APPENDIX

Roadmap.

We order the appendix as follows: In Section A, we provide the preliminaries to be used in our proofs, such facts for basic algebras and inequalities. In Section B, we compute the gradient for our loss function step by step and reform it for prooving its lipschitz. In Section C we prove that the gradient for our loss function is lipschitz. In Section D, we compute the gradient of our loss function with respect to Q and proved the lipschitz property for gradient. In Section E, we repeat the analysis for Q and proved lipschitz property for gradient with respect to K. In Section F we provide systematic analysis on logistic function and proved the lipschitz property for the gradient of the loss function based on logistic function. In Section G, we prove our main results. In Section H, we provide a brief analysis on the hessian of our loss function.

A PRELIMINARY

In this section, we provide the preliminaries to be used in our proofs. In Section A.1, we provide some facts for exact computations. In Section A.2, we provide some inequalities with respect to vector's norms. In Section A.3, we provide some inequalities with respect to matrix's norms. In Section A.4, we provide some facts for computing gradient.

A.1 BASIC ALGEBRAS

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

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

A.2 BASIC VECTOR NORM BOUNDS

Fact A.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 $u, v \in \mathbb{R}^d$ 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 A.3, the 3rd step follows from $|\exp(x) - 1| \leq 2x$ for all $x \in (0, 0.1)$. □

A.3 Basic Matrix Norm Bounds

Fact A.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* $\|uv^\top\| \leq \|u\|_2 \|v\|_2$

Fact A.5. *If $\|Q\|_F \leq R$, then* $\|Qe_{i_2} e_{k_2}^\top\|_F = \mathrm{vec}(e_{i_2} e_{k_2} Q) \leq R$.

If $\|K\|_F \leq R$, then $\|e_{i_2} e_{k_2}^\top K^\top\|_F = \|\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_2 \leq R$

If $\|Q\|_F \leq R$, $\|K\|_F \leq R$, then $\|QK\|_F \leq R^2$

A.4 Basic Calculus

Fact A.6.

$$\frac{\mathrm{d}^2 A(x) B(x)}{\mathrm{d}s \mathrm{d}t} = \frac{\mathrm{d}^2 A(x)}{\mathrm{d}s \mathrm{d}t} B(x) + \frac{\mathrm{d}A(x)}{\mathrm{d}s} \frac{\mathrm{d}B(x)}{\mathrm{d}t} + \frac{\mathrm{d}A(x)}{\mathrm{d}t} \frac{\mathrm{d}B(x)}{\mathrm{d}s} + \frac{\mathrm{d}^2 B(x)}{\mathrm{d}s \mathrm{d}t} A(x)$$

Proof.

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

where the first step is an expansion of hessian, the second step follows from differential chain rule, the last step follows from differential chain rule. □

Fact A.7. *Let $A(x) \in \mathbb{R}$.*

$$\frac{\mathrm{d}^2 A(x)^2}{\mathrm{d}t \mathrm{d}t} = 2A(x)\frac{\mathrm{d}^2 A(x)}{\mathrm{d}t^2} + 2(\frac{\mathrm{d}A(x)}{\mathrm{d}t})^2$$

Proof.

$$\begin{aligned}\frac{\mathrm{d}^2 A(x)^2}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}(\frac{\mathrm{d}A(x)^2}{\mathrm{d}t}) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}(2A(x)\frac{\mathrm{d}A(x)}{\mathrm{d}t}) \\ &= 2A(x)\frac{\mathrm{d}^2 A(x)}{\mathrm{d}t^2} + 2(\frac{\mathrm{d}A(x)}{\mathrm{d}t})^2\end{aligned}$$

where the first step is an expansion of hessian, the second step follows from basic derivative, the third step follows from differential chain rule. □

Fact A.8. *Let $A(x) \in \mathbb{R}$, then we have*

$$\frac{\mathrm{d}^2 A^2(x)}{\mathrm{d}s \mathrm{d}t} = 2\frac{\mathrm{d}A(x)}{\mathrm{d}s}\frac{\mathrm{d}A(x)}{\mathrm{d}t} + 2A(x)\frac{\mathrm{d}^2 A(x)}{\mathrm{d}s \mathrm{d}t}$$

Proof. We can show

$$\begin{aligned}\frac{\mathrm{d}^2 A^2(x)}{\mathrm{d}s \mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}A^2(x)}{\mathrm{d}s} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}(2A(x)\frac{\mathrm{d}A(x)}{\mathrm{d}s}) \\ &= 2\frac{\mathrm{d}A(x)}{\mathrm{d}t}\frac{\mathrm{d}A(x)}{\mathrm{d}s} + 2A(x)\frac{\mathrm{d}}{\mathrm{d}t}(\frac{\mathrm{d}A(x)}{\mathrm{d}s}) \\ &= 2\frac{\mathrm{d}A(x)}{\mathrm{d}s}\frac{\mathrm{d}A(x)}{\mathrm{d}t} + 2A(x)\frac{\mathrm{d}^2 A(x)}{\mathrm{d}s \mathrm{d}t}\end{aligned}$$

where the first step is an expansion of hessian, the second step follows from basic derivative, the third step follows from differential chain rule, the last step follows from simple algebra. □

B Gradient Computation

In this section, we compute the gradient for our loss function step by step. In Section B.1, we define the definitions to be used in this section and the problem we would like to address in this section. In Section B.2, we compute the gradient with respect to x step by step. In Section B.3, we compute the gradient with respect to y step by step. In Section B.4, we reform the gradient with respect to x for the convenience of proving its lipschitz property in Section C.

