\|_2, \|y\|_2 \leq R$ and $\|A(x - y)\|_\infty < 0.01$, we can show that the Hessian matrix is Lipschitz, i.e.,

$$\|\frac{\mathrm{d}^2 L}{\mathrm{d}x^2}(x) - \frac{\mathrm{d}^2 L}{\mathrm{d}x^2}(y)\| \leq \beta^{-2} n \exp(20R^2) \cdot \|x - y\|_2,$$

for some small constant $\beta \in (0, 0.1)$, which implies the Lipschitz property for the Hessian.

4.4 APPROXIMATED NEWTON METHOD WITH SPARSIFICATION TOOL

Newton method is a widely-used and traditional tool used in optimization questions. For a target function $L(x)$, one can compute its gradient $g(x) : \mathbb{R}^d \to \mathbb{R}^d$ and Hessian matrix $H(x) : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ as

$$g(x) := \nabla L(x), H(x) := \nabla^2 L(x)$$

and use them to update the data point as follows

$$x_{t+1} \leftarrow x_t - H(x_t)^{-1} \cdot g(x_t).$$

But in many optimization applications, computing $\nabla^2 L(x_t)$ or $(\nabla^2 L(x_t))^{-1}$ is quite expensive. Therefore, a natural motivation is to approximately formulate its Hessian or inverse of Hessian. In our setting, we want a faster implementation of Newton method. By above steps, we show our Hessian can be approximated by a matrix in the form of $A^\top DA$, where $D = W^2$ is a diagonal matrix. This inspires us to implement a standard tool Deng et al. (2022); Song et al. (2022) that can generate a sparse matrix $\widetilde{D}$ such that

$$(1 - \epsilon) \cdot A^\top DA \preceq A^\top \widetilde{D} A \preceq (1 + \epsilon) \cdot A^\top DA$$

in near input-sparsity time of A. By this tool, we can reduce the time for Hessian calculation of each iteration to the time of $\widetilde{O}(\mathrm{nnz}(A) + d^\omega)$. Here $\mathrm{nnz}(A)$ denotes the number of non-zero entries in matrix A. Let ω denote the exponent of matrix multiplication. Currently, $\omega \approx 2.373$ Williams (2012); Le Gall (2014); Alman & Williams (2021).

Algorithm 1 Informal version of Algorithm 2

procedure ITERATIVESOFTMAXREGRESSION($A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n, w \in \mathbb{R}^n, \epsilon, \delta$) ▷ Theorem 7.1
  We choose x_0 (suppose it satisfies Definition F.1)
  We use $T \leftarrow \log(\|x_0 - x^*\|_2/\epsilon)$ to denote the number of iterations.
  for $t = 0 \to T$ **do**
    $D \leftarrow B_{\mathrm{diag}}(x_t) + \mathrm{diag}(w \circ w)$
    $\widetilde{D} \leftarrow \textsc{SubSample}(D, A, \epsilon_1 = \Theta(1), \delta_1 = \delta/T)$ ▷ Lemma F.5
    Compute gradient
    $\widetilde{H} \leftarrow A^\top \widetilde{D} A$
    $x_{t+1} \leftarrow x_t + \widetilde{H}^{-1} g$
  end for
  $\widetilde{x} \leftarrow x_{T+1}$
  return $\widetilde{x}$
end procedure

5 ANALYSIS OF HESSIAN

In this section we discover two key properties of L that enables us to use computational efficient algorithm to tackle the softmax regression problem. Specifically, in Section 5.1 we simplify $\nabla^2 L$ and decompose it into the sum of several diagonal matrices and low rank matrices. In Section 5.2 we prove that L is convex. In Section 5.3 we prove that $\nabla^2 L$ is lipschitz. With these two key properties, we can use the approximate newton method to solve the softmax regression problem efficiently.

5.1 SPLITTING THE HESSIAN

In this section, we simplify $\nabla^2 L$ and decompose it into the sum of several diagonal matrices and low rank matrices. As described in Eq. (1), we decompose the Hessian into the specific norm with $B(x)$. Now in the following lemma, we provide the result that, $B(x)$ can be decomposed into summation of low-rank matrices and diagonal matrices. By doing so, we simplify the analysis afterwards. The formal version of this Lemma with detailed analysis can be found in Section C.

Lemma 5.1 (Decomposition of $B(x)$, informal version of Lemma C.15)**.** *Let $B(x) = B_1(x) + B_2(x)$. where $B_1(x) \in \mathbb{R}^{n \times n}$ is defined as follows:*

$$B_1(x) = \underbrace{\langle f(x), f(x) \rangle}_{\text{scalar}} \cdot \underbrace{f(x)}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} + \underbrace{\mathrm{diag}(f(x) \circ f(x))}_{n \times n \text{ diagonal matrix}} + \underbrace{(f(x) \circ f(x))}_{n \times 1} \cdot \underbrace{f(x)^\top}_{1 \times n} + \underbrace{(f(x) \circ f(x))}_{n \times 1} \cdot \underbrace{f(x)^\top}_{1 \times n}$$

and $B_2(x) \in \mathbb{R}^{n \times n}$ is defined as follows:

$$B_2(x) = \underbrace{2\langle c(x), f(x) \rangle}_{\text{scalar}} \cdot \underbrace{f(x)}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} + \underbrace{\langle c(x), f(x) \rangle}_{\text{scalar}} \cdot \underbrace{\mathrm{diag}(f(x))}_{n \times n \text{ diagonal matrix}} + \underbrace{\mathrm{diag}(c(x) \circ f(x))}_{n \times n \text{ diagonal matrix}} - \underbrace{(c(x) \circ f(x))}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} - \underbrace{f(x)}_{n \times 1} \underbrace{(f(x) \circ c(x))^\top}_{1 \times n}$$

Finally, we can show that $B(x) \in \mathbb{R}^{n \times n}$ satisfies that

$$B(x) = \underbrace{\langle 3f(x) - 2b, f(x) \rangle}_{\text{scalar}} \cdot \underbrace{f(x)}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} + \underbrace{\langle f(x) - b, f(x) \rangle}_{\text{scalar}} \cdot \underbrace{\mathrm{diag}(f(x))}_{n \times n \text{ diagonal matrix}} + \underbrace{\mathrm{diag}((2f(x) - b) \circ f(x))}_{n \times n \text{ diagonal matrix}} + \underbrace{(b \circ f(x))}_{n \times 1} \cdot \underbrace{f(x)^\top}_{1 \times n} + \underbrace{f(x)}_{n \times 1} \cdot \underbrace{(b \circ f(x))^\top}_{1 \times n}$$

In summary, $B_1(x) \in \mathbb{R}^{n \times n}$ is constructed by three rank-1 matrices and one diagonal matrix; $B_2(x) \in \mathbb{R}^{n \times n}$ is constructed by three rank-1 matrices and two diagonal matrices; $B(x) \in \mathbb{R}^{n \times n}$ is constructed by three rank-1 matrices and two diagonal matrices.

5.2 Hessian is Positive Semidefinite

In this section, we obtained the positive lower bound of $\nabla^2 L$ and thus proved that L is a convex function, which is one property required to use approximate newton method. We also find the upper and lower bound of W^2. The formal version of this lemma with detailed analysis can be found in Section D. The idea is decomposing the Hessian into two terms of $B(x)$ and W, and provide analysis respectively to show their sum is positive definite.

Lemma 5.2 (Informal version of Lemma D.3)**.** *Let $l > 0$ denote a scalar. If all $i \in [n]$, $w_i^2 \geq 4 + l/\sigma_{\min}(A)^2$, then it holds that*

$$\frac{\mathrm{d}^2 L}{\mathrm{d}x^2} \succeq l \cdot I_d.$$

5.3 Hessian is Lipschitz

In this section, we proved that $\nabla^2 L$ is Lipschitz by finding the upper bound of $\|\nabla^2 L(x) - \nabla^2 L(y)\|$, which is another property requires by the approximate newton method. Lemma 5.3 states the main result of this subsection, and we provide a detailed version with proof in Lemma E.1.

Lemma 5.3 (Informal version of Lemma E.1)**.** *Let $R > 2$ be a constant, we show that under certain conditions, it holds that*

$$\|H(x) - H(y)\| \leq \beta^{-2} n \exp(20R^2) \cdot \|x - y\|_2.$$

6 Approximate Newton Method

In this section, we provide an approximate version of the newton method for convex optimization. Traditional Newton methods utilize the exact Hessian matrix to update the target variable, i.e., for each step, we use the following equation to update: $x_{t+1} = x_t - H(x_t)^{-1} \cdot g(x_t)$. While in many real-world tasks, it is very hard and expensive to compute exact $\nabla^2 L(x_t)$ or $(\nabla^2 L(x_t))^{-1}$. Thus, it is natural to consider the approximated computation of the gradient and Hessian. We define the approximate Hessain computation as

Definition 6.1 (Approximate Hessian)**.** *For any Hessian $H(x_t) \in \mathbb{R}^{d \times d}$, we define the approximated Hessian $\widetilde{H}(x_t) \in \mathbb{R}^{d \times d}$ to be a matrix such that the following holds,*

$$(1 - \epsilon_0) \cdot H(x_t) \preceq \widetilde{H}(x_t) \preceq (1 + \epsilon_0) \cdot H(x_t).$$

In order to get the approximated Hessian $\widetilde{H}(x_t)$ efficiently, here we state a standard tool (see Lemma 4.5 in Deng et al. (2022)).

Lemma 6.2 (Deng et al. (2022); Song et al. (2022))**.** *Let $\epsilon_0 = 0.01$ be a constant precision parameter. Let $A \in \mathbb{R}^{n \times d}$ be a real matrix, then for any positive diagonal (PD) matrix $D \in \mathbb{R}^{n \times n}$, there exists an algorithm which runs in time $O((\mathrm{nnz}(A) + d^\omega) \operatorname{poly}(\log(n/\delta)))$ and it outputs an $O(d \log(n/\delta))$ sparse diagonal matrix $\widetilde{D} \in \mathbb{R}^{n \times n}$ for which*

$$(1 - \epsilon_0) A^\top D A \preceq A^\top \widetilde{D} A \preceq (1 + \epsilon_0) A^\top D A.$$

Note that, ω denotes the exponent of matrix multiplication, currently $\omega \approx 2.373$ Williams (2012); Le Gall (2014); Alman & Williams (2021).

Following the standard of Approximate Newton Hessian literature Anstreicher (2000); Jiang et al. (2020a); Brand et al. (2021); Song et al. (2021); Huang et al. (2022); Li et al. (2023b), we consider the following "Approximate update" process, i.e. $x_{t+1} = x_t - \widetilde{H}(x_t)^{-1} \cdot g(x_t)$. Combining this step with the previous analysis on the Hessian matrix, we can get the guarantee of our main algorithm. For the full detail and proof of the main theorem, please refer to Appendix 7.

7 Formal Result

In this section, we present our main result formally in Theorem 7.1 and our efficient algorithm in Algorithm 2.

Algorithm 2 Here, we present our main algorithm informally. Formal version of Algorithm 1

procedure ITERATIVESOFTMAXREGRESSION($A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n, w \in \mathbb{R}^n, \epsilon, \delta$) ▷ Theorem 7.1
  We choose x_0 (suppose it satisfies Definition F.1)
  We use $T \leftarrow \log(\|x_0 - x^*\|_2/\epsilon)$ to denote the number of iterations.
  for $t = 0 \to T$ **do**
    $D \leftarrow B_{\mathrm{diag}}(x_t) + \mathrm{diag}(w \circ w)$
    $\widetilde{D} \leftarrow \textsc{SubSample}(D, A, \epsilon_1 = \Theta(1), \delta_1 = \delta/T)$ ▷ Lemma F.5
    $g \leftarrow A^\top (f(x_t)\langle c(x_t), f(x_t)\rangle + \mathrm{diag}(f(x_t))c(x_t))$
    $\widetilde{H} \leftarrow A^\top \widetilde{D} A$
    $x_{t+1} \leftarrow x_t + \widetilde{H}^{-1} g$
  end for
  $\widetilde{x} \leftarrow x_{T+1}$
  return $\widetilde{x}$
end procedure

Theorem 7.1. *Suppose we have matrix $A \in \mathbb{R}^{n \times d}$, and vectors $b, w \in \mathbb{R}^n$.*

Let $f(x) := \langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \exp(Ax)$. Let x^* be the optimal solution of $\min_{x \in \mathbb{R}^d} 0.5\|f(x) - b\|_2^2 + 0.5\|\mathrm{diag}(w)Ax\|_2^2$, where $g(x^*) = \mathbf{0}_d$. and $\|x^*\|_2 \leq R$. Let $R \geq 10$ be a positive scalar. Suppose we have $\|A\| \leq R$ Suppose it holds that $b \geq \mathbf{0}_n$, and $\|b\|_1 \leq 1$. Suppose it holds that $w_i^2 \geq 100 + l/\sigma_{\min}(A)^2$ for all $i \in [n]$ Suppose it holds that $M = n^{1.5}\exp(30R^2)$. Let x_0 denote an initial point for which it holds that $M\|x_0 - x^*\|_2 \leq 0.1l$.

Then for any accuracy parameter $\epsilon \in (0, 0.1)$ and failure probability $\delta \in (0, 0.1)$, there exists a randomized algorithm (Algorithm 2) such that, with probability at least $1 - \delta$, it runs $T = \log(\|x_0 - x^*\|_2/\epsilon)$ iterations and outputs a vector $\widetilde{x} \in \mathbb{R}^d$ such that $\|\widetilde{x} - x^*\|_2 \leq \epsilon$, and the time cost per iteration is $O((\mathrm{nnz}(A) + d^\omega) \cdot \mathrm{poly}(\log(n/\delta))$. Here ω denotes the exponent of matrix multiplication. Currently $\omega \approx 2.373$ Williams (2012); Le Gall (2014); Alman & Williams (2021).

Proof. It follows from combining Lemma D.3, Lemma F.8, Lemma F.5, Lemma E.1 and Lemma F.7.

The proof of Upper bound on M follows from Lemma F.10; the proof of Hessian is PD follows from Lemma D.3; the proof of Hessian is Lipschitz follows from Lemma E.1; the proof of Cost per iteration follows from Lemma F.5; the proof of Convergence per Iteration follows from Lemma F.7, where we have $\|x_k - x^*\|_2 \leq 0.4 \cdot \|x_{k-1} - x^*\|_2$. For the proof of the Number of Iterations, we can show that after T iterations, we have $\|x_T - x^*\|_2 \leq 0.4^T \cdot \|x_0 - x^*\|_2$ By choice of T, we get the desired bound. The failure probability is following from union bound over T iterations.

□

8 CONCLUSION

This study delves into the intricacies of the softmax regression problem, drawing inspiration from the attention paradigm prevalent in LLMs. We specifically focus on the attention mechanism utilized in these models and redefine the softmax regression issue incorporating an exponential activation function. This choice is motivated by the computational processes underpinning the attention mechanisms in LLMs. Our exploration into this area not only sheds light on the operational dynamics of attention units but also paves the way for more nuanced understandings and applications in the realm of language models. Building on this foundation, we introduce a regularized variant of the softmax regression problem, tailored to enhance its applicability and efficiency. This regularization is a critical step in refining the problem to better suit real-world scenarios where data sparsity and computational efficiency are key concerns. Alongside this, we have developed and propose an algorithm capable of solving this regularized problem in input-sparsity time. In summary, our work makes a notable contribution to the fields of natural language processing and optimization. By offering a novel perspective on the mechanisms driving LLMs and presenting a fast, efficient method for solving the adapted softmax regression problem, we open up new avenues for application across a broad spectrum of NLP challenges.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *Advances in neural information processing systems*, 32, 2019b.

Josh Alman and Zhao Song. Fast attention requires bounded entries. *arXiv preprint arXiv:2302.13214*, 2023.

Josh Alman and Virginia Vassilevska Williams. A refined laser method and faster matrix multiplication. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 522–539. SIAM, 2021.

Josh Alman, Jiehao Liang, Zhao Song, Ruizhe Zhang, and Danyang Zhuo. Bypass exponential time preprocessing: Fast neural network training via weight-data correlation preprocessing. *arXiv preprint arXiv:2211.14227*, 2022.

Kurt M Anstreicher. The volumetric barrier for semidefinite programming. *Mathematics of Operations Research*, 2000.

Jan van den Brand. A deterministic linear program solver in current matrix multiplication time. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 259–278. SIAM, 2020.

Jan van den Brand, Binghui Peng, Zhao Song, and Omri Weinstein. Training (overparametrized) neural networks in near-linear time. In *ITCS*, 2021.

Jan van den Brand, Zhao Song, and Tianyi Zhou. Algorithm and hardness for dynamic attention maintenance in large language models. *arXiv preprint arXiv:2304.02207*, 2023.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

ChatGPT. Optimizing language models for dialogue. *OpenAI Blog*, November 2022. URL https://openai.com/blog/chatgpt/.

Beidi Chen, Zichang Liu, Binghui Peng, Zhaozhuo Xu, Jonathan Lingjie Li, Tri Dao, Zhao Song, Anshumali Shrivastava, and Christopher Re. Mongoose: A learnable lsh framework for efficient neural network training. In *International Conference on Learning Representations*, 2021.

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

Michael B Cohen, Ben Cousins, Yin Tat Lee, and Xin Yang. A near-optimal algorithm for approximating the john ellipsoid. In *Conference on Learning Theory*, pp. 849–873. PMLR, 2019a.

Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. In *STOC*, 2019b.

Yichuan Deng, Zhao Song, and Omri Weinstein. Discrepancy minimization in input-sparsity time. *arXiv preprint arXiv:2210.12468*, 2022.

Yichuan Deng, Zhihang Li, and Zhao Song. An improved sample complexity for rank-1 matrix sensing. *arXiv preprint arXiv:2303.06895*, 2023a.

Yichuan Deng, Sridhar Mahadevan, and Zhao Song. Randomized and deterministic attention sparsification algorithms for over-parameterized feature dimension. *arxiv preprint: arxiv 2304.03426*, 2023b.

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

Yeqi Gao, Sridhar Mahadevan, and Zhao Song. An over-parameterized exponential regression. *arXiv preprint arXiv:2303.16504*, 2023.

Yuzhou Gu and Zhao Song. A faster small treewidth sdp solver. *arXiv preprint arXiv:2211.06033*, 2022.

Yuzhou Gu, Zhao Song, Junze Yin, and Lichen Zhang. Low rank matrix completion via robust alternating minimization in nearly linear time. *arXiv preprint arXiv:2302.11068*, 2023.

Weihua He, Yongyun Wu, and Xiaohua Li. Attention mechanism for neural machine translation: A survey. In *2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 5, pp. 1485–1489. IEEE, 2021.

Baihe Huang, Shunhua Jiang, Zhao Song, Runzhou Tao, and Ruizhe Zhang. Solving sdp faster: A robust ipm framework and efficient implementation. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 233–244. IEEE, 2022.

Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song. A faster interior point method for semidefinite programming. In *2020 IEEE 61st annual symposium on foundations of computer science (FOCS)*, pp. 910–918. IEEE, 2020a.