B.1 Problem Formulation

Definition B.1. *We define $c(x,y)_{l_0,j_0,i_0} \in \mathbb{R}$ as follows*

$$c(x,y)_{l_0,j_0,i_0} := \langle \underbrace{f(x)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle - b_{l_0,j_0,i_0}$$

Definition B.2. *If the following conditions hold*

- *Let c be defined as Definition B.1*

For each $l_0 \in [m]$, $j_0 \in [n]$, $i_0 \in [d]$. We define L_{l_0,j_0,i_0} as follows

$$L(x,y)_{l_0,j_0,i_0} := 0.5c(x,y)^2_{l_0,j_0,i_0}$$

Definition B.3. *The final loss is*

$$L(x,y) := \sum_{l_0=1}^{m} \sum_{j_0=1}^{n} \sum_{i_0=1}^{d} L_{l_0,j_0,i_0}(x,y).$$

Not hard to see that $L(x,y)$ is equivalent

$$\|D(X)^{-1} \exp(A_1 X A_2^\top) A_3 Y - B\|_F^2 \tag{1}$$

Let $X \in \mathbb{R}^{d \times d}$ denote the matrix view of $x \in \mathbb{R}^{d^2}$. Here X can be viewed as $QK^\top$ in attention computation. Let $y_{i_0} \in \mathbb{R}^d$ denote the i_0-th column of $Y \in \mathbb{R}^{d \times d}$.

By using well-known tensor-trick, we can rewrite Eq. (1) in the following vector version

$$\|\mathrm{mat}(D(x)^{-1} \exp(\mathsf{A}\, x)) \underbrace{A_3}_{n \times d} \underbrace{Y}_{d \times d} - B\|_2^2$$

Here the diagonal matrix $D(x) \in \mathbb{R}^{n^2 \times n^2}$ can be written as $D(x) := D(X) \otimes I_n$

We give our formal definition of the optimization formulation

Definition B.4. *Let $A_1, A_2 \in \mathbb{R}^{n \times d}$. Let $X \in \mathbb{R}^{d \times d}$ denote the matrix view of $x \in \mathbb{R}^{d^2}$. We define the optimization formulation as the following:*

$$\min_{X \in \mathbb{R}^{d \times d}} L(X) = \min_{X \in \mathbb{R}^{d \times d}} \|D(X)^{-1} \exp(A_1 X A_2^\top) A_3 Y - B\|_2^2$$

Definition B.5. *Let $A_1, A_2 \in \mathbb{R}^{n \times d}$. Let $\mathsf{A} = A_1 \otimes A_2 \in \mathbb{R}^{n^2 \times d^2}$. Let $X \in \mathbb{R}^{d \times d}$ denote the matrix view of $x \in \mathbb{R}^{d^2}$. Let $D(x) \in \mathbb{R}^{n^2 \times n^2}$ denote the diagonal matrix $D(x) := D(X) \otimes I_n$. We define the vector version of optimization formulation as the following:*

$$\min_{x \in \mathbb{R}^{d^2}} L(x) = \min_{x \in \mathbb{R}^{d^2}} \|\mathrm{mat}(D(x)^{-1} \exp(\mathsf{A}\, x)) A_3 Y - B\|_2^2$$

B.2 Gradient Computation with respect to x

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

- *Let f be defined in Definition 3.3*
- *Let h be defined in Definition 3.4*
- *Let α be defined in Definition 3.2*
- *Let c be defined in Definition B.1*
- *Let L be defined in Definition B.3*

Then, we can show

- *Part 1. For each $i \in [d^2]$*

$$\frac{\mathrm{d}\, \mathsf{A}_{l_0,j_0}\, x}{\mathrm{d} x_i} = \mathsf{A}_{l_0,j_0,i}$$

- *Part 2. For each $i \in [d^2]$*

$$\frac{\mathrm{d} \exp(\mathsf{A}_{l_0,j_0} x)}{\mathrm{d} x_i} = \exp(\mathsf{A}_{l_0,j_0} x) \circ \mathsf{A}_{l_0,j_0,i}$$

- *Part 4. For $i \in [d^2]$*

$$\frac{\mathrm{d} u(x)_{l_0,j_0}}{\mathrm{d} x_i} = u(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}$$

- *Part 5. For $i \in [d^2]$*

$$\frac{\mathrm{d}\alpha(x)_{l_0,j_0}}{\mathrm{d}x_i} = \langle u(x)_{l_0,j_0}, A_{l_0,j_0,i} \rangle$$

- *Part 6. For $i \in [d^2]$*

$$\frac{\mathrm{d}\alpha(x)^{-1}_{l_0,j_0}}{\mathrm{d}x_i} = -\alpha(x)^{-1}_{l_0,j_0} \cdot \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle$$

- *Part 7. For each $i \in [d^2]$*

$$\frac{\mathrm{d}f(x)_{l_0,j_0}}{\mathrm{d}x_i} = f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} + f(x)_{l_0,j_0} \cdot \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle$$

- *Part 8. For $i \in [d^2]$*

$$\frac{\mathrm{d}c(x)_{l_0,j_0,i_0}}{\mathrm{d}x_i} = \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle$$

(This is similar to **Part 5** *of Lemma 5.1 in page 16 of Gao et al. (2023a))*

- *Part 9. For each $i \in [d^2]$,*

$$\frac{\mathrm{d}L_{l_0,j_0,i_0}(x,y)}{\mathrm{d}x_i} = c(x,y)_{l_0,j_0,i_0}(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle)$$

(This is similar to **Part 6** *of Lemma 5.1 in page 16 of Gao et al. (2023a))*

- *For each $i \in [d^2]$,*

$$\frac{\mathrm{d}L(x,y)}{\mathrm{d}x_i} = \sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d c(x,y)_{l_0,j_0,i_0}(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle)$$

Proof. **Proof of Part 1.**

We have

$$\frac{\mathrm{d}(\mathsf{A}_{l_0,j_0} x)}{\mathrm{d}x_i} = \mathsf{A}_{l_0,j_0,i}$$

this follows from simple algebra.