Haotian Jiang, Yin Tat Lee, Zhao Song, and Sam Chiu-wai Wong. An improved cutting plane method for convex optimization, convex-concave games and its applications. In *STOC*, 2020b.

Haotian Jiang, Yin Tat Lee, Zhao Song, and Lichen Zhang. Convex minimization with integer minima in $\widetilde{O}(n^4)$ time. *arXiv preprint arXiv:2304.03426*, 2023.

Shunhua Jiang, Zhao Song, Omri Weinstein, and Hengjie Zhang. Faster dynamic matrix inverse for faster lps. In *STOC*, 2021.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

François Le Gall. Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th international symposium on symbolic and algebraic computation*, pp. 296–303, 2014.

Yin Tat Lee, Zhao Song, and Qiuyi Zhang. Solving empirical risk minimization in the current matrix multiplication time. In *Conference on Learning Theory (COLT)*, pp. 2140–2157. PMLR, 2019.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. *arXiv preprint arXiv:2303.04245*, 2023a.

Zhihang Li, Zhao Song, and Tianyi Zhou. Solving regularized exp, cosh and sinh regression problems. *arXiv preprint, 2303.15725*, 2023b.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suarez, Yoann Dupont, Laurent Romary, Eric Villemonte de La Clergerie, Djame Seddah, and Benoit Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Lianke Qin, Zhao Song, Lichen Zhang, and Danyang Zhuo. An online and unified algorithm for projection matrix vector multiplication with application to empirical risk minimization. In *AISTATS*, 2023a.

Lianke Qin, Zhao Song, and Ruizhe Zhang. A general algorithm for solving rank-one matrix sensing. *arXiv preprint arXiv:2303.12298*, 2023b.

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

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

Charlie Snell, Ruiqi Zhong, Dan Klein, and Jacob Steinhardt. Approximating how single head attention learns. *arXiv preprint arXiv:2103.07601*, 2021.

Zhao Song and Zheng Yu. Oblivious sketching-based central path method for linear programming. In *International Conference on Machine Learning*, pp. 9835–9847. PMLR, 2021.

Zhao Song, Lichen Zhang, and Ruizhe Zhang. Training multi-layer over-parametrized neural network in subquadratic time. *arXiv preprint arXiv:2112.07628*, 2021.

Zhao Song, Xin Yang, Yuanyuan Yang, and Tianyi Zhou. Faster algorithm for structured john ellipsoid computation. *arXiv preprint arXiv:2211.14407*, 2022.

Mohd Usama, Belal Ahmad, Enmin Song, M Shamim Hossain, Mubarak Alrashoud, and Ghulam Muhammad. Attention-based sentiment analysis using convolutional and recurrent neural network. *Future Generation Computer Systems*, 113:571–578, 2020.

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.

Nikhil Vyas, Sham Kakade, and Boaz Barak. Provable copyright protection for generative models. *arXiv preprint arXiv:2302.10870*, 2023.

Virginia Vassilevska Williams. Multiplying matrices faster than coppersmith-winograd. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 887–898, 2012.

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

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Lichen Zhang. Speeding up optimizations via data structures: Faster search, sample and maintenance. Master's thesis, Carnegie Mellon University, 2022.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Appendix

Roadmap. In Section A, we introduce the related works of our paper. In Section B we define the notations used in our work and provide some useful tools for exact algebra, approximate algebra and differential computation. In Section C we provide detailed analysis of $L_{\exp}$, including its gradient and hessian. In Section D we proved that $L = L_{\exp} + L_{\reg}$ is a convex function. In Section E we proved that the hessian of $L_{\exp}$ is Lipschitz. In Section F we provide an approximate version of newton method for solving convex optimization problem which is more efficient under certain assumptions. In Section 7 we state our result of this paper and provide the algorithm for tackling the softmax regression problem in an formal way.

A Related Work

Computation. Since the explosion of LLM, there have been a lot of theoretical works about the computation of attention Kitaev et al. (2020); Chen et al. (2021); Zandieh et al. (2023); Alman & Song (2023); Brand et al. (2023); Li et al. (2023b); Deng et al. (2023b). Locality sensitive hashing (LSH) techniques have been employed in research to approximate attention. Kitaev et al. (2020); Chen et al. (2021); Zandieh et al. (2023). Based on it, Zandieh et al. (2023) proposed KDEformer, an efficient approximation algorithm for the dot-product attention mechanism, with provable spectral norm bounds and superior performance on various pre-trained models. Recent research has investigated both static and dynamic approaches to attention computation Alman & Song (2023); Brand et al. (2023). Additionally, Li et al. (2023b) delved into regularized hyperbolic regression problems involving exponential, cosh, and sinh functions. Deng et al. (2023b) proposed randomized and deterministic algorithms to sparsify the attention matrix in large language models, achieving high accuracy with significantly reduced feature dimension.

Convergence and Optimization. There have been works trying to understanding attention computation on optimization and convergence perspective Zhang et al. (2020); Snell et al. (2021); Gao et al. (2023); Li et al. (2023b;a). In practical attention models, adaptive methods often performs better than SGD. To understand this, Zhang et al. (2020) showed that heavy-tailed distribution of the noise is one of the reason of the bad performance of SGD compared to adaptive methods, and provided new upper and lower bounds for convergence of adaptive methods under heavy-tailed noise in attention models. This answered the question of why adaptive methods performs better in attention models. Snell et al. (2021) explained why models sometimes attend to salient words and how the attention mechanism evolves throughout training, using a model property they defined, named Knowledge to Translate Individual Words (KTIW), which is learned early on from word co-occurrence statistics and later used to attend to input words while predicting the output. Recently, Gao et al. (2023) studied the regression problem inspired by the neural network with exponential activation function, and showed the convergence of a two-layer NN with large width (over-parameterized), while Li et al. (2023b) focused on solving regularized exp, cosh and sinh regression problems inspired by Attention computation. Li et al. (2023a) explored how transformers learn the co-occurrence structure of words by examining attention-based network size, depth, and complexity through experiments and mathematical analysis, showing that the embedding and self-attention layers encode topical structure with higher average inner product and pairwise attention between same-topic words.

Privacy and Security. With the fast development of LLMs, the potential negative impact of abusing LLM has also been considered. To overcome this, without influencing the quality of the generated text, Kirchenbauer et al. (2023) proposed a novel method to add watermark in LLM-generated text. The method needs no access to the parameters or API of the LLM. Vyas et al. (2023) introduced a formal definition of near access-freeness (NAF) and develops generative model learning algorithms to ensure that the model outputs do not resemble copyrighted data by more than k-bits, with experiments on language (transformers) and image (diffusion) generative models demonstrating strong protection against sampling protected content.

Applications of Exponential Functions There are many theory problems use $\exp, \sinh, \cosh$ function as potential functions to prove the convergence of iterative optimization algorithms. In

the works of Cohen et al. (2019b); Brand (2020); Jiang et al. (2021), they use $\cosh$ function to define a potential function for measuring the central path. Such design can guarantee the central path method is robust and stable. Let $x \in \mathbb{R}^n$ denote the primal variables and let s denote the slack variables of the central path algorithm. The central path is defined as tuple (x, y, s, t) that satisfies

$$\begin{aligned} Ax &= b, x > 0 \\ A^\top y + s &= c, s > 0 \\ x_i s_i &= t \text{ for all } i \in [n]. \end{aligned}$$

Let t denote the target at one step of central (also mathematically called the complementary gap). The xs can viewed as real circumstances. In the ideal case, they hope $xs = t$. However, this is unlikely to happen. They use the potential function $\Phi(xs) = \sum_{i=1}^n \cosh(x_i s_i - t)$ to measure the difference between reality and target.

In Qin et al. (2023b), they use $\cosh$ function to build a potential for rank-1 matrix sensing problem. Given a matrix $A \in \mathbb{R}^{d \times n}$, there are n observations x_i, y_i and $b_i = x_i^\top A y_i$. The goal of matrix sensing is to recover A by using observations $\{(x_i, y_i, b_i)\}_{i \in [n]}$. They use the potential function $\Phi(x, y) = \sum_{i=1}^n \cosh(x_i^\top A y_i - b_i)$.

In standard linear ℓ_2 regression, given matrix $A \in \mathbb{R}^{n \times d}$ and vector $b \in \mathbb{R}^n$, the formulation is usually $L(x) = \|Ax - b\|_2^2 = (\sum_{i=1}^n (Ax)_i - b_i)^2$. In Li et al. (2023b), they use $\cosh$ function to construct a ℓ_2 loss such that $L(x) = \sum_{i=1}^n (\cosh((Ax)_i) - b_i)^2$. Furthermore, Li et al. (2023b) also studied $\exp$ and $\sinh$ functions.

Sketching for Convex Optimization. Sketching techniques has been widely-used in optimization problems such as integral minimization problem Jiang et al. (2023), cutting plane method Jiang et al. (2020b), training over-parameterized neural tangent kernel regression Brand et al. (2021); Song et al. (2021); Zhang (2022); Alman et al. (2022), linear programming Cohen et al. (2019b); Jiang et al. (2021); Song & Yu (2021); Gu & Song (2022), empirical risk minimization Lee et al. (2019); Qin et al. (2023a), computing John Ellipsoid Cohen et al. (2019a); Song et al. (2022), matrix sensing Deng et al. (2023a), matrix completion Gu et al. (2023).

B PRELIMINARY

In this section, we provide the preliminaries used in our paper. In Section B.1 we introduce the notations we use. In Section B.2 we provide some basic facts for exact computation. In Section B.3 we provide some tools for finding the bound of norms based on vectors. In Section B.4 we provide some tools for finding the bound of norms related to matrices. In Section B.5, we provide basic inequalities for psd matrices. In Section B.6, we state several basic rules for calculus. In Section B.7 we provide the regularization term L_{reg} and compute ∇L_{reg} and $\nabla^2 L_{\mathrm{reg}}$.

B.1 NOTATIONS

We denote the ℓ_p norm of a vector x by $\|x\|_p$, 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 vector $x \in \mathbb{R}^n$, $\exp(x) \in \mathbb{R}^n$ denotes a vector where $\exp(x)_i$ is $\exp(x_i)$ for all $i \in [n]$. For $n > k$, for any matrix $A \in \mathbb{R}^{n \times k}$, we denote the spectral norm of A by $\|A\|$, i.e., $\|A\| := \sup_{x \in \mathbb{R}^k} \|Ax\|_2 / \|x\|_2$. We use $\sigma_{\min}(A)$ to denote the minimum singular value of A. Given two vectors $x, y \in \mathbb{R}^n$, we use $\langle x, y \rangle$ to denote $\sum_{i=1}^n x_i y_i$. Given two vectors $x, y \in \mathbb{R}^n$, we use $x \circ y$ to denote a vector that its i-th entry is $x_i y_i$ for all $i \in [n]$. We use $e_i \in \mathbb{R}^n$ to denote a vector where i-th entry is 1 and all the other entries are 0. Let $x \in \mathbb{R}^n$ be a vector. We define $\mathrm{diag}(x) \in \mathbb{R}^{n \times n}$ as the diagonal matrix whose diagonal entries are given by $\mathrm{diag}(x)_{i,i} = x_i$ for $i = 1, \ldots, n$, and all off-diagonal entries are zero. For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, we say $A \succ 0$ (positive definite (PD)), if for all $x \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$, we have $x^\top A x > 0$. For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, we say $A \succeq 0$ (positive semidefinite (PSD)), if for all $x \in \mathbb{R}^n$, we have $x^\top A x \geq 0$. The Taylor Series for $\exp(x)$ is $\exp(x) = \sum_{i=0}^\infty \frac{x^i}{i!}$. We use $b \geq \mathbf{0}_n$ to denote that $b_i \geq 0$ for all $i \in [n]$.

B.2 Basic Algebra

Fact B.1. *For vectors $u, v, w \in \mathbb{R}^n$. We have*

- $\langle u, v \rangle = \langle u \circ v, \mathbf{1}_n \rangle$
- $\langle u \circ v, w \rangle = \langle u \circ v \circ w, \mathbf{1}_n \rangle$
- $\langle u, v \rangle = \langle v, u \rangle$
- $\langle u, v \rangle = u^\top v = v^\top u$

Fact B.2. *For any vectors $u, v, w \in \mathbb{R}^n$, we have*

- $u \circ v = v \circ u = \mathrm{diag}(u) \cdot v = \mathrm{diag}(v) \cdot u$
- $u^\top (v \circ w) = u^\top \mathrm{diag}(v) w$
- $u^\top (v \circ w) = v^\top (u \circ w) = w^\top (u \circ v)$
- $u^\top \mathrm{diag}(v) w = v^\top \mathrm{diag}(u) w = u^\top \mathrm{diag}(w) v$
- $\mathrm{diag}(u) \cdot \mathrm{diag}(v) \cdot \mathbf{1}_n = \mathrm{diag}(u) v$
- $\mathrm{diag}(u \circ v) = \mathrm{diag}(u) \, \mathrm{diag}(v)$
- $\mathrm{diag}(u) + \mathrm{diag}(v) = \mathrm{diag}(u + v)$

B.3 Basic Vector Norm Bounds

Fact B.3. *For vectors $u, v \in \mathbb{R}^n$, we have*

- $\langle u, v \rangle \leq \|u\|_2 \cdot \|v\|_2$ *(Cauchy-Schwarz inequality)*
- $\| \mathrm{diag}(u) \| \leq \|u\|_\infty$
- $\|u \circ v\|_2 \leq \|u\|_\infty \cdot \|v\|_2$
- $\|u\|_\infty \leq \|u\|_2 \leq \sqrt{n} \cdot \|u\|_\infty$
- $\|u\|_2 \leq \|u\|_1 \leq \sqrt{n} \cdot \|u\|_2$
- $\| \exp(u) \|_\infty \leq \exp(\|u\|_\infty) \leq \exp(\|u\|_2)$
- *Let α be a scalar, then* $\|\alpha \cdot u\|_2 = |\alpha| \cdot \|u\|_2$
- $\|u + v\|_2 \leq \|u\|_2 + \|v\|_2$.
- *For any* $\|u - v\|_\infty \leq 0.01$*, we have* $\| \exp(u) - \exp(v) \|_2 \leq \| \exp(u) \|_2 \cdot 2 \|u - v\|_\infty$
- *For any $u, v \in \mathbb{R}^d$ such that $\|u\|_2, \|v\|_2 \leq R$, we have* $\| \exp(u) - \exp(v) \| \leq \exp(R) \|u - v\|_2$

Proof. For all the other facts we omit the details. We will only prove the last fact.

We have

$$\begin{aligned} \| \exp(u) - \exp(v) \|_2 &= \| \exp(u) \circ (\mathbf{1}_n - \exp(v - u)) \|_2 \\ &\leq \| \exp(u) \|_2 \cdot \| \mathbf{1}_n - \exp(v - u) \|_\infty \\ &\leq \| \exp(u) \|_2 \cdot 2 \|u - v\|_\infty, \end{aligned}$$

where the 1st step follows from definition of $\circ$ operation and $\exp()$, the 2nd step follows from Fact B.3, the 3rd step follows from $|\exp(x) - 1| \leq 2x$ for all $x \in (0, 0.1)$.

□

B.4 Basic Matrix Norm Bounds

Fact B.4. *For matrices U, V, we have*

- $\|U^\top\| = \|U\|$
- $\|U\| \geq \|V\| - \|U - V\|$
- $\|U + V\| \leq \|U\| + \|V\|$
- $\|U \cdot V\| \leq \|U\| \cdot \|V\|$
- *If $U \preceq \alpha \cdot V$, then $\|U\| \leq \alpha \cdot \|V\|$*
- *For scalar $\alpha \in \mathbb{R}$, we have $\|\alpha \cdot U\| \leq |\alpha| \cdot \|U\|$*
- *For any vector v, we have $\|Uv\|_2 \leq \|U\| \cdot \|v\|_2$.*
- *Let $u, v \in \mathbb{R}^n$ denote two vectors, then we have $\|uv^\top\| \leq \|u\|_2 \|v\|_2$*

B.5 Basic PSD

Fact B.5. *Let $u, v \in \mathbb{R}^n$, We have:*

- $uu^\top \preceq \|u\|_2^2 \cdot I_n$.
- $\mathrm{diag}(u) \preceq \|u\|_2 \cdot I_n$
- $\mathrm{diag}(u \circ u) \preceq \|u\|_2^2 \cdot I_n$
- $uv^\top + vu^\top \preceq uu^\top + vv^\top$
- $uv^\top + vu^\top \succeq -(uu^\top + vv^\top)$
- $(v \circ u)(v \circ u)^\top \preceq \|v\|_\infty^2 uu^\top$

B.6 Basic Derivative Rules

Fact B.6. *Let f be a differentiable function.*

We have

- *Part 1.* $\frac{\mathrm{d}}{\mathrm{d}x} \exp(x) = \exp(x)$
- *Part 2. For any $j \neq i$,* $\frac{\mathrm{d}}{\mathrm{d}x_i} f(x_j) = 0$

Fact B.7 (Rules of differentiation)**.** *Let f denote a differentiable function.*

For all $n, i \in \mathbb{Z}_+$, we have

- *Sum rule 1.* $\frac{\mathrm{d}}{\mathrm{d}t} \sum_{l=1}^n f(x_l) = \sum_{l=1}^n \frac{\mathrm{d}}{\mathrm{d}t} f(x_i)$
- *Sum rule 2.* $\frac{\mathrm{d}}{\mathrm{d}x_i} \sum_{l=1}^n f(x_l) = \frac{\mathrm{d}}{\mathrm{d}x_i} f(x_i)$
- *Chain rule.* $\frac{\mathrm{d}}{\mathrm{d}x_i} f(g(x_i)) = f'(g(x_i)) \cdot g'(x_i)$
- *Difference rule.* $\frac{\mathrm{d}}{\mathrm{d}x_i}(f(x_i) - g(x_i)) = \frac{\mathrm{d}}{\mathrm{d}x_i} f(x_i) - \frac{\mathrm{d}}{\mathrm{d}x_i} g(x_i)$
- *Product rule.* $\frac{\mathrm{d}}{\mathrm{d}x_i}(f(x_i)g(x_i)) = f'(x_i)g(x_i) + f(x_i)g'(x_i)$
- *Constant multiple rule. For any $x \neq y$,* $\frac{\mathrm{d}}{\mathrm{d}x_i}(y_i \cdot f(x_i)) = y_i \cdot \frac{\mathrm{d}}{\mathrm{d}x_i} f(x_i)$

B.7 REGULARIZATION

Definition B.8 (Formal version of Definition 2.5)**.** *Given matrix $A \in \mathbb{R}^{n \times d}$. For a given vector $w \in \mathbb{R}^n$, let $W = \mathrm{diag}(w)$. We define $L_{\mathrm{reg}} : \mathbb{R}^d \to \mathbb{R}$ as follows*

$$L_{\mathrm{reg}}(x) := 0.5 \|WAx\|_2^2$$

Lemma B.9 (Folklore, see Li et al. (2023b) as an example)**.** *For a given vector $w \in \mathbb{R}^n$, let $W = \mathrm{diag}(w)$. Let $L_{\mathrm{reg}} : \mathbb{R}^d \to \mathbb{R}$ be defined as Definition B.8.*

Then, we have

- *The gradient is*

$$\frac{\mathrm{d}L_{\mathrm{reg}}}{\mathrm{d}x} = A^\top W^2 Ax$$

- *The Hessian is*

$$\frac{\mathrm{d}^2 L_{\mathrm{reg}}}{\mathrm{d}x^2} = A^\top W^2 A$$

C SOFTMAX REGRESSION LOSS

In this section, we provide detailed computation for $\nabla L_{\exp}$ and $\nabla^2 L_{\exp}$. In Section C.1, we define $f(x)$ and $\alpha(x)$ to simplify the computation for $\nabla L_{\exp}$ and $\nabla^2 L_{\exp}$. In Section C.2, we compute $\nabla L_{\exp}$ step by step. In Section C.3, we define the gradient of Loss function and also prove the Lipschitz property for gradient. In Section C.4-C.8, we compute $\nabla^2 L_{\exp}$ step by step. To be specific, in Section C.4, we compute $\nabla^2 \exp(Ax)$; in Section C.5, we compute $\nabla^2 \alpha(x)$; in Section C.6, we compute $\nabla^2 \alpha(x)^{-1}$; in Section C.7, we compute $\nabla^2 f(x)$; in Section C.8, we compute $\nabla^2 L_{\exp}$. In Section C.9, we provide some result to aid the computation in Section C.10. In Section C.10, we split $\nabla^2 L_{\exp}$ into several low rank matrices and diagonal matrices.

C.1 DEFINITIONS

We define function softmax f as follows

Definition C.1 (Function f)**.** *Given a matrix $A \in \mathbb{R}^{n \times d}$. Let $\mathbf{1}_n$ denote a length-n vector that all entries are ones. We define prediction function $f : \mathbb{R}^d \to \mathbb{R}^n$ as follows*

$$f(x) := \langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \cdot \exp(Ax).$$

Then we have

Lemma C.2. *Let $f : \mathbb{R}^d \to \mathbb{R}^n$ be defined as Definition C.1, then we have for all $x \in \mathbb{R}^d$,*

- $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$.
- $0 \preceq f(x) f(x)^\top \preceq I_n$.
- *For any vector b,* $0 \preceq (b \circ f(x))(b \circ f(x))^\top \preceq \|b\|_\infty^2 f(x) f(x)^\top \preceq \|b\|_\infty^2 I_n$
- *For any vector b,* $\mathrm{diag}(b \circ b) \preceq \|b\|_\infty^2 I_n$
- $0 \preceq \mathrm{diag}(f(x)) \preceq \|f(x)\|_\infty I_n \preceq \|f(x)\|_2 I_n$.
- $0 \preceq \mathrm{diag}(f(x) \circ f(x)) \preceq \|f(x)\|_\infty^2 I_n \preceq \|f(x)\|_2 I_n$.

Proof. The proofs are very straightforward, so we omitted the details here. □

Definition C.3 (Loss function $L_{\exp}$)**.** *Given a matrix $A \in \mathbb{R}^{n \times d}$ and a vector $b \in \mathbb{R}^n$. We define loss function $L_{\exp} : \mathbb{R}^d \to \mathbb{R}$ as follows*

$$L_{\exp}(x) := 0.5 \cdot \|\langle \exp(Ax), \mathbf{1}_n \rangle^{-1} \exp(Ax) - b\|_2^2.$$

For convenient, we define two helpful notations α and c

Definition C.4 (Normalized coefficients)**.** *We define $\alpha : \mathbb{R}^d \to \mathbb{R}$ as follows*

$$\alpha(x) := \langle \exp(Ax), \mathbf{1}_n \rangle.$$

Then, we can rewrite $f(x)$ (see Definition C.1) and $L_{\exp}(x)$ (see Definition C.3) as follows

- $f(x) = \alpha(x)^{-1} \cdot \exp(Ax)$.
- $L_{\exp}(x) = 0.5 \cdot \|\alpha(x)^{-1} \cdot \exp(Ax) - b\|_2^2$.
- $L_{\exp}(x) = 0.5 \cdot \|f(x) - b\|_2^2$.

Definition C.5. *We define function $c : \mathbb{R}^d \in \mathbb{R}^n$ as follows*

$$c(x) := f(x) - b.$$

Then we can rewrite $L_{\exp}(x)$ (see Definition C.3) as follows

- $L_{\exp}(x) = 0.5 \cdot \|c(x)\|_2^2$.

C.2 Gradient

Lemma C.6 (Gradient)**.** *If the following conditions hold*

- *Given matrix $A \in \mathbb{R}^{n \times d}$ and a vector $b \in \mathbb{R}^n$.*
- *Let $\alpha(x)$ be defined in Definition C.4.*
- *Let $f(x)$ be defined in Definition C.1.*
- *Let $c(x)$ be defined in Definition C.5.*
- *Let $L_{\exp}(x)$ be defined in Definition C.3.*

For each $i \in [d]$, we have

- *Part 1.*
$$\frac{\mathrm{d} \exp(Ax)}{\mathrm{d} x_i} = \exp(Ax) \circ A_{*,i}$$
- *Part 2.*
$$\frac{\mathrm{d} \langle \exp(Ax), \mathbf{1}_n \rangle}{\mathrm{d} x_i} = \langle \exp(Ax), A_{*,i} \rangle$$
- *Part 3.*
$$\frac{\mathrm{d} \alpha(x)^{-1}}{\mathrm{d} x_i} = -\alpha(x)^{-1} \cdot \langle f(x), A_{*,i} \rangle$$
- *Part 4.*
$$\frac{\mathrm{d} f(x)}{\mathrm{d} x_i} = \frac{\mathrm{d} c(x)}{\mathrm{d} x_i} = -\langle f(x), A_{*,i} \rangle \cdot f(x) + f(x) \circ A_{*,i}$$
- *Part 5.*
$$\frac{\mathrm{d} \langle f(x), A_{*,i} \rangle}{\mathrm{d} x_i} = -\langle f(x), A_{*,i} \rangle^2 + \langle f(x), A_{*,i} \circ A_{*,i} \rangle$$
- *Part 6. For each $j \neq i$*
$$\frac{\mathrm{d} \langle f(x), A_{*,i} \rangle}{\mathrm{d} x_j} = -\langle f(x), A_{*,i} \rangle \cdot \langle f(x), A_{*,j} \rangle + \langle f(x), A_{*,i} \circ A_{*,j} \rangle$$

- *Part 7.*

$$\frac{\mathrm{d}L_{\exp}(x)}{\mathrm{d}x_i} = A_{*,i}^\top \cdot (-f(x)(f(x)-b)^\top f(x) + \mathrm{diag}(f(x))(f(x)-b))$$

Proof. **Proof of Part 1.** For each $j \in [n]$, we have

$$\begin{aligned}\frac{\mathrm{d}(\exp(Ax))_j}{\mathrm{d}x_i} =& \exp(Ax)_j \cdot \frac{\mathrm{d}(Ax)_j}{\mathrm{d}x_i} \\ =& \exp(Ax)_j \cdot \frac{(A\mathrm{d}x)_j}{\mathrm{d}x_i} \\ =& \exp(Ax)_j \cdot A_{j,i}\end{aligned}$$

where the first step follows from simple algebra, the second step follows from Fact B.6, the third step follows from simple algebra.

Thus, we have

$$\frac{\mathrm{d}\exp(Ax)}{\mathrm{d}x_i} = \exp(Ax) \circ A_{*,i}$$

Proof of Part 2. It trivially follows from arguments in Part 1.

Proof of Part 3.

$$\begin{aligned}\frac{\mathrm{d}\alpha(x)^{-1}}{\mathrm{d}x_i} =& \frac{\mathrm{d}\langle \exp(Ax), \mathbf{1}_n\rangle^{-1}}{\mathrm{d}x_i} \\ =& -1 \cdot \langle \exp(Ax), \mathbf{1}_n\rangle^{-1-1} \cdot \frac{\mathrm{d}}{\mathrm{d}x_i}(\langle \exp(Ax), \mathbf{1}_n\rangle) \\ =& -\langle \exp(Ax), \mathbf{1}_n\rangle^{-2} \langle \exp(Ax), A_{*,i}\rangle \\ =& -\alpha(x)^{-1} \langle f(x), A_{*,i}\rangle\end{aligned}$$

where the first step follows from $\frac{\mathrm{d}y^z}{\mathrm{d}x} = z \cdot y^{z-1} \frac{\mathrm{d}y}{\mathrm{d}x}$, the second step follows from results in **Part 2**, the third step follows from simple algebra, the last step follows from the definition of α and f.

Proof of Part 4.

$$\begin{aligned}\frac{\mathrm{d}f(x)}{\mathrm{d}x_i} =& \frac{\mathrm{d}\langle \exp(Ax), \mathbf{1}_n\rangle^{-1} \exp(Ax)}{\mathrm{d}x_i} \\ =& \exp(Ax) \cdot \frac{\mathrm{d}}{\mathrm{d}x_i}(\langle \exp(Ax), \mathbf{1}_n\rangle^{-1}) + \langle \exp(Ax), \mathbf{1}_n\rangle^{-1} \cdot \frac{\mathrm{d}}{\mathrm{d}x_i} \exp(Ax) \\ =& -\langle \exp(Ax), \mathbf{1}_n\rangle^{-2} \cdot \langle \exp(Ax), A_{*,i}\rangle \cdot \exp(Ax) \\ & + \langle \exp(Ax), \mathbf{1}_n\rangle^{-1} \cdot \exp(Ax) \circ A_{*,i} \\ =& -\langle f(x), A_{*,i}\rangle \cdot f(x) + f(x) \circ A_{*,i}\end{aligned}$$

where the first step follows from Definition of f, the second step follows from differential chain rule, the third step follows from the result from **Part 2** and **Part 3**, the forth step follows from definition of f (see Definition C.1).

Proof of Part 5

$$\begin{aligned}\frac{\mathrm{d}\langle f(x), A_{*,i}\rangle}{\mathrm{d}x_i} =& A_{*,i}^\top \frac{\mathrm{d}f(x)}{\mathrm{d}x_i} \\ =& A_{*,i}^\top(-\langle f(x), A_{*,i}\rangle \cdot f(x) + f(x) \circ A_{*,i}) \\ =& -\langle f(x), A_{*,i}\rangle \cdot A_{*,i}^\top f(x) + A_{*,i}^\top f(x) \circ A_{*,i} \\ =& -\langle f(x), A_{*,i}\rangle^2 + \langle f(x), A_{*,i} \circ A_{*,i}\rangle\end{aligned}$$

where the first step follows from extracting $A_{*,i}$ and Fact B.1 , the second step follows from result of **Part 4**, the third step follows from simple algebra, the last step follows from Fact B.1 .

Proof of Part 6.

$$\begin{aligned}\frac{\mathrm{d}\langle f(x), A_{*,i}\rangle}{\mathrm{d}x_j} &= A_{*,i}^\top \frac{\mathrm{d}f(x)}{\mathrm{d}x_j} \\ &= A_{*,i}^\top(-\langle f(x), A_{*,j}\rangle \cdot f(x) + f(x) \circ A_{*,j}) \\ &= -\langle f(x), A_{*,j}\rangle \cdot A_{*,i}^\top f(x) + A_{*,i}^\top f(x) \circ A_{*,j} \\ &= -\langle f(x), A_{*,j}\rangle \langle f(x), A_{*,i}\rangle + \langle A_{*,i}, f(x) \circ A_{*,j}\rangle \\ &= -\langle f(x), A_{*,i}\rangle \cdot \langle f(x), A_{*,j}\rangle + \langle f(x), A_{*,i} \circ A_{*,j}\rangle\end{aligned}$$

where the 1st step follows from extracting $A_{*,i}$ and $\langle a, b\rangle = a^\top b = b^\top a$, the 2nd step follows from result of **Part 4**, the 3rd step follows from simple algebra, the 4th step follows from $a^\top b = \langle a, b\rangle = \langle b, a\rangle$, the last step follows from Fact B.1.

Proof of Part 7.

$$\begin{aligned}\frac{\mathrm{d}L_{\exp}(x)}{\mathrm{d}x_i} &= \frac{\mathrm{d}}{\mathrm{d}x_i}(0.5 \cdot \|f(x) - b\|_2^2) \\ &= (f(x) - b)^\top \frac{\mathrm{d}}{\mathrm{d}x_i}(f(x) - b) \\ &= (f(x) - b)^\top(-\langle f(x), A_{*,i}\rangle \cdot f(x) + f(x) \circ A_{*,i}) \\ &= -A_{*,i}^\top f(x)(f(x) - b)^\top f(x) + (f(x) - b)^\top f(x) \circ A_{*,i} \\ &= -A_{*,i}^\top f(x)(f(x) - b)^\top f(x) + A_{*,i}^\top f(x) \circ (f(x) - b) \\ &= A_{*,i}^\top(-f(x)(f(x) - b)^\top f(x) + \mathrm{diag}(f(x))(f(x) - b))\end{aligned}$$

where the 1st step follows from the definition of f, the 2nd step follows from $\frac{\mathrm{d}\|y\|_2^2}{\mathrm{d}x} = 2y^\top \frac{\mathrm{d}y}{\mathrm{d}x}$, the 3rd step follows from the result in **Part 4**, the forth step follows from $\langle a, b\rangle = a^\top b$, the 5th step step follows from Fact B.2, the last step follows from extracting $A_{*,i}$ and Fact B.2.

□

C.3 Definition of Gradient

In this section, we use $g(x)$ to denote the gradient of $L_{\exp}(x)$.

Definition C.7. *If the following conditions hold*

- *Let $L_{\exp}(x)$ be defined as Definition C.3.*
- *Let $c(x)$ be defined as Definition C.5.*
- *Let $f(x)$ be defined as Definition C.1.*

We define $g(x) \in \mathbb{R}^d$ as follows

$$g(x) := \underbrace{A^\top}_{d \times n} \cdot \Big(- \underbrace{f(x)}_{n \times 1} \underbrace{\langle c(x), f(x)\rangle}_{\mathrm{scalar}} + \underbrace{\mathrm{diag}(f(x))}_{n \times n} \underbrace{c(x)}_{n \times 1} \Big)$$

Equivalently, for each $i \in [d]$, we define

$$g(x)_i := - \underbrace{\langle A_{*,i}, f(x)\rangle}_{\mathrm{scalar}} \cdot \underbrace{\langle c(x), f(x)\rangle}_{\mathrm{scalar}} + \underbrace{\langle A_{*,i}, f(x) \circ c(x)\rangle}_{\mathrm{scalar}}.$$

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

- *Let $g_1 : \mathbb{R}^d \to \mathbb{R}^n$ be defined as $g_1(x) := -f(x)\langle c(x), f(x)\rangle$*
- *Let $g_2 : \mathbb{R}^d \to \mathbb{R}^n$ be defined as $g_2(x) := \mathrm{diag}(f(x))c(x)$*

- *Let R_f be parameter such that*
  - $\|f(x) - f(y)\|_2 \leq R_f \cdot \|x - y\|_2$
  - $\|c(x) - c(y)\|_2 \leq R_f \cdot \|x - y\|_2$
- *Let $R_\infty \in (0, 2]$ be parameter such that*

$$R_\infty := \max\{\|f(x)\|_2, \|f(y)\|_2, \|c(x)\|_2, \|c(y)\|_2\}$$

We can show

- *Part 1.*

$$\|g_1(x) - g_1(y)\|_2 \leq 3R_f R_\infty^2 \|x - y\|_2$$

- *Part 2.*

$$\|g_2(x) - g_2(y)\|_2 \leq 2R_f R_\infty \|x - y\|_2$$

- *Part 3.*

$$\|g_1(x) + g_2(x) - g_1(y) - g_2(y)\|_2 \leq 8R_f R_\infty \|x - y\|_2$$

- *Part 4.*

$$\|g(x) - g(y)\|_2 \leq 8 \cdot \|A\| \cdot R_f \cdot R_\infty \|x - y\|_2$$

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

$$\begin{aligned}
\|g_1(x) - g_1(y)\|_2 = & \|f(x)\langle c(x), f(x)\rangle - f(y)\langle c(y), f(y)\rangle\|_2 \\
= & \|f(x)\langle c(x), f(x)\rangle - f(y)\langle c(x), f(x)\rangle \\
& + f(y)\langle c(x), f(x)\rangle - f(y)\langle c(y), f(x)\rangle \\
& + f(y)\langle c(y), f(x)\rangle - f(y)\langle c(y), f(y)\rangle\|_2 \\
\leq & \|f(x)\langle c(x), f(x)\rangle - f(y)\langle c(x), f(x)\rangle\|_2 \\
& + \|f(y)\langle c(x), f(x)\rangle - f(y)\langle c(y), f(x)\rangle\|_2 \\
& + \|f(y)\langle c(y), f(x)\rangle - f(y)\langle c(y), f(y)\rangle\|_2
\end{aligned}$$

where the 1st step follows from the definition of g_1, the second step follows from simple algebra, the 3rd step follows from Fact B.3.

For the first term, we have

$$\begin{aligned}
\|f(x)\langle c(x), f(x)\rangle - f(y)\langle c(x), f(x)\rangle\|_2 \leq & \|f(x) - f(y)\|_2 \cdot |\langle c(x), f(x)\rangle| \\
\leq & \|f(x) - f(y)\|_2 \cdot \|c(x)\|_2 \cdot \|f(x)\|_2 \\
\leq & R_f \cdot \|x - y\|_2 \cdot \|c(x)\|_2 \cdot \|f(x)\|_2
\end{aligned}$$

where the 1st step follows from $\|\alpha a\|_2 \leq |\alpha| \|a\|_2$(Fact B.3), the 2nd step follows from $\langle a, b\rangle \leq \|a\|_2 \|b\|_2$(Fact B.3), the 3rd step follows from the definition of R_f.

For the second term, we have

$$\begin{aligned}
\|f(y)\langle c(x), f(x)\rangle - f(y)\langle c(y), f(x)\rangle\|_2 \leq & \|f(y)\|_2 \cdot |\langle c(x) - c(y), f(x)\rangle| \\
\leq & \|f(y)\|_2 \cdot \|c(x) - c(y)\|_2 \cdot \|f(x)\|_2 \\
\leq & \|f(y)\|_2 \cdot R_f \cdot \|x - y\|_2 \cdot \|f(x)\|_2
\end{aligned}$$

where the 1st step follows from $\|\alpha a\|_2 \leq |\alpha| \|a\|_2$(Fact B.3), the 2nd step follows from $\langle a, b\rangle \leq \|a\|_2 \|b\|_2$(Fact B.3), the 3rd step follows from the definition of R_f.

For the third term, we have

$$\begin{aligned}
\|f(y)\langle c(y), f(x)\rangle - f(y)\langle c(y), f(y)\rangle\|_2 \leq & \|f(y)\|_2 \cdot |\langle c(y), f(x) - f(y)\rangle| \\
\leq & \|f(y)\|_2 \cdot \|c(y)\|_2 \cdot \|f(x) - f(y)\|_2
\end{aligned}$$

$$\leq \|f(y)\|_2 \cdot \|c(y)\|_2 \cdot R_f \cdot \|x - y\|_2$$

the 1st step follows from Fact B.3, the 2nd step follows from $\langle a, b \rangle \leq \|a\|_2 \|b\|_2$(Fact B.3), the 3rd step follows from the definition of R_f.