Proof of Part 2.

We have

$$\begin{aligned} \frac{\mathrm{d}\exp(\mathsf{A}_{l_0,j_0} x)}{\mathrm{d}x_i} = & \exp(\mathsf{A}_{l_0,j_0} x) \circ \frac{\mathrm{d}(\mathsf{A}_{l_0,j_0} x)}{\mathrm{d}x_i} \\ = & \exp(\mathsf{A}_{l_0,j_0} x) \circ \mathsf{A}_{l_0,j_0,i} \end{aligned}$$

where the first step follows from differential chain rule, the second step follows from **Part 1**.

Proof of Part 4. We have

$$\begin{aligned} \frac{\mathrm{d}u(x)_{l_0,j_0}}{\mathrm{d}x_i} = & \frac{\mathrm{d}\exp(\mathsf{A}_{l_0,j_0} x)}{\mathrm{d}x_i} \\ = & u(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \end{aligned}$$

where the first step follows from the definition of $u(x)_{l_0,j_0}$, the second step follows from basic calculus.

Proof of Part 5.

Let $j_0 \in [n]$. Let $i \in [d^2]$.

We have

$$\begin{aligned}\frac{\mathrm{d}\alpha(x)_{l_0,j_0}}{\mathrm{d}x_i} = & \frac{\mathrm{d}\langle \exp(\mathsf{A}_{l_0,j_0}x), \mathbf{1}_n\rangle}{\mathrm{d}x_i} \\ = & \langle \frac{\mathrm{d}\exp(\mathsf{A}_{l_0,j_0}x)}{\mathrm{d}x_i}, \mathbf{1}_n\rangle \\ = & \langle \exp(\mathsf{A}_{l_0,j_0}x) \circ (\mathsf{A}_{l_0,j_0,i}), \mathbf{1}_n\rangle \\ = & \langle u(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle\end{aligned}$$

where the first step follows from the definition of $\alpha(x)_{l_0,j_0}$, the second step follows from simple algebra, the third step follows from **Part 2**, the last step follows from Fact A.1.

Proof of Part 6. We have

$$\begin{aligned}\frac{\mathrm{d}\alpha(x)^{-1}_{l_0,j_0}}{\mathrm{d}x_i} = & -1 \cdot \alpha(x)^{-2}_{l_0,j_0} \cdot \frac{\mathrm{d}\alpha(x)_{l_0,j_0}}{\mathrm{d}x_i} \\ = & -\alpha(x)^{-1}_{l_0,j_0} \cdot \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle\end{aligned}$$

where the first step follows from differential chain rule, the second step follows from **Part 5**.

Proof of Part 7.

We have

$$\begin{aligned}\frac{\mathrm{d}f(x)_{l_0,j_0}}{\mathrm{d}x_i} = & \frac{\mathrm{d}(\alpha(x)^{-1}_{l_0,j_0}u(x)_{l_0,j_0})}{\mathrm{d}x_i} \\ = & \alpha(x)^{-1}_{l_0,j_0} \cdot \frac{\mathrm{d}u(x)_{l_0,j_0}}{\mathrm{d}x_i} + \frac{\mathrm{d}\alpha(x)^{-1}_{l_0,j_0}}{\mathrm{d}x_i}u(x)_{l_0,j_0} \\ = & \alpha(x)^{-1}_{l_0,j_0} \cdot u(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} + \frac{\mathrm{d}\alpha(x)^{-1}_{l_0,j_0}}{\mathrm{d}x_i}u(x)_{l_0,j_0} \\ = & \alpha(x)^{-1}_{l_0,j_0} \cdot u(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} - \alpha(x)^{-1}_{l_0,j_0} \cdot \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle \cdot u(x)_{l_0,j_0} \\ = & f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} - f(x)_{l_0,j_0} \cdot \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle\end{aligned}$$

where the first step follows from Definition 3.3, the second step follows from differential chain rule, the third step follows from **Part 4**, the fourth step follows from **Part 6**, the last step follows from definition of function f.

Proof of Part 8.

$$\begin{aligned}\frac{\mathrm{d}c(x)_{l_0,j_0,i_0}}{\mathrm{d}x_i} = & \frac{\mathrm{d}}{\mathrm{d}x_i}(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle - b_{l_0,j_0,i_0}) \\ = & \frac{\mathrm{d}}{\mathrm{d}x_i}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle \\ = & \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0}\rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle\langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle\end{aligned}$$

where the first step follows from the definition of $c(x)$, the second step follows from simple algebra and the last step follows from **Part 7**.