Combining three terms together, we complete the proof.

Proof of Part 2.

We have

$$\begin{aligned} & \| \operatorname{diag}(f(x))c(x) - \operatorname{diag}(f(y))c(y)\|_2 \\ = & \| \operatorname{diag}(f(x))c(x) - \operatorname{diag}(f(x))c(y) + \operatorname{diag}(f(x))c(y) - \operatorname{diag}(f(y))c(y)\|_2 \\ \leq & \| \operatorname{diag}(f(x))c(x) - \operatorname{diag}(f(x))c(y)\|_2 + \| \operatorname{diag}(f(x))c(y) - \operatorname{diag}(f(y))c(y)\|_2 \end{aligned}$$

where the first step follows from simple algebra, the second step follows from Fact B.3.

For the first term, we have

$$\begin{aligned} \| \operatorname{diag}(f(x))c(x) - \operatorname{diag}(f(x))c(y)\|_2 = & \| \operatorname{diag}(f(x))(c(x) - c(y))\|_2 \\ \leq & \| \operatorname{diag}(f(x))\| \cdot \|c(x) - c(y)\|_2 \\ \leq & \|f(x)\|_\infty \cdot \|c(x) - c(y)\|_2 \\ \leq & \|f(x)\|_2 \cdot \|c(x) - c(y)\|_2 \\ \leq & \|f(x)\|_2 \cdot R_f \cdot \|x - y\|_2 \end{aligned}$$

where the 1st step follows from Fact B.4, the 2nd step follows from Fact B.3, the 3rd step follows from Fact B.3, the 4th step follows from the definition of R_f.

For the second term, we have

$$\begin{aligned} \| \operatorname{diag}(f(x))c(y) - \operatorname{diag}(f(y))c(y)\|_2 = & \|(\operatorname{diag}(f(x) - f(y)))c(y)\|_2 \\ \leq & \| \operatorname{diag}(f(x) - f(y))\| \|c(y)\|_2 \\ \leq & \|f(x) - f(y)\|_2 \cdot \|c(y)\|_2 \\ \leq & R_f \cdot \|x - y\|_2 \cdot \|c(y)\|_2 \end{aligned}$$

where the first step follows from Fact B.2, the second step follows from Fact B.4, the third step follows from Fact B.3, the last step follows from the definition of R_f.

Combining two terms together, then we complete the proof. □

Proof of Part 3.

It follows from combining **Part 1** and **Part 2**.

Proof of Part 4.

It follows from **Part 3**.

C.4 Hessian Calculations: Step 1, Hessian of $\exp(Ax)$

Lemma C.9 (Hessian of $\exp(Ax)$)**.** *If the following condition holds*

- *Given a matrix $A \in \mathbb{R}^{n \times d}$.*

Then, we have, for each $i \in [d]$

- *Part 1.*

$$\frac{\mathrm{d}^2 \exp(Ax)}{\mathrm{d} x_i^2} = A_{*,i} \circ \exp(Ax) \circ A_{*,i}$$

- *Part 2.*

$$\frac{\mathrm{d}^2 \exp(Ax)}{\mathrm{d} x_i \mathrm{d} x_j} = A_{*,j} \circ \exp(Ax) \circ A_{*,i}$$

Proof. **Proof of Part 1.**

$$\begin{aligned}\frac{\mathrm{d}^2(\exp(Ax))}{\mathrm{d}x_i^2} &= \frac{\mathrm{d}}{\mathrm{d}x_i}(\frac{\mathrm{d}(\exp(Ax))}{\mathrm{d}x_i}) \\ &= \frac{\mathrm{d}(\exp(Ax) \circ A_{*,i})}{\mathrm{d}x_i} \\ &= A_{*,i} \circ \frac{\mathrm{d}\exp(Ax)}{\mathrm{d}x_i} \\ &= A_{*,i} \circ \exp(Ax) \circ A_{*,i}\end{aligned}$$

where the 1st step is an expansion of the Hessian, the 2nd step follows from **Part 1** in Lemma C.6, the 3rd step extracts the matrix $A_{*,i}$ with constant entries out of the derivative, and the last step also follows from **Part 1** in Lemma C.6.

Proof of Part 2.

$$\begin{aligned}\frac{\mathrm{d}^2(\exp(Ax))}{\mathrm{d}x_i \mathrm{d}x_j} &= \frac{\mathrm{d}}{\mathrm{d}x_i}(\frac{\mathrm{d}}{\mathrm{d}x_j}(\exp(Ax))) \\ &= \frac{\mathrm{d}}{\mathrm{d}x_i}(\exp(Ax) \circ A_{*,j}) \\ &= A_{*,j} \circ \exp(Ax) \circ A_{*,i}\end{aligned}$$

where the 1st step is an expansion of the Hessian, the 2nd step follows from **Part 1** in Lemma C.6, the 3rd step follows extracting $A_{*,j}$ and **Part 1** in Lemma C.6.

□

C.5 Hessian Calculations: Step 2, Hessian of $\alpha(x)$

Lemma C.10. *If the following conditions hold*

- *Let $\alpha(x)$ be defined as Definition C.4.*

Then, we have

- *Part 1.*

$$\frac{\mathrm{d}^2 \alpha(x)}{\mathrm{d} x_i^2} = \langle \exp(Ax), A_{*,i} \circ A_{*,i} \rangle$$

- *Part 2.*

$$\frac{\mathrm{d}^2 \alpha(x)}{\mathrm{d} x_i \mathrm{d} x_j} = \langle \exp(Ax), A_{*,i} \circ A_{*,j} \rangle$$

Proof. **Proof of Part 1.**

$$\begin{aligned}
\frac{\mathrm{d}^2 \alpha(x)}{\mathrm{d} x_i^2} = & \frac{\mathrm{d}}{\mathrm{d} x_i} (\frac{\mathrm{d}}{\mathrm{d} x_i} \langle \exp(Ax), \mathbf{1}_n \rangle) \\
= & \frac{\mathrm{d}}{\mathrm{d} x_i} (\langle \exp(Ax) \circ A_{*,i}, \mathbf{1}_n \rangle) \\
= & \langle A_{*,i} \circ \exp(Ax) \circ A_{*,i}, \mathbf{1}_n \rangle \\
= & \langle \exp(Ax), A_{*,i} \circ A_{*,i} \rangle
\end{aligned}$$

, where the 1st step follows from the expansion of hessian, the 2nd step follows from **Part 3** of Lemma C.6, the 3rd step follows from simple algebra, and the last step follows from Fact B.1.

Proof of Part 2.

$$\begin{aligned}
\frac{\mathrm{d}^2 \alpha(x)}{\mathrm{d} x_i \mathrm{d} x_j} = & \frac{\mathrm{d}}{\mathrm{d} x_j} (\frac{\mathrm{d}}{\mathrm{d} x_i} \langle \exp(Ax), \mathbf{1}_n \rangle) \\
= & \frac{\mathrm{d}}{\mathrm{d} x_j} (\langle \exp(Ax) \circ A_{*,i}, \mathbf{1}_n \rangle) \\
= & \langle A_{*,j} \circ \exp(Ax) \circ A_{*,i}, \mathbf{1}_n \rangle \\
= & \langle \exp(Ax), A_{*,i} \circ A_{*,j} \rangle
\end{aligned}$$

where the 1st step follows from the expansion of hessian, the 2nd step follows from **Part 2** of Lemma C.6, the 3rd step follows from simple algebra, the last step follows from Fact B.1.

□

C.6 Hessian Calculations: Step 3, Hessian of $\alpha(x)^{-1}$

Lemma C.11 (Hessian of $\alpha(x)^{-1}$)**.** *If the following conditions hold*

- *Let $\alpha(x)$ be defined as Definition C.4*
- *Let $f(x)$ be defined in Definition C.1.*

We have

- *Part 1.*

$$\begin{aligned}
\frac{\mathrm{d}^2 \alpha(x)^{-1}}{\mathrm{d} x_i^2} = & 2\alpha(x)^{-1} \cdot \langle f(x), A_{*,i} \rangle^2 - \alpha(x)^{-1} \cdot \langle f(x), A_{*,i} \circ A_{*,i} \rangle \\
= & 2\alpha(x)^{-1} A_{*,i}^\top f(x) f(x)^\top A_{*,i} - A_{*,i}^\top \operatorname{diag}(f(x)) A_{*,i}
\end{aligned}$$

- *Part 2.*

$$\begin{aligned}\frac{\mathrm{d}^2\alpha(x)^{-1}}{\mathrm{d}x_i \mathrm{d}x_j} &= 2\alpha(x)^{-1}\langle f(x), A_{*,i}\rangle\langle f(x), A_{*,j}\rangle - \alpha(x)^{-1}\langle f(x), A_{*,i}\circ A_{*,j}\rangle \\ &= 2\alpha(x)^{-1}A_{*,i}^\top f(x) f(x)^\top A_{*,j} - A_{*,i}^\top \operatorname{diag}(f(x)) A_{*,j}\end{aligned}$$

Proof. **Proof of Part 1.**

$$\begin{aligned}&\frac{\mathrm{d}^2\alpha(x)^{-1}}{\mathrm{d}x_i^2} \\ =& \frac{\mathrm{d}}{\mathrm{d}x_i}(\frac{\mathrm{d}}{\mathrm{d}x_i}\alpha(x)^{-1}) \\ =& \frac{\mathrm{d}}{\mathrm{d}x_i}(-\alpha(x)^{-1}\langle f(x), A_{*,i}\rangle) \\ =& -(\frac{\mathrm{d}}{\mathrm{d}x_i}\alpha(x)^{-1})\cdot\langle f(x), A_{*,i}\rangle - \alpha(x)^{-1}\frac{\mathrm{d}}{\mathrm{d}x_i}\langle f(x), A_{*,i}\rangle \\ =& 2\alpha(x)^{-1}\langle f(x), A_{*,i}\rangle^2 - \alpha(x)^{-1}\langle f(x), A_{*,i}\circ A_{*,i}\rangle\end{aligned}$$

where the 1st step follows from the expansion of hessian, the 2nd step follows from **Part 3** of Lemma C.6, the 3rd step follows from differential chain rule, the 4th step follows from simple algebra, the last step follows from Fact B.1.

Proof of Part 2.

$$\begin{aligned}\frac{\mathrm{d}^2\alpha(x)^{-1}}{\mathrm{d}x_i \mathrm{d}x_j} =& \frac{\mathrm{d}}{\mathrm{d}x_j}(\frac{\mathrm{d}}{\mathrm{d}x_i}\alpha(x)^{-1}) \\ =& \frac{\mathrm{d}}{\mathrm{d}x_j}(-\alpha(x)^{-1}\langle f(x), A_{*,i}\rangle) \\ =& -(\frac{\mathrm{d}}{\mathrm{d}x_j}\alpha(x)^{-1})\cdot\langle f(x), A_{*,i}\rangle - \alpha(x)^{-1}\frac{\mathrm{d}}{\mathrm{d}x_j}\langle f(x), A_{*,i}\rangle \\ =& -(-\alpha(x)^{-1}\langle f(x), A_{*,j}\rangle)\langle f(x), A_{*,i}\rangle - \alpha(x)^{-1}(-\langle f(x), A_{*,j}\rangle\langle f(x), A_{*,i}\rangle + \langle f(x), A_{*,i}\circ A_{*,j}\rangle) \\ =& 2\alpha(x)^{-1}\langle f(x), A_{*,i}\rangle\langle f(x), A_{*,j}\rangle - \alpha(x)^{-1}\langle f(x), A_{*,j}\circ A_{*,i}\rangle\end{aligned}$$

where the 1st step follows from the expansion of hessian, the 2nd step follows from **Part 3** of Lemma C.6, the 3rd step follows from differential chain rule, the 4th step follows from **Part 5** and **Part 3** in Lemma C.6, the last step step follows from simple algebra.

□

C.7 Hessian Calculations: Step 4, Hessian of $f(x)$

Lemma C.12 (Hessian of $f(x)$)**.** *If the following conditions hold*

- *Let* $f(x) = \langle \exp(Ax), \mathbf{1}_n\rangle^{-1}\exp(Ax)$ *(see Definition C.1).*

Then, we have

- *Part 1.*

$$\begin{aligned}\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x_i^2} =& \ 2\langle f(x), A_{*,i}\rangle^2 \cdot f(x) - \langle f(x), A_{*,i}\circ A_{*,i}\rangle \cdot f(x) \\ & - 2\langle f(x), A_{*,i}\rangle f(x)\circ A_{*,i} + A_{*,i}\circ f(x)\circ A_{*,i}\end{aligned}$$

- *Part 2.*

$$\begin{aligned}\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x_i \mathrm{d}x_j} =& 2\langle f(x), A_{*,i}\rangle\langle f(x), A_{*,j}\rangle f(x) - \langle f(x), A_{*,i}\circ A_{*,j}\rangle f(x) \\ & - \langle f(x), A_{*,i}\rangle f(x)\circ A_{*,j} - \langle f(x), A_{*,j}\rangle f(x)\circ A_{*,i} + A_{*,i}\circ f(x)\circ A_{*,j}\end{aligned}$$

Proof. **Proof of Part 1.**

$$\begin{aligned}\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x_i^2} &= \frac{\mathrm{d}}{\mathrm{d}x_i}(\frac{\mathrm{d}}{\mathrm{d}x_i} f(x)) \\ &= \frac{\mathrm{d}}{\mathrm{d}x_i}(-\langle f(x), A_{*,i}\rangle \cdot f(x) + f(x) \circ A_{*,i}) \\ &= 2\langle f(x), A_{*,i}\rangle^2 \cdot f(x) - \langle f(x), A_{*,i} \circ A_{*,i}\rangle \cdot f(x) \\ &\quad - 2\langle f(x), A_{*,i}\rangle f(x) \circ A_{*,i} + A_{*,i} \circ f(x) \circ A_{*,i}\end{aligned}$$

where the 1st step follows from the expansion of hessian, the 2nd step follows from **Part 4** of Lemma C.6, the third step follows from differential chain rule and **Part 4**, **Part 5** in Lemma C.6.

Proof of Part 2.

$$\begin{aligned}&\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x_i \mathrm{d}x_j} \\ =& \frac{\mathrm{d}}{\mathrm{d}x_j}(\frac{\mathrm{d}}{\mathrm{d}x_i} f(x)) \\ =& \frac{\mathrm{d}}{\mathrm{d}x_j}(-\langle f(x), A_{*,i}\rangle \cdot f(x) + f(x) \circ A_{*,i}) \\ =& 2\langle f(x), A_{*,i}\rangle \langle f(x), A_{*,j}\rangle f(x) - \langle f(x), A_{*,i} \circ A_{*,j}\rangle f(x) \\ &- \langle f(x), A_{*,i}\rangle f(x) \circ A_{*,j} - \langle f(x), A_{*,j}\rangle f(x) \circ A_{*,i} + A_{*,i} \circ f(x) \circ A_{*,j}\end{aligned}$$

where the 1st step follows from the expansion of hessian, the 2nd step follows from **Part 4** of Lemma C.6, the 3rd step follows from differential chain rule and **Part 4**, **Part 5** in Lemma C.6.

□

C.8 Hessian Calculations: Step 5, Hessian of $L_{\exp}(x)$

Lemma C.13 (Hessian of $L_{\exp}(x)$)**.** *We define*

- $B_1(x) \in \mathbb{R}^{n \times n}$ *such that*

$$A_{*,i}^\top B_1(x) A_{*,j} := (-\langle f(x), A_{*,j}\rangle f(x) + f(x) \circ A_{*,j})^\top \cdot (-\langle f(x), A_{*,i}\rangle f(x)) + f(x) \circ A_{*,i}))$$

- $B_2(x) \in \mathbb{R}^{n \times n}$ *such that*

$$\begin{aligned}A_{*,i}^\top B_2(x) A_{*,j} := c^\top \cdot (&2\langle f(x), A_{*,i}\rangle \langle f(x), A_{*,j}\rangle f(x) - \langle f(x), A_{*,i} \circ A_{*,j}\rangle f(x) \\ &- \langle f(x), A_{*,i}\rangle f(x) \circ A_{*,j} - \langle f(x), A_{*,j}\rangle f(x) \circ A_{*,i} + A_{*,i} \circ f(x) \circ A_{*,j})\end{aligned}$$

Then we have

- *Part 1.*

$$\frac{\mathrm{d}^2 L_{\exp}}{\mathrm{d}x_i^2} = A_{*,i}^\top B_1(x) A_{*,i} + A_{*,i}^\top B_2(x) A_{*,i}$$

- *Part 2.*

$$\frac{\mathrm{d}^2 L_{\exp}}{\mathrm{d}x_i \mathrm{d}x_j} = A_{*,i}^\top B_1(x) A_{*,j} + A_{*,i}^\top B_2(x) A_{*,j}$$

Proof. **Proof of Part 1.**

$$\begin{aligned}&\frac{\mathrm{d}^2 L_{\exp}}{\mathrm{d}x_i^2} \\ =& \frac{\mathrm{d}}{\mathrm{d}x_i}(\frac{\mathrm{d}L_{\exp}}{\mathrm{d}x_i})\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathrm{d}}{\mathrm{d}x_i}(\underbrace{(f(x)-b)^\top}_{1\times n}\underbrace{(-\langle f(x)\circ A_{*,i}, \mathbf{1}_n\rangle f(x)) + f(x)\circ A_{*,i})}_{n\times 1}) \\
&= (-\langle f(x), A_{*,i}\rangle f(x) + f(x)\circ A_{*,i})^\top \cdot (-\langle f(x), A_{*,i}\rangle f(x) + f(x)\circ A_{*,i}) \\
&\quad + c^\top \cdot (2\langle f(x), A_{*,i}\rangle^2 f(x) - \langle f(x), A_{*,i}\circ A_{*,i}\rangle f(x) - 2\langle f(x), A_{*,i}\rangle f(x)\circ A_{*,i} + A_{*,i}\circ f(x)\circ A_{*,i}) \\
&= A_{*,i}^\top B_1(x) A_{*,i} + A_{*,i}^\top B_2(x) A_{*,i}
\end{aligned}
$$

where the 1st step follows from the expansion of hessian, the 2nd step follows from **Part 7** of Lemma C.6, the third step follows from arguments in Lemma C.15.

Proof of Part 2.

$$
\begin{aligned}
&\frac{\mathrm{d}^2 L_{\exp}}{\mathrm{d}x_i \mathrm{d}x_j} \\
&= \frac{\mathrm{d}}{\mathrm{d}x_j}(\frac{\mathrm{d}L_{\exp}}{\mathrm{d}x_i}) \\
&= \frac{\mathrm{d}}{\mathrm{d}x_j}(\underbrace{(f(x)-b)^\top}_{1\times n}\underbrace{(-\langle f(x), A_{*,i}\rangle f(x) + f(x)\circ A_{*,i})}_{n\times 1}) \\
&= (-\langle f(x), A_{*,j}\rangle f(x) + f(x)\circ A_{*,j})^\top \cdot (-\langle f(x), A_{*,i}\rangle f(x) + f(x)\circ A_{*,i}) \\
&\quad + c^\top \cdot (2\langle f(x), A_{*,i}\rangle\langle f(x), A_{*,j}\rangle f(x) - \langle f(x), A_{*,i}\circ A_{*,j}\rangle f(x) - \langle f(x), A_{*,i}\rangle f(x)\circ A_{*,j} \\
&\quad - \langle f(x), A_{*,j}\rangle f(x)\circ A_{*,i} + A_{*,i}\circ f(x)\circ A_{*,j}) \\
&= A_{*,i}^\top B_1(x) A_{*,j} + A_{*,i}^\top B_2(x) A_{*,j}
\end{aligned}
$$

where the 1st step follows from the expansion of hessian, the 2nd step follows from **Part 7** of Lemma C.6, the 3rd step is a simplification of step 2 by applying notations α (Definition C.4) and c (Definition C.5) and arguments in Lemma C.15.

□

C.9 Helpful Lemma

The goal of this section to prove Lemma C.14. We remark that in this lemma, we can replace $f(x)$ by any vector. However, for easy of presentation, we use $f(x)$.

Lemma C.14. *For any length-n vector $c \in \mathbb{R}^n$ and any vector $f(x) \in \mathbb{R}^n$, we have*

- *Part 1.*
$$c^\top (A_{*,i}\circ f(x)\circ A_{*,j}) = A_{*,i}^\top \underbrace{\mathrm{diag}(c\circ f(x))}_{n\times n} A_{*,j}$$
- *Part 2.*
$$c^\top f(x)\langle f(x), A_{*,i}\rangle\langle f(x), A_{*,j}\rangle = A_{*,i}^\top \underbrace{f(x)}_{n\times 1}\underbrace{\langle c, f(x)\rangle}_{\text{scalar}}\underbrace{f(x)^\top}_{1\times n} A_{*,j}$$
- *Part 3.*
$$c^\top \langle f(x), A_{*,i}\circ A_{*,j}\rangle f(x) = A_{*,i}^\top \underbrace{\mathrm{diag}(\langle c, f(x)\rangle f(x))}_{n\times n} A_{*,j}$$
- *Part 4.*
$$c^\top \langle f(x), A_{*,j}\rangle f(x)\circ A_{*,i} = A_{*,i}^\top \underbrace{(c\circ f(x))}_{n\times 1}\underbrace{f(x)^\top}_{1\times n} A_{*,j}$$
- *Part 5.*
$$c^\top \langle f(x), A_{*,i}\rangle f(x)\circ A_{*,j} = A_{*,i}^\top \underbrace{f(x)}_{n\times 1}\underbrace{(f(x)\circ c)^\top}_{1\times n} A_{*,j}$$

- *Part 6.*
$$(\langle f(x), A_{*,j} \rangle f(x))^\top f(x) \circ A_{*,i} = A_{*,i} \underbrace{(f(x) \circ f(x))}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} A_{*,j}$$
- *Part 7.*
$$(f(x) \circ A_{*,i})^\top (f(x) \circ A_{*,j}) = A_{*,i}^\top \underbrace{\mathrm{diag}(f(x) \circ f(x))}_{n \times n} A_{*,j}$$
- *Part 8.*
$$(\langle f(x), A_{*,j} \rangle f(x))^\top (\langle f(x), A_{*,i} \rangle f(x)) = A_{*,i}^\top \underbrace{f(x)}_{n \times 1} \underbrace{\langle f(x), f(x) \rangle}_{\mathrm{scalar}} \underbrace{f(x)^\top}_{1 \times n} A_{*,j}$$
- *Part 9.*
$$(f(x) \circ A_{*,i})^\top (f(x) \circ A_{*,j}) = A_{*,i}^\top \mathrm{diag}(f(x) \circ f(x)) A_{*,j}$$

Proof. **Proof of Part 1.**
$$\begin{aligned} c^\top (A_{*,i} \circ f(x) \circ A_{*,j}) &= A_{*,i}^\top (c \circ f(x) \circ A_{*,j}) \\ &= A_{*,i}^\top \mathrm{diag}(c \circ f(x)) \circ A_{*,j} \end{aligned}$$
where the 1st step follows from Fact B.2, the 2nd step follows from Fact B.2.

Proof of Part 2.
$$\begin{aligned} c^\top f(x) \langle f(x), A_{*,i} \rangle \langle f(x), A_{*,j} \rangle &= \langle c, f(x) \rangle \langle f(x), A_{*,i} \rangle \langle f(x), A_{*,j} \rangle \\ &= A_{*,i}^\top f(x) \langle c, f(x) \rangle (f(x))^\top A_{*,j} \end{aligned}$$
where the 1st step follows from $a^\top b = \langle a, b \rangle$ (Fact B.1), the 2nd step follows from $\langle a, b \rangle = a^\top b$ (Fact B.1).

Proof of Part 3.
$$\begin{aligned} c^\top \langle f(x), A_{*,i} \circ A_{*,j} \rangle f(x) &= c^\top (f(x))^\top A_{*,i} \circ A_{*,j} f(x) \\ &= A_{*,i}^\top (f(x))^\top c \circ A_{*,j} f(x) \\ &= A_{*,i}^\top \langle f(x), c \rangle \circ A_{*,j} f(x) \\ &= A_{*,i}^\top \mathrm{diag}(\langle f(x), c \rangle) f(x) A_{*,j} \end{aligned}$$
where the 1st step follows from $\langle a, b \rangle = a^\top b$ (Fact B.1), the 2nd step follows from Fact B.2, the 3rd step follows from $a^\top b = \langle a, b \rangle$ (Fact B.1), the last step follows from Fact B.2.

Proof of Part 4.
$$\begin{aligned} c^\top \langle f(x), A_{*,j} \rangle f(x) \circ A_{*,i} &= c^\top (f(x))^\top A_{*,j} f(x) \circ A_{*,i} \\ &= A_{*,i}^\top (f(x))^\top A_{*,j} f(x) \circ c \\ &= A_{*,i}^\top (f(x) \circ c) (f(x))^\top A_{*,j} \end{aligned}$$
where the 1st step follows from $\langle a, b \rangle = a^\top b$ (Fact B.1), the 2nd step follows from Fact B.2, the 3rd step follows from $f(x)^\top A_{*,j} = \langle f(x), A_{*,j} \rangle$ (Fact B.1) is a scalar.

Proof of Part 5.
$$\begin{aligned} c^\top \langle f(x), A_{*,i} \rangle f(x) \circ A_{*,j} &= (f(x))^\top A_{*,i} c^\top f(x) \circ A_{*,j} \\ &= (f(x))^\top A_{*,i} A_{*,j}^\top f(x) \circ c \\ &= (f(x))^\top A_{*,i} (f(x) \circ c)^\top A_{*,j} \\ &= A_{*,i}^\top f(x) (f(x) \circ c)^\top A_{*,j} \end{aligned}$$

where the 1st step follows from $\langle a, b \rangle = a^\top b$ (Fact B.1), the 2nd step follows from Fact B.2, the 3rd step follows from $a^\top b = b^\top a$ (Fact B.1), the last step follows from $a^\top b = b^\top a$ (Fact B.1).

Proof of Part 6

$$\begin{aligned}(\langle f(x), A_{*,j} \rangle f(x))^\top f(x) \circ A_{*,i} &= A_{*,i}^\top f(x) \circ \langle f(x), A_{*,j} \rangle f(x) \\ &= A_{*,i}^\top f(x) \circ (f(x))^\top A_{*,j} f(x) \\ &= A_{*,i}^\top f(x) \circ f(x) (f(x))^\top A_{*,j}\end{aligned}$$

where the 1st step follows from Fact B.2, the 2nd step follows from $\langle a, b \rangle = a^\top b$ (Fact B.1), the 3rd step follows from $f(x)^\top A_{*,j} = \langle f(x), A_{*,j} \rangle$ is a scalar (Fact B.1).

Proof of Part 7.

$$\begin{aligned}(f(x) \circ A_{*,i})^\top (f(x) \circ A_{*,j}) &= \langle f(x) \circ A_{*,i}, f(x) \circ A_{*,j} \rangle \\ &= \langle f(x) \circ f(x), A_{*,i} \circ A_{*,j} \rangle \\ &= (f(x) \circ f(x))^\top (A_{*,i} \circ A_{*,j}) \\ &= A_{*,i}^\top (f(x) \circ f(x) \circ A_{*,j}) \\ &= A_{*,i}^\top \operatorname{diag}(f(x) \circ f(x)) A_{*,j}\end{aligned}$$

where the 1st step follows from $a^\top b = \langle a, b \rangle$ (Fact B.1), the 2nd step follows from Fact B.1, the 3rd step follows from $\langle a, b \rangle = a^\top b$ (Fact B.1), the 4th step follows from Fact B.2, the last step follows from Fact B.2.

Proof of Part 8.

$$\begin{aligned}(\langle f(x), A_{*,j} \rangle f(x))^\top (\langle f(x), A_{*,i} \rangle f(x)) &= \langle f(x), A_{*,j} \rangle f(x)^\top (\langle f(x), A_{*,i} \rangle f(x)) \\ &= f(x)^\top A_{*,j} f(x)^\top f(x)^\top A_{*,i} f(x) \\ &= f(x)^\top A_{*,i} f(x)^\top A_{*,j} f(x)^\top f(x) \\ &= A_{*,i}^\top f(x) f(x)^\top A_{*,j} f(x)^\top f(x) \\ &= A_{*,i}^\top f(x) f(x)^\top f(x) f(x)^\top A_{*,j} \\ &= A_{*,i}^\top f(x) \langle f(x), f(x) \rangle f(x)^\top A_{*,j}\end{aligned}$$

where the 1st step follows from $a^\top b = b^\top a$ (Fact B.1), the 2nd step follows from $\langle a, b \rangle = a^\top b$ (Fact B.1), the 3rd step follows from $a^\top b = b^\top a$ (Fact B.1), the 4th step follows from $A_{*,i}^\top f(x) = \langle A_{*,i}, f(x) \rangle$ is a scalar (Fact B.1), the 5th step step follows from $f(x)^\top A_{*,j} = \langle f(x), A_{*,j} \rangle$ is a scalar (Fact B.1), the last step follows from $a^\top b = \langle a, b \rangle$ (Fact B.1).

Proof of Part 9.

$$\begin{aligned}(f(x) \circ A_{*,i})^\top (f(x) \circ A_{*,j}) &= \langle f(x) \circ A_{*,i}, f(x) \circ A_{*,j} \rangle \\ &= \langle f(x) \circ f(x), A_{*,i} \circ A_{*,j} \rangle \\ &= (f(x) \circ f(x))^\top (A_{*,i} \circ A_{*,j}) \\ &= A_{*,i}^\top (f(x) \circ f(x) \circ A_{*,j}) \\ &= A_{*,i}^\top \operatorname{diag}(f(x) \circ f(x)) A_{*,j}\end{aligned}$$

where the 1st step follows from $a^\top b = \langle a, b \rangle$ (Fact B.1), the 2nd step follows from Fact B.1 , the 3rd step follows from $\langle a, b \rangle = a^\top b$ (Fact B.1), the 4th step follows from Fact B.2, the last step follows from Fact B.2. □

C.10 Decomposing $B_1(x)$, $B_2(x)$ and $B(x)$ into Low Rank Plus Diagonal

Lemma C.15 (Rewriting $B_1(x)$ and $B_2(x)$, formal version of Lemma 5.1)**.** *If the following conditions hold*

- *Given matrix $A \in \mathbb{R}^{n \times d}$.*
- *Let $f(x)$ be defined as Definition C.1.*
- *Let $c(x)$ be defined as Definition C.5.*
- *Let $B(x) = B_1(x) + B_2(x)$.*

Then, we can show that

- **Part 1.** *For $B_1(x) \in \mathbb{R}^{n \times n}$, we have*

$$\begin{aligned} B_1(x) = & \underbrace{\langle f(x), f(x) \rangle}_{\text{scalar}} \cdot \underbrace{f(x)}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} + \underbrace{\mathrm{diag}(f(x) \circ f(x))}_{n \times n \text{ diagonal matrix}} \\ & + \underbrace{(f(x) \circ f(x))}_{n \times 1} \cdot \underbrace{f(x)^\top}_{1 \times n} + \underbrace{(f(x) \circ f(x))}_{n \times 1} \cdot \underbrace{f(x)^\top}_{1 \times n} \end{aligned}$$

  - *In summary, $B_1(x) \in \mathbb{R}^{n \times n}$ is constructed by three rank-1 matrices and a diagonal matrix.*

- **Part 2.** *For $B_2(x) \in \mathbb{R}^{n \times n}$, we have*

$$\begin{aligned} B_2(x) = & \underbrace{2\langle c(x), f(x) \rangle}_{\text{scalar}} \cdot \underbrace{f(x)}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} + \underbrace{\langle c(x), f(x) \rangle}_{\text{scalar}} \cdot \underbrace{\mathrm{diag}(f(x))}_{n \times n \text{ diagonal matrix}} + \underbrace{\mathrm{diag}(c(x) \circ f(x))}_{n \times n \text{ diagonal matrix}} \\ & - \underbrace{(c(x) \circ f(x))}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} - \underbrace{f(x)}_{n \times 1} \underbrace{(f(x) \circ c(x))^\top}_{1 \times n} \end{aligned}$$

  - *In summary, $B_2(x) \in \mathbb{R}^{n \times n}$ is constructed by three rank-1 matrices and two diagonal matrices.*

- **Part 3.** *For $B(x) \in \mathbb{R}^{n \times n}$, we have*

$$\begin{aligned} B(x) = & \underbrace{\langle 3f(x) - 2b, f(x) \rangle}_{\text{scalar}} \cdot \underbrace{f(x)}_{n \times 1} \underbrace{f(x)^\top}_{1 \times n} \\ & + \underbrace{\langle f(x) - b, f(x) \rangle}_{\text{scalar}} \cdot \underbrace{\mathrm{diag}(f(x))}_{n \times n \text{ diagonal matrix}} \\ & + \underbrace{\mathrm{diag}((2f(x) - b) \circ f(x))}_{n \times n \text{ diagonal matrix}} \\ & + \underbrace{(b \circ f(x))}_{n \times 1} \cdot \underbrace{f(x)^\top}_{1 \times n} + \underbrace{f(x)}_{n \times 1} \cdot \underbrace{(b \circ f(x))^\top}_{1 \times n} \end{aligned}$$

  - *In summary, $B(x) \in \mathbb{R}^{n \times n}$ is constructed by three rank-1 matrices and two diagonal matrices.*

Proof. **Proof of Part 1. $B_1(x)$.**

For $B_1(x)$, we have:

$$\begin{aligned} A_{*,i}^\top B_1(x) A_{*,j} = & (-\langle f(x), A_{*,j} \rangle f(x) + f(x) \circ A_{*,j})^\top \cdot (-\langle f(x), A_{*,i} \rangle f(x)) + f(x) \circ A_{*,i})) \\ = & (-(\langle f(x), A_{*,j} \rangle f(x))^\top + (f(x) \circ A_{*,j})^\top) \cdot (-\langle f(x), A_{*,i} \rangle f(x)) + f(x) \circ A_{*,i})) \\ = & (\langle f(x), A_{*,j} \rangle f(x))^\top \langle f(x), A_{*,i} \rangle f(x) + (f(x) \circ A_{*,j})^\top (f(x) \circ A_{*,i})) \\ & - (\langle f(x), A_{*,j} \rangle f(x))^\top (f(x) \circ A_{*,i}) - (f(x) \circ A_{*,j})^\top \langle f(x), A_{*,i} \rangle f(x) \\ = & A_{*,i}^\top f(x) \langle f(x), f(x) \rangle f(x)^\top A_{*,j} + A_{*,i}^\top \mathrm{diag}(f(x) \circ f(x)) A_{*,j} \\ & - A_{*,i}^\top (f(x) \circ f(x)) f(x)^\top A_{*,j} - A_{*,i}^\top (f(x) \circ f(x))^\top f(x) A_{*,j} \end{aligned} \tag{2}$$

where the 1st step follows from the definition of $B_1(x)$, the 2nd step follows from $(A + B)^\top = A^\top + B^\top$, the 3rd step follows from simple algebra, the last step follows from Lemma C.14.

Thus, by extracting $A_{*,i}^\top$ and $A_{*,j}$, we have:

$$\begin{aligned} B_1(x) = & \langle f(x), f(x) \rangle \cdot f(x) f(x)^\top + \mathrm{diag}(f(x) \circ f(x)) \\ & + (f(x) \circ f(x)) f(x)^\top + f(x)(f(x) \circ f(x))^\top \end{aligned}$$

Proof of Part 2. $B_2(x)$**.**

For $B_2(x) \in \mathbb{R}^{n \times n}$, we have:

$$\begin{aligned} & A_{*,i}^\top B_2(x) A_{*,j} \\ = & c(x)^\top \cdot (2\langle f(x), A_{*,i} \rangle \langle f(x), A_{*,j} \rangle f(x) - \langle f(x), A_{*,i} \circ A_{*,j} \rangle f(x) - \langle f(x), A_{*,i} \rangle f(x) \circ A_{*,j} \\ & - \langle f(x), A_{*,j} \rangle f(x) \circ A_{*,i} + A_{*,i} \circ f(x) \circ A_{*,j}) \end{aligned}$$

Thus, we can rewrite $B_2(x)$ as

$$\begin{aligned} & A_{*,i}^\top B_2(x) A_{*,j} \\ = & c(x)^\top \cdot (2\langle f(x), A_{*,i} \rangle \langle f(x), A_{*,j} \rangle f(x) - \langle f(x), A_{*,i} \circ A_{*,j} \rangle f(x) - \langle f(x), A_{*,i} \rangle f(x) \circ A_{*,j} \\ & - \langle f(x), A_{*,j} \rangle f(x) \circ A_{*,i} + A_{*,i} \circ f(x) \circ A_{*,j}) \\ = & 2c(x)^\top \langle f(x), A_{*,i} \rangle \langle f(x), A_{*,j} \rangle f(x) - c(x)^\top \langle f(x), A_{*,i} \circ A_{*,j} \rangle f(x) + c(x)^\top A_{*,i} \circ f(x) \circ A_{*,j} \\ & - c(x)^\top \langle f(x), A_{*,i} \rangle f(x) \circ A_{*,j} - c(x)^\top \langle f(x), A_{*,j} \rangle f(x) \circ A_{*,i} \\ = & 2A_{*,i}^\top f(x) \langle c(x), f(x) \rangle f(x)^\top A_{*,j} - A_{*,i}^\top \mathrm{diag}(\langle c(x), f(x) \rangle f(x)) A_{*,j} + A_{*,i}^\top \mathrm{diag}(c(x) \circ f(x)) A_{*,j} \\ & - A_{*,i}^\top f(x)(f(x) \circ c)^\top A_{*,j} - A_{*,i}(c(x) \circ f(x)) f(x)^\top A_{*,j} \end{aligned} \tag{3}$$

where the 1st step follows from definition of $B_2(x)$, the 2nd step follows from simple algebra, the 3rd step follows from simple algebra, the last step follows from Lemma C.14.