Proof of Part 9.

$$\begin{aligned}\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}x_i} = & \frac{\mathrm{d}}{\mathrm{d}x_i}0.5c(x,y)^2_{l_0,j_0,i_0} \\ = & c(x,y)_{l_0,j_0,i_0}\frac{\mathrm{d}}{\mathrm{d}x_i}c(x,y)_{l_0,j_0,i_0} \\ = & c(x,y)_{l_0,j_0,i_0}(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0}\rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle\langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle)\end{aligned}$$

where the first step follows from the definition of $L_{l_0,j_0,i_0}(x,y)$, the second step follows from simple algebra, the third step follows from **Part 8**. □

Proof of Part 10

$$\frac{\mathrm{d}L(x,y)}{\mathrm{d}x_i} = \sum_{l_0=1}^{m}\sum_{j_0=1}^{n}\sum_{i_0=1}^{d} c(x,y)_{l_0,j_0,i_0}(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0}\rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle\langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle)$$

This trivially follows from **Part 9**.

B.3 Gradient Computation with respect to y

Lemma B.7. *If the following conditions hold*

- *Let f be defined in Definition 3.3*
- *Let h be defined in Definition 3.4*
- *Let α be defined in Definition 3.2*
- *Let c be defined in Definition B.1*
- *Let L be defined in Definition B.3*

For $i_1 \in [d], i_0 \in [d]$, $i_2 \in [d]$ we have

- *Part 1.* $i_0 = i_1$

$$\frac{\mathrm{d}h(y)_{l_0,i_0}}{\mathrm{d}y_{i_1,i_2}} = \underbrace{A_{l_0,3,i_2}}_{n\times 1}$$

  where y_{i_1,i_2} is the i_2-th entry in vector $y_{i_1} \in \mathbb{R}^d$

- *Part 2.* $i_0 \neq i_1$

$$\frac{\mathrm{d}h(y)_{l_0,i_0}}{\mathrm{d}y_{i_1,i_2}} = \mathbf{0}_n$$

- *Part 3.* $i_0 = i_1$

$$\frac{\mathrm{d}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle}{\mathrm{d}y_{i_1,i_2}} = \langle f(x)_{l_0,j_0}, A_{l_0,3,i_2}\rangle$$

- *Part 4.* $i_0 \neq i_1$

$$\frac{\mathrm{d}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle}{\mathrm{d}y_{i_1,i_2}} = 0$$

- *Part 5.* $i_0 = i_1$

$$\frac{\mathrm{d}c(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} = \langle f(x)_{l_0,j_0}, A_{l_0,3,i_2}\rangle$$

- *Part 6.* $i_0 \neq i_1$

$$\frac{\mathrm{d}c(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} = 0$$

- *Part 7.* $i_0 = i_1$

$$\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} = c(x,y)_{l_0,j_0,i_0}\langle f(x)_{l_0,j_0}, A_{l_0,3,i_2}\rangle$$

- *Part 8.* $i_0 \neq i_1$

$$\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} = 0$$

Proof. **Proof of Part 1.** For $\forall i_1 \in [d], i_2 \in [d]$,

$$\begin{aligned}\frac{\mathrm{d}h(y)_{l_0,i_0}}{\mathrm{d}y_{i_1,i_2}} &= \frac{\mathrm{d}}{\mathrm{d}y_{i_1,i_2}} A_{l_0,3} y_{i_0} \\ &= \underbrace{A_{l_0,3,i_2}}_{n\times 1}\end{aligned}$$

where the first step follows from simple calculus.

Proof of Part 2.

For $i_1 \neq i_0$,

$$\begin{aligned}\frac{\mathrm{d}h(y)_{l_0,i_0}}{\mathrm{d}y_{i_1,i_2}} &= \frac{\mathrm{d}}{\mathrm{d}y_{i_1,i_2}} A_{l_0,3} y_{i_0} \\ &= \underbrace{\mathbf{0}_n}_{n\times 1}\end{aligned}$$

Proof of Part 3

$$\begin{aligned}\frac{\mathrm{d}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle}{\mathrm{d}y_{i_1,i_2}} &= \langle f(x)_{l_0,j_0}, \frac{h(y)_{l_0,i_0}}{\mathrm{d}y_{i_1,i_2}}\rangle \\ &= \langle f(x)_{l_0,j_0}, A_{l_0,3,i_2}\rangle\end{aligned}$$

where the first step follows from simple algebra, the second step follows from **Part 1**.

Proof of Part 4

$$\begin{aligned}\frac{\mathrm{d}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle}{\mathrm{d}y_{i_1,i_2}} &= \langle f(x)_{l_0,j_0}, \frac{h(y)_{l_0,i_0}}{\mathrm{d}y_{i_1,i_2}}\rangle \\ &= 0\end{aligned}$$

where the first step follows from simple algebra, the second step follows from **Part 2**.

Proof of Part 5

$$\begin{aligned}\frac{\mathrm{d}c(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} &= \frac{\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle - b_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} \\ &= \frac{\mathrm{d}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle}{\mathrm{d}y_{i_1,i_2}} \\ &= \langle f(x)_{l_0,j_0}, A_{l_0,3,i_2}\rangle\end{aligned}$$

where the first step follows from the definition of $c(x,y)_{l_0,j_0,i_0}$, the second step follows from simple algebra, the third step follows from **Part 3**.

Proof of Part 6

$$\begin{aligned}\frac{\mathrm{d}c(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} &= \frac{\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle - b_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} \\ &= \frac{\mathrm{d}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle}{\mathrm{d}y_{i_1,i_2}} \\ &= \mathbf{0}_n\end{aligned}$$

where the first step follows from simple algebra, the second step follows from simple algebra, the third step follows from **Part 4**.