By extracting $A_{*,i}^\top$ and $A_{*,j}$, we have

$$\begin{aligned} B_2(x) = & 2\langle c, f(x) \rangle f(x) f(x)^\top + \mathrm{diag}(\langle c, f(x) \rangle f(x)) + \mathrm{diag}(c \circ f(x)) \\ & - (c(x) \circ f(x)) f(x)^\top - f(x)(c(x) \circ f(x))^\top \end{aligned}$$

Proof of Part 3. $B(x)$

We define

$$\begin{aligned} B_{1,1}(x) := & \langle f(x), f(x) \rangle \cdot f(x) f(x)^\top \\ B_{1,2}(x) := & \mathrm{diag}(f(x) \circ f(x)) \\ B_{1,3}(x) := & (f(x) \circ f(x)) f(x)^\top \\ B_{1,4}(x) := & f(x)(f(x) \circ f(x))^\top \end{aligned}$$

Thus, we have:

$$B_1(x) = B_{1,1}(x) + B_{1,2}(x) + B_{1,3}(x) + B_{1,4}(x)$$

Similarly, we define

$$\begin{aligned} B_{2,1}(x) := & 2\langle c, f(x) \rangle f(x) f(x)^\top \\ B_{2,2}(x) := & \mathrm{diag}(\langle c, f(x) \rangle f(x)) \\ B_{2,3}(x) := & \mathrm{diag}(c \circ f(x)) \\ B_{2,4}(x) := & -(c(x) \circ f(x)) f(x)^\top \\ B_{2,5}(x) := & -f(x)(c(x) \circ f(x))^\top \end{aligned}$$

Thus, we have:

$$B_2(x) = B_{2,1}(x) + B_{2,2}(x) + B_{2,3}(x) + B_{2,4}(x) + B_{2,5}(x)$$

Merge $B_{1,1}(x)$ and $B_{2,1}(x)$:

$$\begin{aligned} B_{1,1}(x) + B_{2,1}(x) &= \langle f(x), f(x) \rangle \cdot f(x) f(x)^\top + 2\langle c(x), f(x) \rangle f(x) f(x)^\top \\ &= \langle 3f(x) - 2b, f(x) \rangle f(x) f(x)^\top \end{aligned}$$

Maintain $B_{2,2}(x)$ itself:

$$\begin{aligned} B_{2,2}(x) = &\ \mathrm{diag}(\langle f(x) - b, f(x) \rangle f(x)) \\ = &\ \langle f(x) - b, f(x) \rangle \, \mathrm{diag}(f(x)) \end{aligned}$$

Merge $B_{1,2}(x)$ and $B_{2,3}(x)$:

$$\begin{aligned} B_{1,2}(x) + B_{2,3}(x) = &\ \mathrm{diag}((f(x) - b) \circ f(x)) + \mathrm{diag}(f(x) \circ f(x)) \\ = &\ \mathrm{diag}((2f(x) - b) \circ f(x)) \end{aligned}$$

Merge $B_{1,3}(x)$ and $B_{2,4}(x)$:

$$\begin{aligned} B_{1,3}(x) + B_{2,4}(x) &= (f(x) \circ f(x)) f(x)^\top - ((f(x) - b) \circ f(x)) f(x)^\top \\ &= (f(x) \circ f(x) - f(x) \circ f(x) + b \circ f(x)) f(x)^\top \\ &= (b \circ f(x)) f(x)^\top \end{aligned}$$

Merge $B_{1,4}(x)$ and $B_{2,5}(x)$:

$$\begin{aligned} B_{1,4}(x) + B_{2,5}(x) &= f(x)(f(x) \circ f(x))^\top - f(x)((f(x) - b) \circ f(x))^\top \\ &= f(x)(f(x)^\top \circ f(x)^\top - f(x)^\top \circ f(x)^\top + b^\top \circ f(x)^\top) \\ &= f(x)(b \circ f(x))^\top \end{aligned}$$

By combining all the above equations, we have

$$\begin{aligned} B(x) = &\ \underbrace{\langle 3f(x) - 2b, f(x) \rangle f(x) f(x)^\top}_{B_{1,1} + B_{2,1}} \\ &+ \underbrace{\langle f(x) - b, f(x) \rangle \, \mathrm{diag}(f(x))}_{B_{2,2}} \\ &+ \underbrace{\mathrm{diag}((2f(x) - b) \circ f(x))}_{B_{1,2} + B_{2,3}} \\ &+ \underbrace{(b \circ f(x)) f(x)^\top}_{B_{1,3} + B_{2,4}} + \underbrace{f(x)(b \circ f(x))^\top}_{B_{1,4} + B_{2,5}} \end{aligned}$$

Thus, we complete the proof. □

D HESSIAN IS POSITIVE DEFINITE

In this section, we prove that $\nabla^2 L \succeq 0$ and thus L is convex. In Section D.1, we find the lower bound of $B(x)$. To be specific, we split $B(x)$ into several terms and find their lower bounds separately. In Section D.2, we use the result of Section D.1 to prove that lower bound of $\nabla^2 L \succeq 0$ and thus L is convex.

D.1 PSD LOWER BOUND

For convenient, we define $B(x)$

Definition D.1. *We define $B(x)$ as follows*

$$B(x) := \langle 3f(x) - 2b, f(x) \rangle f(x) f(x)^\top$$

$$
\begin{aligned}
&+ (b \circ f(x)) f(x)^\top + f(x)(b \circ f(x))^\top \\
&+ \langle f(x) - b, f(x) \rangle \cdot \mathrm{diag}(f(x)) \\
&+ \mathrm{diag}((2f(x) - b) \circ f(x))
\end{aligned}
$$

Further, we define

$$
\begin{aligned}
B_{\mathrm{rank}}(x) &:= \underbrace{\langle 3f(x) - 2b, f(x) \rangle f(x) f(x)^\top}_{:= B^1_{\mathrm{rank}}(x)} + \underbrace{(b \circ f(x)) f(x)^\top + f(x)(b \circ f(x))^\top}_{:= B^2_{\mathrm{rank}}(x)} \\
B_{\mathrm{diag}}(x) &:= \underbrace{\langle f(x) - b, f(x) \rangle \cdot \mathrm{diag}(f(x))}_{:= B^1_{\mathrm{diag}}(x)} + \underbrace{\mathrm{diag}((2f(x) - b) \circ f(x))}_{:= B^2_{\mathrm{diag}}(x)}
\end{aligned}
$$

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

- $\|f(x)\|_1 = 1$ *(see Definition C.1).*
- *Let* $B(x) \in \mathbb{R}^{n \times n}$ *be defined as Definition D.1.*
- *Let* $f(x) \geq \mathbf{0}_n$.
- *Let* $b \geq \mathbf{0}_n$.
- *Let* B^1_{rank}, B^2_{rank} *be defined as Definition D.1.*
- *Let* B^1_{diag}, B^2_{diag} *be defined as Definition D.1.*

Then we have

- *Part 1.*
$$-0.5\|b\|_2^2 \cdot f(x) f(x)^\top \preceq B^1_{\mathrm{rank}}(x) \preceq (3\|f(x)\|_2^2) \cdot f(x) f(x)^\top$$
- *Part 2.*
$$-(1 + \|b\|_\infty^2) \cdot f(x) f(x)^\top \preceq B^2_{\mathrm{rank}}(x) \preceq (1 + \|b\|_\infty^2) \cdot f(x) f(x)^\top$$
- *Part 3.*
$$-0.25\|b\|_2^2 \cdot \mathrm{diag}(f(x)) \preceq B^1_{\mathrm{diag}} \preceq (\|f(x)\|_2^2) \cdot \mathrm{diag}(f(x))$$
- *Part 4.*
$$-0.5 \cdot \mathrm{diag}(b \circ b) \preceq B^2_{\mathrm{diag}} \preceq 2 \cdot \mathrm{diag}(f(x) \circ f(x))$$
- *Part 5. If* $\|b\|_1 \leq 1$ *and* $\|f(x)\|_1 \leq 1$*, then we have*
$$-4I_n \preceq B(x) \preceq 8I_n$$

Proof. Recall that in Definition D.1, we split $B(x)$ into four terms

$$B(x) = B^1_{\mathrm{rank}} + B^2_{\mathrm{rank}} + B^1_{\mathrm{diag}} + B^2_{\mathrm{diag}},$$

where B^i_{rank} and B^i_{diag} are defined as

$$
\begin{aligned}
B^1_{\mathrm{rank}} &:= \langle 3f(x) - 2b, f(x) \rangle f(x) f(x)^\top, \\
B^2_{\mathrm{rank}} &:= (b \circ f(x)) f(x)^\top + f(x)(b \circ f(x))^\top, \\
B^1_{\mathrm{diag}} &:= \langle f(x) - b, f(x) \rangle \, \mathrm{diag}(f(x)), \\
B^2_{\mathrm{diag}} &:= \mathrm{diag}(f(x) \circ (2f(x) - b)).
\end{aligned}
$$

Proof of B^1_{rank}.

On one hand, we can lower bound the coefficient, we have

$$\begin{aligned}\langle 3f(x) - 2b, f(x)\rangle &\geq 2\langle f(x) - b, f(x)\rangle \\ &= 2\langle f(x) - b, f(x)\rangle + 0.5\|b\|_2^2 - 0.5\|b\|_2^2 \\ &= 0.5\|2f(x) - b\|_2^2 - 0.5\|b\|_2^2 \\ &\geq -0.5\|b\|_2^2.\end{aligned}$$

Thus,

$$B_{\mathrm{rank}}^1 \succeq -0.5\|b\|_2^2 f(x)f(x)^\top.$$

On the other hard, we have

$$\begin{aligned}\langle 3f(x) - 2b, f(x)\rangle &= 3\|f(x)\|_2^2 - 2\langle b, f(x)\rangle \\ &\leq 3\|f(x)\|_2^2\end{aligned}$$

Thus,

$$B_{\mathrm{rank}}^1 \preceq 3\|f(x)\|_2^2 \cdot f(x)f(x)^\top.$$

Proof of $B_{\mathrm{rank}}^2(x)$.

On one hand, we have

$$\begin{aligned}B_{\mathrm{rank}}^2(x) &\succeq -(b \circ f(x))^\top (b \circ f(x)) - f(x)f(x)^\top \\ &= -(\|b\|_\infty^2 + 1) \cdot f(x)f(x)^\top,\end{aligned}$$

where the 1st step follows from Fact B.5, , the last step follows from Fact B.5.

On the other hand, we have

$$\begin{aligned}B_{\mathrm{rank}}^2(x) &\preceq (b \circ f(x))^\top (b \circ f(x)) + f(x)f(x)^\top \\ &\preceq (\|b\|_\infty^2 + 1) \cdot f(x)f(x)^\top\end{aligned}$$

where the 1st step follows from Fact B.5 , the 2nd step follows from Fact B.5 .

Proof of $B_{\mathrm{diag}}^1(x)$.

For the coefficient, we have

$$\begin{aligned}\langle f(x) - b, f(x)\rangle &= \langle f(x) - b, f(x)\rangle + \frac{1}{4}\|b\|_2^2 - \frac{1}{4}\|b\|_2^2 \\ &= \|f(x) - \frac{1}{2}b\|_2^2 - \frac{1}{4}\|b\|_2^2 \\ &\geq -\frac{1}{4}\|b\|_2^2\end{aligned}$$

Thus, we have

$$B_{\mathrm{diag}}^1 \succeq -\frac{1}{4}\|b\|_2^2 \cdot \mathrm{diag}(f(x)).$$

We can show

$$\begin{aligned}\langle f(x) - b, f(x)\rangle &= \|f(x)\|_2^2 - \langle b, f(x)\rangle \\ &\leq \|f(x)\|_2^2\end{aligned}$$

We have,

$$B_{\mathrm{diag}}^2 \preceq (\|f(x)\|_2^2) \cdot \mathrm{diag}(f(x))$$

Proof of $B_{\mathrm{diag}}^2(x)$.

For the third term, we have

$$\begin{aligned} B_{\mathrm{diag}}^2 = & \operatorname{diag}(f(x) \circ (2f(x) - b) + \frac{1}{2} b \circ b) - \frac{1}{2} \operatorname{diag}(b \circ b) \\ \succeq & -\frac{1}{2} \operatorname{diag}(b \circ b) \\ \succeq & -\frac{1}{2} \|b\|_2^2 I_n \end{aligned}$$

where the 1st step follows from simple algebra, the 2nd step follows from simple algebra, the last step follows from Fact B.5.

Proof of $B(x)$. It trivially follows from

$$\|f(x)\|_1 \leq 1, \|b\|_1 \leq 1$$

and using Lemma C.2 and Fact B.5

$$\max\{f(x)f(x)^\top, \operatorname{diag}(f(x)), \operatorname{diag}(f(x) \circ f(x)), \operatorname{diag}(b \circ b)\} \preceq I_n.$$

□

D.2 Lower bound on Hessian

The goal of this section is to prove Lemma D.3.

Lemma D.3 (Formal version of Lemma 5.2)**.** *If the following conditions hold*

- *Given matrix $A \in \mathbb{R}^{n \times d}$.*
- *Let $L_{\exp}(x)$ be defined as Definition C.3.*
- *Let $L_{\mathrm{reg}}(x)$ be defined as Definition B.8.*
- *Let $L(x) = L_{\exp}(x) + L_{\mathrm{reg}}(x)$.*
- *Let $W = \operatorname{diag}(w) \in \mathbb{R}^{n \times n}$. Let $W^2 \in \mathbb{R}^{n \times n}$ denote the matrix that i-th diagonal entry is $w_{i,i}^2$.*
- *Let $\sigma_{\min}(A)$ denote the minimum singular value of A.*
- *Let $l > 0$ denote a scalar.*

Then, we have

- *Part 1. If all $i \in [n]$, $w_i^2 \geq 4 + l/\sigma_{\min}(A)^2$, then*
$$\frac{\mathrm{d}^2 L}{\mathrm{d}x^2} \succeq l \cdot I_d$$
- *Part 2. If all $i \in [n]$, $w_i^2 \geq 100 + l/\sigma_{\min}(A)^2$, then*
$$(1 - 1/10) \cdot (B(x) + W^2) \preceq W^2 \preceq (1 + 1/10) \cdot (B(x) + W^2)$$

Proof. By applying Lemma C.13 and Lemma C.15, we have

$$\frac{\mathrm{d}^2 L_{\exp}}{\mathrm{d}x^2} = A^\top B(x) A$$

where

$$B(x) \succeq -4 I_n \tag{4}$$

Also, it's trivial that

$$\frac{\mathrm{d}^2 L}{\mathrm{d}x^2} = \frac{\mathrm{d}^2 L_{\mathrm{reg}}}{\mathrm{d}x^2} + \frac{\mathrm{d}^2 L_{\exp}}{\mathrm{d}x^2} \tag{5}$$

Thus, by applying Lemma B.9, Eq. (5) can be written as

$$\begin{aligned}\frac{\mathrm{d}^2 L}{\mathrm{d}x^2} &= A^\top B(x) A + A^\top W^2 A \\ &= A^\top (B(x) + W^2) A\end{aligned}$$

Let

$$D = B(x) + W^2$$

Then, $\frac{\mathrm{d}^2 L}{\mathrm{d}x^2}$ can be rewrite as

$$\frac{\mathrm{d}^2 L}{\mathrm{d}x^2} = A^\top D A$$

Now, we can bound D as follows

$$\begin{aligned}D \succeq & -4 I_n + w_{\min}^2 I_n \\ = & (-4 + w_{\min}^2) I_n \\ \succeq & \frac{l}{\sigma_{\min}(A)^2} I_n\end{aligned}$$

where 2nd step follows from simple algebra, the 3rd step follows from $w_{\min}^2 \geq 4 + l/\sigma_{\min}(A)^2$.

Since D is positive definite, then we have

$$A^\top D A \succeq \sigma_{\min}(D) \cdot \sigma_{\min}(A)^2 I_d \succeq l \cdot I_d$$

Thus, Hessian is positive definite forever and thus the function is convex. □

E HESSIAN IS LIPSCHITZ

In this section, we find the upper bound of $\|\nabla^2 L(x) - \nabla^2 L(y)\|$ and thus proved that $\nabla^2 L$ is lipschitz. In Section E.2, we prove that some basic terms satisfy the property of Lipschitz. In Section E.3, we provide a sketch of how we find the bound of $\|\nabla^2 L(x) - \nabla^2 L(y)\|$, to be specific, we split $\|\nabla^2 L(x) - \nabla^2 L(y)\|$ into 8 terms and state that all these terms can be bound by using $\|f(x) - f(y)\|$. In Section E.4, we use $\|f(x) - f(y)\|$ to bound the first term. In Section E.5, we use $\|f(x) - f(y)\|$ to bound the second term. In Section E.6, we use $\|f(x) - f(y)\|$ to bound the third term. In Section E.7, we use $\|f(x) - f(y)\|$ to bound the fourth term. In Section E.8, we use $\|f(x) - f(y)\|$ to bound the fifth term. In Section E.9, we use $\|f(x) - f(y)\|$ to bound the sixth term. In Section E.10, we use $\|f(x) - f(y)\|$ to bound the seventh term. In Section E.11, we use $\|f(x) - f(y)\|$ to bound the last term.

E.1 MAIN RESULT

Lemma E.1 (Formal version of Lemma 5.3)**.** *If the following condition holds*

- *Let* $H(x) = \frac{\mathrm{d}^2 L}{\mathrm{d}x^2}$
- *Let* $R > 2$
- $\|x\|_2 \leq R, \|y\|_2 \leq R$
- $\|A(x-y)\|_\infty < 0.01$
- $\|A\| \leq R$
- $\|b\|_2 \leq R$
- $\langle \exp(Ax), \mathbf{1}_n \rangle \geq \beta$ *and* $\langle \exp(Ay), \mathbf{1}_n \rangle \geq \beta$

Then we have

$$\|H(x) - H(y)\| \leq \beta^{-2} n \exp(20R^2) \cdot \|x - y\|_2$$

Proof.

$$\begin{aligned}
&\|H(x) - H(y)\| \\
\leq &\|A\| \cdot (2\|G_1\| + \|G_2\| + \cdots + \|G_8\|)\|A\| \\
\leq &R^2 \cdot (2\|G_1\| + \|G_2\| + \cdots + \|G_8\|) \\
\leq &R^2 \cdot 100R \cdot \|f(x) - f(y)\|_2 \\
\leq &R^2 \cdot 100R \cdot \beta^{-2} n \exp(3R^2)\|x - y\|_2 \\
\leq &\beta^{-2} n \exp(20R^2)\|x - y\|_2
\end{aligned}$$

where the 1st step follows definition of G_i and matrix spectral norm, the 2nd step follows from $\|A\| \leq R$, the 3rd step follows from Lemma E.3, the 4th step follows from Lemma E.2, and the last step follows from simple algebra. □

E.2 A Core Tool: Lipschitz Property for Several Basic Functions

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

- *Let* $A \in \mathbb{R}^{n \times d}$
- *Let* $b \in \mathbb{R}^n$ *satisfy that* $\|b\|_1 \leq 1$
- *Let* $\beta \in (0, 0.1)$
- *Let* $R \geq 4$
- *Let* $x, y \in \mathbb{R}^d$ *satisfy* $\|A(x - y)\|_\infty < 0.01$
- $\|A\| \leq R$
- $\langle \exp(Ax), \mathbf{1}_n \rangle \geq \beta$
- $\langle \exp(Ay), \mathbf{1}_n \rangle \geq \beta$
- *Let* $R_f := \beta^{-2} n \exp(3R^2)$
- *Let* $\alpha(x)$ *be defined as Definition C.4*
- *Let* $c(x)$ *be defined as Definition C.5*
- *Let* $f(x)$ *be defined as Definition C.1*
- *Let* $g(x)$ *be defined as Definition C.7*

We have

- *Part 0.* $\| \exp(Ax)\|_2 \leq \sqrt{n} \exp(R^2)$
- *Part 1.* $\| \exp(Ax) - \exp(Ay)\|_2 \leq R \exp(R^2) \cdot \|x - y\|_2$
- *Part 2.* $|\alpha(x) - \alpha(y)| \leq \sqrt{n} \cdot \| \exp(Ax) - \exp(Ay)\|_2$
- *Part 3.* $|\alpha(x)^{-1} - \alpha(y)^{-1}| \leq \beta^{-2} \cdot |\alpha(x) - \alpha(y)|$
- *Part 4.* $\|f(x) - f(y)\|_2 \leq R_f \cdot \|x - y\|_2$
- *Part 5.* $\|c(x) - c(y)\|_2 \leq R_f \cdot \|x - y\|_2$
- *Part 6.* $\|g(x) - g(y)\|_2 \leq 18 \cdot R \cdot R_f \cdot \|x - y\|_2$

Proof. **Proof of Part 0.**

We can show that

$$\begin{aligned}\|\exp(Ax)\|_2 &\leq \sqrt{n} \cdot \|\exp(Ax)\|_\infty \\ &\leq \sqrt{n} \cdot \exp(\|Ax\|_\infty) \\ &\leq \sqrt{n} \cdot \exp(\|Ax\|_2) \\ &\leq \sqrt{n} \cdot \exp(R^2),\end{aligned}$$

where the first step follows from **Part 4** of Fact B.3, the second step follows from **Part6** Fact B.3, the third step follows from **Part** Fact B.3, and the last step follows from $\|A\| \leq R$ and $\|x\|_2 \leq R$.

Proof of Part 1. We have

$$\begin{aligned}\|\exp(Ax) - \exp(Ay)\|_2 \leq & \exp(R^2)\|Ax - Ay\|_2 \\ \leq & \exp(R^2)\|A\|\|x - y\|_2 \\ \leq & R\exp(R^2)\|x - y\|_2\end{aligned}$$

where the first step follows from **Part 10** of Fact B.3, the second step follows from **Part 4** of Fact B.4, the third step follows from $\|A\| \leq R$.

Proof of Part 2.

$$\begin{aligned}|\alpha(x) - \alpha(y)| &= |\langle \exp(Ax) - \exp(Ay), \mathbf{1}_n \rangle| \\ &\leq \|\exp(Ax) - \exp(Ay)\|_2 \cdot \sqrt{n}\end{aligned}$$

where the 1st step follows from the definition of $\alpha(x)$, the 2nd step follows from Cauchy-Schwarz inequality (**Part 1** of Fact B.3).

Proof of Part 3.

We can show that

$$\begin{aligned}|\alpha(x)^{-1} - \alpha(y)^{-1}| &= \alpha(x)^{-1}\alpha(y)^{-1} \cdot |\alpha(x) - \alpha(y)| \\ &\leq \beta^{-2} \cdot |\alpha(x) - \alpha(y)|\end{aligned}$$

where the 1st step follows from simple algebra, the 2nd step follows from $\alpha(x), \alpha(y) \geq \beta$.

Proof of Part 4.

We can show that

$$\begin{aligned}\|f(x) - f(y)\|_2 &= \|\alpha(x)^{-1}\exp(Ax) - \alpha(y)^{-1}\exp(Ay)\|_2 \\ &\leq \|\alpha(x)^{-1}\exp(Ax) - \alpha(x)^{-1}\exp(Ay)\|_2 + \|\alpha(x)^{-1}\exp(Ay) - \alpha(y)^{-1}\exp(Ay)\|_2 \\ &\leq \alpha(x)^{-1}\|\exp(Ax) - \exp(Ay)\|_2 + |\alpha(x)^{-1} - \alpha(y)^{-1}| \cdot \|\exp(Ay)\|_2\end{aligned}$$

where the 1st step follows from the definition of $f(x)$ and $\alpha(x)$, the 2nd step follows from triangle inequality (**Part 3** of Fact B.3), the 3rd step follows from $\|\alpha A\| \leq |\alpha|\|A\|$(**Part 5** of Fact B.4).

For the first term in the above, we have

$$\begin{aligned}\alpha(x)^{-1}\|\exp(Ax) - \exp(Ay)\|_2 &\leq \beta^{-1}\|\exp(Ax) - \exp(Ay)\|_2 \\ &\leq \beta^{-1} \cdot R\exp(R^2) \cdot \|x - y\|_2 \qquad (6)\end{aligned}$$

where the 1st step follows from $\alpha(x) \geq \beta$, the 2nd step follows from **Part 1**.

For the second term in the above, we have

$$\begin{aligned}|\alpha(x)^{-1} - \alpha(y)^{-1}| \cdot \|\exp(Ay)\|_2 &\leq \beta^{-2} \cdot |\alpha(x) - \alpha(y)| \cdot \|\exp(Ay)\|_2 \\ &\leq \beta^{-2} \cdot |\alpha(x) - \alpha(y)| \cdot \sqrt{n}\exp(R^2) \\ &\leq \beta^{-2} \cdot \sqrt{n} \cdot \|\exp(Ax) - \exp(Ay)\|_2 \cdot \sqrt{n}\exp(R^2) \\ &\leq \beta^{-2} \cdot \sqrt{n} \cdot R\exp(R^2)\|x - y\|_2 \cdot \sqrt{n}\exp(R^2)\end{aligned}$$

$$= \beta^{-2} \cdot nR \exp(2R^2) \|x - y\|_2 \tag{7}$$

where the 1st step follows from the result of **Part 3**, the 2nd step follows from **Part 0**, the 3rd step follows from the result of **Part 2**, the 4th step follows from **Part 1**, and the last step follows from simple algebra.

Combining Eq. (6) and Eq. (7) together, we have

$$\begin{aligned} \|f(x) - f(y)\|_2 \leq & \beta^{-1} \cdot R \exp(R^2) \cdot \|x - y\|_2 + \beta^{-2} \cdot nR \exp(2R^2) \|x - y\|_2 \\ \leq & 2\beta^{-2} nR \exp(2R^2) \|x - y\|_2 \\ \leq & \beta^{-2} n \exp(3R^2) \|x - y\|_2 \end{aligned}$$

where the 1st step follows from the bound of the first term and the second term, the 2nd step follows from $\beta^{-1} \geq 1$ and $n > 1$ trivially, the 3rd step follows from simple algebra.

Proof of Part 5. We have

$$\|c(x) - c(y)\|_2 = \|f(x) - f(y)\|_2 \leq R_f \cdot \|x - y\|_2,$$

the first step follows from the definition of $c(x)$, the last step follows from **Part 4** and definition of R_f. **Proof of Part 6.**

Using Lemma C.8, we have

$$\begin{aligned} \|g(x) - g(y)\|_2 \leq & 8\|A\| R_f \cdot 2\|x - y\|_2 \\ \leq & 18 R R_f \cdot \|x - y\|_2, \end{aligned}$$

where the second step follows from $\|A\| \leq R$.

Thus, we complete the proof. □

E.3 Summary of Eight Steps

Lemma E.3. *If the following conditions hold*

- $G_1 = \|f(x)\|_2^2 f(x) f(x)^\top - \|f(y)\|_2^2 f(y) f(y)^\top$
- $G_2 = \langle f(x), b \rangle f(x) f(x)^\top - \langle f(y), b \rangle f(y) f(y)^\top$
- $G_3 = \langle f(x), f(x) \rangle \operatorname{diag}(f(x)) - \langle f(y), f(y) \rangle \operatorname{diag}(f(y))$
- $G_4 = \langle f(x), b \rangle \operatorname{diag}(f(x)) - \langle f(y), b \rangle \operatorname{diag}(f(y))$
- $G_5 = \operatorname{diag}(f(x) \circ (f(x) - b)) - \operatorname{diag}(f(y) \circ (f(y) - b))$
- $G_6 = \operatorname{diag}(f(x) \circ f(x)) - \operatorname{diag}(f(y) \circ f(y))$
- $G_7 = f(x)(f(x) \circ b)^\top - f(y)(f(y) \circ b)^\top$
- $G_8 = (f(x) \circ b) f(x)^\top - (f(y) \circ b) f(y)^\top$

We have

$$\|G_1\| + \sum_{i=1}^{8} \|G_i\| \leq 100R \cdot \|f(x) - f(y)\|_2$$

Proof. The proof directly follows from applying Lemma E.4, Lemma E.5, Lemma E.6, Lemma E.7, Lemma E.8, Lemma E.9, Lemma E.10, Lemma E.11. □

E.4 Lipschitz Calculations: Step 1. Lipschitz for Matrix Function $\|f(x)\|_2^2 f(x) f(x)^\top$

Lemma E.4. *If the following condition holds*

- $G_1 = \|f(x)\|_2^2 f(x) f(x)^\top - \|f(y)\|_2^2 f(y) f(y)^\top$

Then

$$\|G_1\| \leq 4\|f(x) - f(y)\|_2$$

Proof. We define

$$\begin{aligned}
G_{1,1} &:= \langle f(x), f(x)\rangle f(x)f(x)^\top - \langle f(x), f(y)\rangle f(x)f(x)^\top \\
G_{1,2} &:= \langle f(x), f(y)\rangle f(x)f(x)^\top - \langle f(y), f(y)\rangle f(x)f(x)^\top \\
G_{1,3} &:= \langle f(y), f(y)\rangle f(x)f(x)^\top - \langle f(y), f(y)\rangle f(y)f(x)^\top \\
G_{1,4} &:= \langle f(y), f(y)\rangle f(y)f(x)^\top - \langle f(y), f(y)\rangle f(y)f(y)^\top
\end{aligned}$$

We have

$$G_1 = G_{1,1} + G_{1,2} + G_{1,3} + G_{1,4}$$

Let us only prove for $G_{1,1}$, the others are similar,

$$\begin{aligned}
\|G_{1,1}\| &\leq |\langle f(x), f(x) - f(y)\rangle| \cdot \|f(x)f(x)^\top\| \\
&\leq \|f(x)\|_2 \cdot \|f(x) - f(y)\|_2 \cdot \|f(x)f(x)^\top\| \\
&= \|f(x)\|_2 \cdot \|f(x) - f(y)\|_2 \cdot \|f(x)\|_2^2 \\
&\leq \|f(x) - f(y)\|_2
\end{aligned}$$

where the 1st step follows from Fact B.4, the 2nd step follows from $|\langle a, b\rangle| \leq \|a\|_2\|b\|_2$ (Fact B.3), the 3rd step follows from $aa^\top \preceq \|a\|_2^2 I_n$(Fact B.3), the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$ (Lemma C.2).

It is obvious that for each $i \in [4]$, we have

$$\begin{aligned}
\|G_{1,i}\| &\leq \|f(x) - f(y)\|_2 \max\{\|f(x)\|_2, \|f(y)\|_2\}^3 \\
&\leq \|f(x) - f(y)\|_2
\end{aligned}$$

where the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$. □

E.5 Lipschitz Calculations: Step 2. Lipschitz for Matrix Function $\langle f(x), b\rangle f(x)f(x)^\top$

Lemma E.5. *If the following condition holds*

- $G_2 := \langle f(x), b\rangle f(x)f(x)^\top - \langle f(y), b\rangle f(y)f(y)^\top$

Then we have

$$\|G_2\| \leq 3\|f(x) - f(y)\|_2 \cdot \|b\|_2$$

Proof. We define

$$\begin{aligned}
G_{2,1} &:= \langle f(x), b\rangle f(x)f(x)^\top - \langle f(x), b\rangle f(y)f(x)^\top \\
G_{2,2} &:= \langle f(x), b\rangle f(y)f(x)^\top - \langle f(x), b\rangle f(y)f(y)^\top \\
G_{2,3} &:= \langle f(x), b\rangle f(y)f(y)^\top - \langle f(y), b\rangle f(y)f(y)^\top
\end{aligned}$$

Then it's apparent that

$$G_2 = G_{2,1} + G_{2,2} + G_{2,3}$$

Since $G_{2,1}, G_{2,2}, G_{2,3}$ are similar, we only have to bound $\|G_{2,1}\|$:

$$\begin{aligned}
\|G_{2,1}\| &= \|\langle f(x), b\rangle f(x)f(x)^\top - \langle f(x), b\rangle f(y)f(x)^\top\| \\
&= \|\langle f(x), b\rangle (f(x) - f(y))f(x)^\top\| \\
&\leq |\langle f(x), b\rangle| \cdot \|(f(x) - f(y))f(x)^\top\| \\
&\leq |\langle f(x), b\rangle| \cdot \|f(x) - f(y)\|_2 \cdot \|f(x)\|_2
\end{aligned}$$

$$\leq \|f(x)\|_2^2 \cdot \|b\|_2 \|f(x) - f(y)\|_2$$
$$\leq \|f(x) - f(y)\|_2 \cdot \|b\|_2$$

where the 1st step follows from the definition of $G_{2,1}$, the 2nd step follows from simple algebra, the 3rd step follows from Fact B.4, the 4th step follows from $\|ab^\top\| \leq \|a\|_2 \|b\|_2$(Fact B.4), the 5th step follows from $\langle a, b \rangle \leq \|a\|_2 \|b\|_2$(Fact B.3), the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$.

Thus, we have

$$\|G_2\| \leq 3\|f(x) - f(y)\| \|b\|_2$$

□

E.6 Lipschitz Calculations: Step 3. Lipschitz for Matrix Function $f(x)f(x)^\top \operatorname{diag}(f(x))$

Lemma E.6. *If the following condition holds*

- $G_3 := \langle f(x), f(x) \rangle \operatorname{diag}(f(x)) - \langle f(y), f(y) \rangle \operatorname{diag}(f(y))$

Then we have

$$\|G_3\| \leq 3\|f(x) - f(y)\|_2$$

Proof. We define

$$G_{3,1} := \langle f(x), f(x) \rangle \operatorname{diag}(f(x)) - \langle f(x), f(y) \rangle \operatorname{diag}(f(x))$$
$$G_{3,2} := \langle f(x), f(y) \rangle \operatorname{diag}(f(x)) - \langle f(x), f(y) \rangle \operatorname{diag}(f(y))$$
$$G_{3,3} := \langle f(x), f(y) \rangle \operatorname{diag}(f(y)) - \langle f(y), f(y) \rangle \operatorname{diag}(f(y))$$

Thus, it's trivial that

$$G_3 = G_{3,1} + G_{3,2} + G_{3,3}$$

Since $G_{3,1}, G_{3,2}, G_{3,3}$ are similar, we only need to bound $\|G_{3,1}\|$:

$$\begin{aligned}
\|G_{3,1}\| &= \|\langle f(x), f(x) \rangle \operatorname{diag}(f(x)) - \langle f(x), f(y) \rangle \operatorname{diag}(f(x))\| \\
&= \|\langle f(x), f(x) - f(y) \rangle \operatorname{diag}(f(x))\| \\
&\leq \|f(x)^\top\|_2 \|f(x) - f(y)\|_2 \|\operatorname{diag}(f(x))\| \\
&= \|f(x)\|_2^2 \|f(x) - f(y)\|_2 \\
&\leq \|f(x) - f(y)\|_2
\end{aligned}$$

where the 1st step follows from the definition of $G_{3,1}$, the 2nd step follows from simple algebra, the 3rd step follows from $\|\alpha A\| \leq |\alpha| \|A\|$(Fact B.4), $\langle a, b \rangle \leq \|a\|_2 \|b\|_2$(Fact B.3), and $\|ab\| \leq \|a\| \|b\|$(Fact B.4), the 4th step follows from $\|\operatorname{diag}(f(x))\| = \|f(x)\|_2$, the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$ (Fact B.3).

Thus, we have

$$\begin{aligned}
\|G_8\| &= \|G_{8,1} + G_{8,2} + G_{3,3}\| \\
&\leq \|G_{8,1}\| + \|G_{8,2}\| + \|G_{3,3}\| \\
&= 3\|f(x) - f(y)\|_2
\end{aligned}$$

where the 1st step follows from the definition of G_3, the 2nd step follows from Fact B.4, the last step follows from the bound of $\|G_{3,1}\|$,$\|G_{3,2}\|$ and $\|G_{3,3}\|$. □

E.7 Lipschitz Calculations: Step 4. Lipschitz for Matrix Function $\langle f(x), b \rangle \operatorname{diag}(f(x))$

Lemma E.7. *If the following condition holds*

- $G_4 := \langle f(x), b \rangle \operatorname{diag}(f(x)) - \langle f(y), b \rangle \operatorname{diag}(f(y))$

Then we have

$$\|G_4\| \leq 2\|f(x) - f(y)\|_2 \|b\|_2$$

Proof. We define:

$$\begin{aligned} G_{4,1} :&= \langle f(x), b \rangle \operatorname{diag}(f(x)) - \langle f(y), b \rangle \operatorname{diag}(f(x)) \\ G_{4,2} :&= \langle f(y), b \rangle \operatorname{diag}(f(x)) - \langle f(y), b \rangle \operatorname{diag}(f(y)) \end{aligned}$$

Thus, it's trivial that

$$G_4 = G_{4,1} + G_{4,2}$$

Since $G_{4,1}$ and $G_{4,2}$ are similar, we only need to bound $\|G_{4,1}\|$:

$$\begin{aligned} \|G_{4,1}\| &= \|\langle f(x), b \rangle \operatorname{diag}(f(x)) - \langle f(y), b \rangle \operatorname{diag}(f(x))\| \\ &= \|b^\top (f(x) - f(y)) \operatorname{diag}(f(x))\| \\ &\leq \|b^\top\|_2 \|f(x) - f(y)\|_2 \|\operatorname{diag}(f(x)\| \\ &\leq \|b\|_2 \|f(x) - f(y)\|_2 \|f(x)\|_2 \\ &\leq \|f(x) - f(y)\|_2 \|b\|_2 \end{aligned}$$

where the 1st step follows from the definition of $G_{4,1}$, the 2nd step follows from simple algebra, the 3rd step follows from $\|ab\| \leq \|a\|\|b\|$(Fact B.4) and , the 4th step follows from $\|\operatorname{diag}(x)\| \leq \|x\|_\infty \leq \|x\|_2$(Fact B.3), the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$(Fact B.3).