Proof of Part 7

$$\begin{aligned}\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} &= \frac{\mathrm{d}0.5c(x,y)^2_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} \\ &= c(x,y)_{l_0,j_0,i_0}\frac{\mathrm{d}c(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} \\ &= c(x,y)_{l_0,j_0,i_0}\langle f(x)_{l_0,j_0}, A_{l_0,3,i_2}\rangle\end{aligned}$$

where the first step follows from simple algebra, the second step follows from simple algebra, the third step follows from **Part 5**.

Proof of Part 8

$$\begin{aligned}\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} &= \frac{\mathrm{d}0.5c(x,y)^2_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} \\ &= c(x,y)_{l_0,j_0,i_0}\frac{\mathrm{d}c(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} \\ &= \mathbf{0}_n\end{aligned}$$

where the first step follows from simple algebra, the second step follows from simple algebra, the third step follows from **Part 6**. □

B.4 Reformulating Gradient with respect to x

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

- $\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}x_i} = c(x,y)_{l_0,j_0,i_0}(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0}\rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle\langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle)$

Then we can rewrite $\frac{\mathrm{d}L_{l_0,j_0,i_0}(x,y)}{\mathrm{d}x_i}$ *as follows:*

$$\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}x_i} = c(x,y)_{l_0,j_0,i_0}\,\mathsf{A}^\top_{l_0,j_0,i}(f(x)_{l_0,j_0} - f(x)_{l_0,j_0}f(x)^\top_{l_0,j_0})h(y)_{l_0,i_0}$$

Proof. Note that by Fact A.1 we have

$$\begin{aligned}\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0}\rangle &= \mathsf{A}^\top_{l_0,j_0,i}\,\mathrm{diag}(f(x)_{l_0,j_0})h(y)_{l_0,i_0} \\ \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle\langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle &= \mathsf{A}^\top_{l_0,j_0,i}\,f(x)_{l_0,j_0}f(x)^\top_{l_0,j_0}h(y)_{l_0,i_0}\end{aligned}$$

By substitute the two terms above into $\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}x_i}$, we completes the proof. □

C Gradient Lipschitz

In this section, we aim to prove the lipschitz property for the gradient of loss function defined in the previous section. In Section C.1, we adopt a result from previous section to reform the gradient with respect to x. In Section C.5, we reform the gradient with respect to y. In Section C.3, we prove the lipschitz property for several basic terms. In Section C.4, we prove the lipschitz property of gradient with respect to x. In Section C.5, we prove the lipschitz property of gradient with respect to y.

C.1 Reformulating Gradient for x

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

- *Let* $L(x,y)_{l_0,j_0,i_0}$ *be computed in Lemma B.6*
- *Let* $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n\times d^2}$
- *Let* f *be defined in Definition 3.3*

- *Let h be defined in Definition 3.4*
- *Let α be defined in Definition 3.2*
- *Let c be defined in Definition B.1*
- *Let L be defined in Definition 3.4*

Then, we have

- $$\underbrace{\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}x}}_{d^2\times 1} = \underbrace{c(x,y)_{l_0,j_0,i_0}}_{\text{scalar}} \underbrace{\mathsf{A}_{l_0,j_0}^\top}_{d^2\times n} (\underbrace{\mathrm{diag}(f(x)_{l_0,j_0})}_{n\times n} - \underbrace{f(x)_{l_0,j_0}}_{n\times 1} \underbrace{f(x)_{l_0,j_0}^\top}_{1\times n}) \underbrace{h(y)_{l_0,i_0}}_{n\times 1}$$

Proof. This trivially follows from Lemma B.8 □

C.2 Reformulating Gradient for y

Lemma C.2. *Let $\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}}$ be computed as in Lemma B.7.*

For the case $i_1 = i_0$, we can rewrite $\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1}}$ as

$$\underbrace{\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1}}}_{d\times 1} = \underbrace{A_{l_0,3}^\top}_{d\times n} \underbrace{f(x)_{l_0,j_0}}_{n\times 1} \underbrace{c(x,y)_{l_0,j_0,i_0}}_{\text{scalar}}$$

For the case $i_1 \neq i_0$, then it's

$$\underbrace{\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1}}}_{d\times 1} = \underbrace{\mathbf{0}_d}_{d\times 1}.$$

Proof.

$$\begin{aligned}\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1,i_2}} &= c(x,y)_{l_0,j_0,i_0}\langle f(x)_{l_0,j_0}, A_{l_0,3,i_2}\rangle \\ &= A_{l_0,3,i_2}^\top f(x)_{l_0,j_0} c(x,y)_{l_0,j_0,i_0}\end{aligned}$$

where the first step follows from **Part 7** of Lemma B.7, the second step follows from simple algebra. Thus, we know

$$\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}y_{i_1}} = A_{l_0,3}^\top f(x)_{l_0,j_0} c(x,y)_{l_0,j_0,i_0}$$

□

C.3 Lipschitz For Some Basic Terms

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

- *Let $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n\times d^2}$*
- *Let $b_{l_0,j_0,i_0} \in \mathbb{R}^n$ satisfy that $\|b\|_1 \leq 1$*
- *Let $\beta \in (0, 0.1)$*
- *Let $R \geq 4$*

- *Let* $x, y \in \mathbb{R}^d$ *satisfy* $\| \mathsf{A}_{l_0,j_0} x\|_2 \leq R$ *and* $\| \mathsf{A}_{l_0,j_0} y\|_2 \leq R$
- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- $\langle \exp(\mathsf{A}_{l_0,j_0} x), \mathbf{1}_n \rangle \geq \beta$
- $\langle \exp(\mathsf{A}_{l_0,j_0} y), \mathbf{1}_n \rangle \geq \beta$
- *Let* $R_f := \beta^{-2} n \exp(3R^2)$
- *Let* $\alpha(x)_{l_0,j_0}$ *be defined as Definition 3.2*
- *Let* $c(x,y)_{l_0,j_0,i_0}$ *be defined as Definition B.1*
- *Let* $f(x)_{l_0,j_0}$ *be defined as Definition 3.3*

We have

- *Part 0.* $\| \exp(\mathsf{A}_{l_0,j_0} x)\|_2 \leq \sqrt{n} \exp(R^2)$
- *Part 1.* $\| \exp(\mathsf{A}_{l_0,j_0} x) - \exp(\mathsf{A}_{l_0,j_0} y)\|_2 \leq R \exp(R^2) \cdot \|x - y\|_2$
- *Part 2.* $|\alpha(x)_{l_0,j_0} - \alpha(y)_{l_0,j_0}| \leq \sqrt{n} \cdot \| \exp(Ax) - \exp(Ay)\|_2$
- *Part 3.* $|\alpha(x)^{-1}_{l_0,j_0} - \alpha(y)^{-1}_{l_0,j_0}| \leq \beta^{-2} \cdot |\alpha(x) - \alpha(y)|$
- *Part 4.* $\|f(x)_{l_0,j_0} - f(y)_{l_0,j_0}\|_2 \leq R_f \cdot \|x - y\|_2$
- *Part 5.* $\|c(x,z)_{l_0,j_0,i_0} - c(y,z)_{l_0,j_0,i_0}\|_2 \leq R^2 \beta^{-2} n \exp(3R^2) \|x - y\|_2$
- *Part 6.* $\| \mathrm{diag}(f(x)_{l_0,j_0}) - \mathrm{diag}(f(y)_{l_0,j_0})\| \leq \beta^{-2} n \exp(3R^2) \|x - y\|_2$
- *Part 7.* $\|f(x)_{l_0,j_0}\|_2 \leq \beta^{-1} n \exp(2R^2)$
- *Part 8.* $f(x)_{l_0,j_0} f(x)^\top_{l_0,j_0} - f(y)_{l_0,j_0} f(y)_{l_0,j_0} \leq 2\beta^{-3} n^2 \exp(5R^2) \|x - y\|_2$
- *Part 9.* $\|(\mathrm{diag}(f(x)_{l_0,j_0}) - f(x)_{l_0,j_0} f(x)^\top_{l_0,j_0}) - (\mathrm{diag}(f(y)_{l_0,j_0}) - f(y)_{l_0,j_0} f(y)^\top_{l_0,j_0})\| \leq 3\beta^{-2} n^2 \exp(5R^2) \|x - y\|_2$
- *Part 10.* $\|c(x,y)_{l_0,j_0,i_0}\| \leq R\beta^{-1} n \exp(2R^2)$
- *Part 11.* $\|c(x,z)_{l_0,j_0,i_0} (\mathrm{diag}(f(x)_{l_0,j_0}) - f(x)_{l_0,j_0} f(x)^\top_{l_0,j_0}) - c(y,z)_{l_0,j_0,i_0} (\mathrm{diag}(f(y)_{l_0,j_0}) - f(y)_{l_0,j_0} f(x)^\top_{l_0,j_0})\| \leq 6R\beta^{-3} \exp(7R^2) \|x - y\|_2$

Proof. **Proof of Part 0.**

We can show that

$$\begin{aligned} \| \exp(\mathsf{A}_{l_0,j_0} x)\|_2 &\leq \sqrt{n} \cdot \| \exp(\mathsf{A}_{l_0,j_0} x)\|_\infty \\ &\leq \sqrt{n} \cdot \exp(\| \mathsf{A}_{l_0,j_0} x\|_\infty) \\ &\leq \sqrt{n} \cdot \exp(\| \mathsf{A}_{l_0,j_0} x\|_2) \\ &\leq \sqrt{n} \cdot \exp(R^2), \end{aligned}$$

where the first step follows from **Part 4** of Fact A.3, the second step follows from **Part6** of Fact A.3, the third step follows from Fact A.3, and the last step follows from $\| \mathsf{A}_{l_0,j_0} \| \leq R$ and $\|x\|_2 \leq R$.

Proof of Part 1. We have

$$\begin{aligned} \| \exp(\mathsf{A}_{l_0,j_0} x) - \exp(\mathsf{A}_{l_0,j_0} y)\|_2 &\leq \exp(R^2) \| \mathsf{A}_{l_0,j_0} x - \mathsf{A}_{l_0,j_0} y\|_2 \\ &\leq \exp(R^2) \| \mathsf{A}_{l_0,j_0} \| \|x - y\|_2 \\ &\leq R \exp(R^2) \|x - y\|_2 \end{aligned}$$

where the first step follows from **Part 10** of Fact A.3, the second step follows from **Part 4** of Fact A.4, the third step follows from $\| \mathsf{A}_{l_0,j_0} \| \leq R$.