Thus, we have

$$\begin{aligned} \|G_4\| &= \|G_{4,1} + G_{4,2}\| \\ &\leq \|G_{4,1}\| + \|G_{4,2}\| \\ &= 2\|f(x) - f(y)\|_2 \cdot \|b\|_2 \end{aligned}$$

where the 1st step follows from the definition of G_4, the 2nd step follows from Fact B.4, the last step follows from the bound of $\|G_{4,1}\|$ and $\|G_{4,2}\|$. □

E.8 Lipschitz Calculations: Step 5. Lipschitz for Matrix Function $\operatorname{diag}(f(x) \circ (f(x) - b))$

Lemma E.8. *If the following condition holds*

- $G_5 := \operatorname{diag}(f(x) \circ (f(x) - b)) - \operatorname{diag}(f(y) \circ (f(y) - b))$

Then we have

$$\|G_5\| \leq 2\|f(x) - f(y)\|_2 + \|f(x) - f(y)\| \cdot \|b\|_2$$

Proof. We define:

$$\begin{aligned} G_{5,1} :&= \operatorname{diag}(f(x) \circ (f(x) - b)) - \operatorname{diag}(f(x) \circ (f(y) - b)) \\ G_{5,2} :&= \operatorname{diag}(f(x) \circ (f(y) - b)) - \operatorname{diag}(f(y) \circ (f(y) - b)) \end{aligned}$$

Then, it's trivial that

$$G_5 = G_{5,1} + G_{5,2}$$

Bound $\|G_{5,1}\|$:

$$\begin{aligned} \|G_{5,1}\| &= \|\operatorname{diag}(f(x) \circ (f(x) - b)) - \operatorname{diag}(f(x) \circ (f(y) - b))\| \\ &= \|\operatorname{diag}(f(x)) \operatorname{diag}(f(x) - f(y))\| \\ &\leq \|\operatorname{diag}(f(x))\| \|\operatorname{diag}(f(x) - f(y))\| \\ &\leq \|f(x)\|_2 \|f(x) - f(y)\|_2 \end{aligned}$$

$$\leq \|f(x) - f(y)\|_2$$

where the 1st step follows from the definition of $G_{5,1}$, the 2nd step follows from Fact B.2, the 3rd step follows from $\|ab\| \leq \|a\|\|b\|$(Fact B.4), the 4th step follows from $\|\operatorname{diag}(a)\| \leq \|a\|_\infty \leq \|a\|_2$(Fact B.3), the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$.

Bound $\|G_{5,2}\|$:

$$\begin{aligned}
\|G_{5,2}\| &= \|\operatorname{diag}(f(x) \circ (f(y) - b)) - \operatorname{diag}(f(y) \circ (f(y) - b))\| \\
&= \|\operatorname{diag}(f(x) - f(y))\operatorname{diag}(f(y) - b)\| \\
&\leq \|\operatorname{diag}(f(x) - f(y))\|\|\operatorname{diag}(f(y) - b)\| \\
&\leq \|f(x) - f(y)\|_2\|f(y)\|_2 + \|f(x) - f(y)\|_2\|b\|_2 \\
&\leq \|f(x) - f(y)\|_2 + \|f(x) - f(y)\|_2\|b\|_2
\end{aligned}$$

where the 1st step follows from the definition of $G_{5,2}$, the 2nd step follows from Fact B.2, the 3rd step follows from Fact B.4, the 4th step follows from $\|\operatorname{diag}(a)\| \leq \|a\|_\infty \leq \|a\|_2$(Fact B.3), the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$.

Thus, we have

$$\begin{aligned}
\|G_5\| &= \|G_{5,1} + G_{5,2}\| \\
&\leq \|G_{5,1}\| + \|G_{5,2}\| \\
&\leq 2\|f(x) - f(y)\|_2 + \|f(x) - f(y)\|_2 \cdot \|b\|_2
\end{aligned}$$

where the 1st step follows from the definition of G_5, the 2nd step follows from Fact B.2, the 3rd step follows from the bound of $\|G_{5,1}\|$ and $\|G_{5,2}\|$. □

E.9 Lipschitz Calculations: Step 6. Lipschitz for Matrix Function $\operatorname{diag}(f(x) \circ f(x))$

Lemma E.9. *If the following condition holds*

- $G_6 := \operatorname{diag}(f(x) \circ f(x)) - \operatorname{diag}(f(y) \circ f(y))$

Then we have

$$\|G_6\| \leq 2\|f(x) - f(y)\|_2$$

Proof. We define:

$$\begin{aligned}
G_{6,1} &:= \operatorname{diag}(f(x) \circ f(x)) - \operatorname{diag}(f(x) \circ f(y)) \\
G_{6,2} &:= \operatorname{diag}(f(x) \circ f(y)) - \operatorname{diag}(f(y) \circ f(y))
\end{aligned}$$

Then, it's trivial that

$$G_6 = G_{6,1} + G_{6,2}$$

Since, $G_{6,1}$ and $G_{6,2}$ are similar, we only need to bound $\|G_{6,1}\|$:

$$\begin{aligned}
\|G_{6,1}\| &= \|\operatorname{diag}(f(x) \circ f(x)) - \operatorname{diag}(f(x) \circ f(y))\| \\
&= \|\operatorname{diag}(f(x))(\operatorname{diag}(f(x) - \operatorname{diag}(f(y)))\| \\
&\leq \|f(x)\|_2\|f(x) - f(y)\|_2 \\
&\leq \|f(x) - f(y)\|
\end{aligned}$$

where the 1st step follows from the definition of $G_{6,1}$, the 2nd step follows from Fact B.2, the 3rd step follows from $\|ab\| \leq \|a\|\|b\|$(Fact B.4) and $\|\operatorname{diag}(a)\| \leq \|a\|_\infty \leq \|a\|_2$(Fact B.3), the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$.

Thus, we have

$$\begin{aligned}
\|G_6\| &= \|G_{6,1} + G_{6,2}\| \\
&\leq \|G_{6,1}\| + \|G_{6,2}\| \\
&= 2\|f(x) - f(y)\|_2
\end{aligned}$$

where the 1st step follows from the definition of G_6, the 2nd step follows from Fact B.4, the last step follows from the bound of $\|G_{6,1}\|$ and $\|G_{6,2}\|$. □

E.10 Lipschitz Calculations: Step 7. Lipschitz for Matrix Function $f(x)(f(x) \circ b)^\top$

Lemma E.10. *If the following condition holds*

- $G_7 := f(x)(f(x) \circ b)^\top - f(y)(f(y) \circ b)^\top$

Then, we have

$$\|G_7\| \leq 2\|f(x) - f(y)\|_2 \cdot \|b\|_2$$

Proof. We define:

$$\begin{aligned} G_{7,1} :&= f(x)(f(x) \circ b)^\top - f(x)(f(y) \circ b)^\top \\ G_{7,2} :&= f(x)(f(y) \circ b)^\top - f(y)(f(y) \circ b)^\top \end{aligned}$$

Since $G_{7,1}$ and $G_{7,2}$ are similar, we only need to bound $\|G_{7,1}\|$:

$$\begin{aligned} \|G_{7,1}\| &= \|f(x)(f(x) \circ b)^\top - f(x)(f(y) \circ b)^\top\| \\ &= \|f(x)((f(x) - f(y)) \circ b)^\top\| \\ &\leq \|f(x)\|_2 \|(f(x) - f(y)) \circ b\|_2 \\ &\leq \|f(x) - f(y)\|_2 \|b\|_2 \end{aligned}$$

where the 1st step follows from the definition of $G_{7,1}$, the 2nd step follows from simple algebra, the 3rd step follows from $\|ab^\top\| \leq \|a\|_2 \|b\|_2$(Fact B.4) and $\|a^\top\|_2 = \|a\|_2$, the last step follows from $\|a \circ b\|_2 \leq \|a\|_\infty \|b\| \leq \|a\|_2 \|b\|_2$(Fact B.3) and $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$.

Thus, we have

$$\begin{aligned} \|G_7\| &= \|G_{7,1} + G_{7,2}\| \\ &\leq \|G_{7,1}\| + \|G_{7,2}\| \\ &= 2\|f(x) - f(y)\|_2 \cdot \|b\|_2 \end{aligned}$$

where the 1st step follows from the definition of G_7, the 2nd step follows from Fact B.4, the last step follows from the bound of $\|G_{7,1}\|$ and $\|G_{7,2}\|$. □

E.11 Lipschitz Calculations: Step 8. Lipschitz for Matrix Function $(f(x) \circ b)f(x)^\top$

Lemma E.11. *If the following condition holds*

- $G_8 := (f(x) \circ b)f(x)^\top - (f(y) \circ b)f(y)^\top$

Then we have

$$\|G_8\| \leq 2\|f(x) - f(y)\|_2 \cdot \|b\|_2$$

Proof. We define:

$$\begin{aligned} G_{8,1} :&= (f(x) \circ b)f(x)^\top - (f(x) \circ b)f(y)^\top \\ G_{8,2} :&= (f(x) \circ b)f(y)^\top - (f(y) \circ b)f(y)^\top \end{aligned}$$

Then, it's trivial that

$$G_8 = G_{8,1} + G_{8,2}$$

Since $G_{8,1}$ and $G_{8,2}$ are similar, we only need to bound $\|G_{8,1}\|$:

$$\begin{aligned} \|G_{8,1}\| &= \|(f(x) \circ b)f(x)^\top - (f(x) \circ b)f(y)^\top\| \\ &= \|(f(x) \circ b)(f(x) - f(y))^\top\| \end{aligned}$$

$$\begin{aligned}
&\leq \|f(x) \circ b\|_2 \|f(x) - f(y)\|_2 \\
&\leq \|f(x)\|_2 \|b\|_2 \|f(x) - f(y)\|_2 \\
&\leq \|f(x) - f(y)\|_2 \|b\|_2
\end{aligned}$$

where the 1st step follows from the definition of $G_{8,1}$, the 2nd step follows from simple algebra, the 3rd step follows from $\|ab^\top\| \leq \|a\|_2 \|b\|_2$(Fact B.4), the 4th step follows from $\|a \circ b\|_2 \leq \|a\|_\infty \|b\| \leq \|a\|_2 \|b\|_2$(Fact B.3), the last step follows from $\|f(x)\|_2 \leq \|f(x)\|_1 \leq 1$ (Lemma C.2).

Thus, we have

$$\begin{aligned}
\|G_8\| &= \|G_{8,1} + G_{8,2}\| \\
&\leq \|G_{8,1}\| + \|G_{8,2}\| \\
&= 2\|f(x) - f(y)\|_2 \cdot \|b\|_2
\end{aligned}$$

where the 1st step follows from the definition of G_8, the 2nd step follows from Fact B.4, the last step follows from the bound of $\|G_{8,1}\|$ and $\|G_{8,2}\|$. □

F APPROXIMATE NEWTON METHOD

In this section, we provide an approximate version of the newton method for convex optimization. In Section F.1, we state some assumptions of the traditional newton method and the exact update rule of the traditional algorithm. In Section F.2, we provide the approximate update rule of the approximate newton method, we also implement a tool for compute the approximation of $\nabla^2 L$ and use some lemmas from Li et al. (2023b) to analyze the approximate newton method. In Section F.3, we find a lower bound of $\alpha(x)$. In Section F.4, we use a Lemma from previous Sections to find a upper bound of M.

F.1 Definition and Update Rule

Here in this section, we focus on the local convergence of the Newton method. We consider the following target function

$$\min_{x \in \mathbb{R}^d} L(x)$$

with these assumptions:

Definition F.1 ((l, M)-good Loss function)**.** *For a function $L : \mathbb{R}^d \to \mathbb{R}$, we say L is (l, M)-good it satisfies the following conditions,*

- **l-local Minimum.** *We define $l > 0$ to be a positive scalar. If there exists a vector $x^* \in \mathbb{R}^d$ such that the following holds*
  - $\nabla L(x^*) = \mathbf{0}_d$.
  - $\nabla^2 L(x^*) \succeq l \cdot I_d$.
- **Hessian is M-Lipschitz.** *If there exists a positive scalar $M > 0$ such that*

$$\|\nabla^2 L(y) - \nabla^2 L(x)\| \leq M \cdot \|y - x\|_2$$

- **Good Initialization Point.** *Let x_0 denote the initialization point. If $r_0 := \|x_0 - x_*\|_2$ satisfies*

$$r_0 M \leq 0.1 l$$

We define gradient and Hessian as follows

Definition F.2 (Gradient and Hessian)**.** *The gradient $g : \mathbb{R}^d \to \mathbb{R}^d$ of the loss function is defined as*

$$g(x) := \nabla L(x)$$

The Hessian $H : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ of the loss function is defined as,

$$H(x) := \nabla^2 L(x)$$

With the gradient function $g : \mathbb{R}^d \to \mathbb{R}^d$ and the Hessian matrix $H : \mathbb{R}^d \to \mathbb{R}^{d \times d}$, we define the exact process of the Newton method as follows:

Definition F.3 (Exact update of the Newton method)**.**

$$x_{t+1} = x_t - H(x_t)^{-1} \cdot g(x_t)$$

F.2 Approximate of Hessian and Update Rule

In many real-world tasks, it is very hard and expensive to compute exact $\nabla^2 L(x_t)$ or $(\nabla^2 L(x_t))^{-1}$. Thus, it is natural to consider the approximated computation of the gradient and Hessian. The computation is defined as

Definition F.4 (Approximate Hessian)**.** *For any Hessian $H(x_t) \in \mathbb{R}^{d \times d}$, we define the approximated Hessian $\widetilde{H}(x_t) \in \mathbb{R}^{d \times d}$ to be a matrix such that the following holds,*

$$(1 - \epsilon_0) \cdot H(x_t) \preceq \widetilde{H}(x_t) \preceq (1 + \epsilon_0) \cdot H(x_t).$$

In order to get the approximated Hessian $\widetilde{H}(x_t)$ efficiently, here we state a standard tool (see Lemma 4.5 in Deng et al. (2022)).

Lemma F.5 (Deng et al. (2022); Song et al. (2022))**.** *Let $\epsilon_0 = 0.01$ be a constant precision parameter. Let $A \in \mathbb{R}^{n \times d}$ be a real matrix, then for any positive diagonal (PD) matrix $D \in \mathbb{R}^{n \times n}$, there exists an algorithm which runs in time*

$$O((\mathrm{nnz}(A) + d^\omega) \, \mathrm{poly}(\log(n/\delta)))$$

and it outputs an $O(d \log(n/\delta))$ sparse diagonal matrix $\widetilde{D} \in \mathbb{R}^{n \times n}$ for which

$$(1 - \epsilon_0) A^\top D A \preceq A^\top \widetilde{D} A \preceq (1 + \epsilon_0) A^\top D A.$$

Note that, ω denotes the exponent of matrix multiplication, currently $\omega \approx 2.373$ Williams (2012); Le Gall (2014); Alman & Williams (2021).

Following the standard of Approximate Newton Hessian literature Anstreicher (2000); Jiang et al. (2020a); Brand et al. (2021); Song et al. (2021); Huang et al. (2022); Li et al. (2023b), we consider the following.

Definition F.6 (Approximate update)**.** *We consider the following process*

$$x_{t+1} = x_t - \widetilde{H}(x_t)^{-1} \cdot g(x_t).$$

We state a tool from prior work,

Lemma F.7 (Iterative shrinking Lemma, Lemma 6.9 on page 32 of Li et al. (2023b))**.** *If the following condition hold*

- *Loss Function L is (l, M)-good (see Definition F.1).*
- *Let $\epsilon_0 \in (0, 0.1)$ (see Definition F.4).*
- *Let $r_t := \|x_t - x^*\|_2$.*
- *Let $\overline{r}_t := M \cdot r_t$*

Then we have

$$r_{t+1} \leq 2 \cdot (\epsilon_0 + \overline{r}_t / (l - \overline{r}_t)) \cdot r_t.$$

Let T denote the total number of iterations of the algorithm, to apply Lemma F.7, we will need the following induction hypothesis lemma. This is very standard in the literature, see Li et al. (2023b).

Lemma F.8 (Induction hypothesis, Lemma 6.10 on page 34 of Li et al. (2023b))**.** *For each $i \in [t]$, we define $r_i := \|x_i - x^*\|_2$. If the following condition hold*

- $\epsilon_0 = 0.01$ *(see Definition F.4 for ϵ_0)*

- $r_i \leq 0.4 \cdot r_{i-1}$, *for all* $i \in [t]$
- $M \cdot r_i \leq 0.1l$, *for all* $i \in [t]$ *(see Definition F.1 for* M*)*

Then we have

- $r_{t+1} \leq 0.4 r_t$
- $M \cdot r_{t+1} \leq 0.1l$

F.3 Lower bound on β

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

- $\|A\| \leq R$
- $\|x\|_2 \leq R$
- *Let* β *be lower bound on* $\langle \exp(Ax), \mathbf{1}_n \rangle$

Then we have

$$\beta \geq \exp(-R^2)$$

Proof. We have

$$\begin{aligned} \langle \exp(Ax), \mathbf{1}_n \rangle \geq & \max_{i \in [n]} \exp(-|(Ax)_i|) \\ \geq & \exp(-\|Ax\|_\infty) \\ \geq & \exp(-\|Ax\|_2) \\ \geq & \exp(-R^2) \end{aligned}$$

the 1st step follows from simple algebra, the 2nd step follows from definition of ℓ_∞ norm, the 3rd step follows from Fact B.3.

□

F.4 Upper bound on M

Lemma F.10. *If the following conditions holds*

- $\|A\| \leq R$.
- $\|x\|_2 \leq R$.
- *Let* H *denote the hessian of loss function* L.
- $\|H(x) - H(y)\| \leq \beta^{-2} n^{1.5} \exp(20R^2) \cdot \|x - y\|_2$ *(Lemma E.1)*

Then, we have

$$M \leq n^{1.5} \exp(30R^2).$$

Proof. It follows from Lemma F.9. □

G Limitation

This work primarily develops a theoretical framework for softmax regression influenced by attention mechanisms in large language models. The focus on theory means that the absence of empirical validation leaves important questions unanswered about the practical effectiveness and robustness of the proposed methods. Real-world applications could exhibit behaviors not predicted by the theoretical models, especially under conditions of data variability and deviations from the model

assumptions that fall outside the scope of this study. The effectiveness of the proposed algorithms is contingent on several assumptions made about the data and model structure, such as sparsity, distribution characteristics, and the relationships between variables. These assumptions might not always hold in practical scenarios, potentially limiting the generalizability and real-world applicability of the results.

In terms of complexity and scalability, while the proposed algorithms are designed to be computationally efficient in theory, their ability to scale effectively when handling extremely large datasets or high-dimensional problems remains unexplored. Practical deployment could face challenges such as memory constraints or performance bottlenecks that were not apparent in the theoretical analysis. Additionally, the findings are specifically tailored to softmax regression problems, and this narrow focus may restrict the direct applicability of the results to other types of regression or machine learning areas that do not use softmax or similar functions. This limitation suggests that further research would be needed to adapt or extend these methods to broader contexts, ensuring their relevance across a wider range of applications.