Proof of Part 2.

$$\begin{aligned} |\alpha(x)_{l_0,j_0} - \alpha(y)_{l_0,j_0}| &= |\langle \exp(\mathsf{A}_{l_0,j_0} x) - \exp(\mathsf{A}_{l_0,j_0} y), \mathbf{1}_n \rangle| \\ &\leq \| \exp(\mathsf{A}_{l_0,j_0} x) - \exp(\mathsf{A}_{l_0,j_0} y) \|_2 \cdot \sqrt{n} \end{aligned}$$

where the 1st step follows from the definition of $\alpha(x)_{l_0,j_0}$, the 2nd step follows from Cauchy-Schwarz inequality (**Part 1** of Fact A.3).

Proof of Part 3.

We can show that

$$\begin{aligned} |\alpha(x)^{-1}_{l_0,j_0} - \alpha(y)^{-1}_{l_0,j_0}| &= \alpha(x)^{-1}_{l_0,j_0} \alpha(y)^{-1}_{l_0,j_0} \cdot |\alpha(x)_{l_0,j_0} - \alpha(y)_{l_0,j_0}| \\ &\leq \beta^{-2} \cdot |\alpha(x)_{l_0,j_0} - \alpha(y)_{l_0,j_0}| \end{aligned}$$

where the 1st step follows from simple algebra, the 2nd step follows from $\alpha(x)_{l_0,j_0}, \alpha(y)_{l_0,j_0} \geq \beta$.

Proof of Part 4.

We can show that

$$\begin{aligned} & \|f(x)_{l_0,j_0} - f(y)_{l_0,j_0}\|_2 \\ = & \|\alpha(x)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} x) - \alpha(y)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} y)\|_2 \\ \leq & \|\alpha(x)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} x) - \alpha(x)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} y)\|_2 + \|\alpha(x)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} y) - \alpha(y)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} y)\|_2 \\ \leq & \alpha(x)^{-1}_{l_0,j_0} \| \exp(\mathsf{A}_{l_0,j_0} x) - \exp(\mathsf{A}_{l_0,j_0} y)\|_2 + |\alpha(x)^{-1}_{l_0,j_0} - \alpha(y)^{-1}_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} y)\|_2 \end{aligned}$$

where the 1st step follows from the definition of $f(x)_{l_0,j_0}$ and $\alpha(x)_{l_0,j_0}$, the 2nd step follows from triangle inequality (**Part 3** of Fact A.4), the 3rd step follows from $\|\alpha A\| \leq |\alpha| \|A\|$(**Part 5** of Fact A.4).

For the first term in the above, we have

$$\begin{aligned} \alpha(x)^{-1}_{l_0,j_0} \| \exp(\mathsf{A}_{l_0,j_0} x) - \exp(\mathsf{A}_{l_0,j_0} y)\|_2 &\leq \beta^{-1} \| \exp(\mathsf{A}_{l_0,j_0} x) - \exp(\mathsf{A}_{l_0,j_0} y)\|_2 \\ &\leq \beta^{-1} \cdot R \exp(R^2) \cdot \|x - y\|_2 \end{aligned} \tag{2}$$

where the 1st step follows from $\alpha(x)_{l_0,j_0} \geq \beta$, the 2nd step follows from **Part 1**.

For the second term in the above, we have

$$\begin{aligned} |\alpha(x)^{-1}_{l_0,j_0} - \alpha(y)^{-1}_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} y)\|_2 &\leq \beta^{-2} \cdot |\alpha(x)_{l_0,j_0} - \alpha(y)_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} y)\|_2 \\ &\leq \beta^{-2} \cdot |\alpha(x)_{l_0,j_0} - \alpha(y)_{l_0,j_0}| \cdot \sqrt{n} \exp(R^2) \\ &\leq \beta^{-2} \cdot \sqrt{n} \cdot \| \exp(\mathsf{A}_{l_0,j_0} x) - \exp(A_{l_0,j_0} y)\|_2 \cdot \sqrt{n} \exp(R^2) \\ &\leq \beta^{-2} \cdot \sqrt{n} \cdot R \exp(R^2) \|x - y\|_2 \cdot \sqrt{n} \exp(R^2) \\ &= \beta^{-2} \cdot nR \exp(2R^2) \|x - y\|_2 \end{aligned} \tag{3}$$

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. (2) and Eq. (3) together, we have

$$\begin{aligned} \|f_{l_0,j_0}(x) - f_{l_0,j_0}(y)\|_2 &\leq \beta^{-1} \cdot R \exp(R^2) \cdot \|x - y\|_2 + \beta^{-2} \cdot nR \exp(2R^2) \|x - y\|_2 \\ &\leq 2\beta^{-2} nR \exp(2R^2) \|x - y\|_2 \\ &\leq \beta^{-2} n \exp(3R^2) \|x - y\|_2 \end{aligned}$$

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

$$\|c(x,z)_{l_0,j_0,i_0} - c(y,z)_{l_0,j_0,i_0}\|_2 = \|\langle f(x)_{l_0,j_0}, h(z)_{l_0,i_0} \rangle - \langle f(y)_{l_0,j_0}, h(z)_{l_0,i_0} \rangle\|_2$$

$$
\begin{aligned}
&= \|\langle (f(x)_{l_0,j_0} - f(y)_{l_0,j_0}), h(z)_{l_0,i_0} \rangle\|_2 \\
&\leq \|h(z)_{l_0,i_0}\|_2 \|f(x)_{l_0,j_0} - f(y)_{l_0,j_0}\|_2 \\
&\leq \|\mathsf{A}_{l_0,3}\, z_{i_0}\|_2 \|f(x)_{l_0,j_0} - f(y)_{l_0,j_0}\|_2 \\
&\leq \|\mathsf{A}_{l_0,3}\, z_{i_0}\|_2 \cdot \beta^{-2} n \exp(3R^2) \|x - y\|_2 \\
&\leq \|\mathsf{A}_{l_0,3}\| \|z_{i_0}\|_2 \beta^{-2} n \exp(3R^2) \|x - y\|_2 \\
&\leq R\beta^{-2} n \exp(3R^2) \|x - y\|_2
\end{aligned}
$$

where the first step follows from the definition of $c(x,y)_{l_0,j_0,i_0}$, the second step follows from simple algebra, the third step follows from Fact A.3, the fourth step follows from the definition of $h(y)_{l_0,i_0}$, the fifth step follows from **Part 4**, the sixth step follows from Fact A.4, the last step follows from $\|A_{l_0,3}\| \leq R$ and $\|z_{i_0}\|_2 \leq R$.

Thus, we complete the proof.

Proof of Part 6

$$
\begin{aligned}
\|\operatorname{diag}(f(x)_{l_0,j_0}) - \operatorname{diag}(f(y)_{l_0,j_0})\| &= \|\operatorname{diag}(f(x)_{l_0,j_0} - f(y)_{l_0,j_0})\| \\
&\leq \|f(x)_{l_0,j_0} - f(y)_{l_0,j_0}\|_\infty \\
&\leq \|f(x)_{l_0,j_0} - f(y)_{l_0,j_0}\|_2 \\
&\leq \beta^{-2} n \exp(3R^2) \|x - y\|_2
\end{aligned}
$$

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

Proof of Part 7

$$
\begin{aligned}
\|f(x)_{l_0,j_0}\|_2 &= \|\alpha(x)^{-1}_{l_0,j_0} \cdot u(x)_{l_0,j_0}\|_2 \\
&\leq \|\alpha(x)^{-1}_{l_0,j_0}\|_2 \|u(x)_{l_0,j_0}\|_2 \\
&\leq \beta \|\alpha(x)_{l_0,j_0}\| \exp(\mathsf{A}_{l_0,j_0}\, x)\|_2 \\
&\leq \beta^{-1} \|\langle \exp(\mathsf{A}_{l_0,j_0}\, x), \mathbf{1}_n \rangle\|_2 \sqrt{n} \cdot \exp(R^2) \\
&\leq \beta^{-1} \|\exp(\mathsf{A}_{l_0,j_0}\, x)\|_2 \|\mathbf{1}_n\|_2 \sqrt{n} \cdot \exp(R^2) \\
&\leq \beta^{-1} \sqrt{n} \cdot \exp(R^2) \sqrt{n} \cdot \exp(R^2) \\
&= \beta^{-1} n \exp(2R^2)
\end{aligned}
$$

where the first step follows from the definition of $f(x)_{l_0,j_0}$, the second step follows from Fact A.3, the third step follows from $\langle \exp(\mathsf{A}_{l_0,j_0}\, x), \mathbf{1}_n \rangle \geq \beta$, the fourth step follows from **Part 0**, the fifth step follows from Fact A.3, the sixth step follows from **Part 0**, the last step follows from simple algebra.

Proof of Part 8 For the simplicity of the proof, we define

$$
\begin{aligned}
C_1 &:= f(x)_{l_0,j_0} f(x)^\top_{l_0,j_0} - f(x)_{l_0,j_0} f(y)^\top_{l_0,j_0} \\
C_2 &:= f(x)_{l_0,j_0} f(y)^\top_{l_0,j_0} - f(y)_{l_0,j_0} f(y)^\top_{l_0,j_0}
\end{aligned}
$$

Then it's obvious that

$$
\|f(x)_{l_0,j_0} f(x)^\top_{l_0,j_0} - f(y)_{l_0,j_0} f(y)^\top_{l_0,j_0}\| = \|C_1 + C_2\|
$$

Since C_1 and C_2 are similar, we only needs to bound $\|C_1\|$:

$$
\begin{aligned}
\|f(x)_{l_0,j_0} f(x)^\top_{l_0,j_0} - f(x)_{l_0,j_0} f(y)^\top_{l_0,j_0}\| &= \|f(x)_{l_0,j_0} (f(x)_{l_0,j_0} - f(y)_{l_0,j_0})^\top\| \\
&\leq \|f(x)_{l_0,j_0}\|_2 \|f(x)_{l_0,j_0} - f(y)_{l_0,j_0}\|_2 \\
&\leq \beta^{-1} n \exp(2R^2) \beta^{-2} n \exp(3R^2) \|x - y\|_2 \\
&= \beta^{-3} n^2 \exp(5R^2) \|x - y\|_2
\end{aligned}
$$

where the first step follows from simple algebra, the second step follows from Fact A.4, the third step follows from **Part 7** and **Part 4**, the last step follows from simple algebra.

Thus, we know

$$\|f(x)_{l_0,j_0}f(x)_{l_0,j_0}^\top - f(y)_{l_0,j_0}f(y)_{l_0,j_0}^\top\| \leq 2\beta^{-3}n^2\exp(5R^2)\|x-y\|_2$$

Proof of Part 9

$$\begin{aligned}
&\|(\mathrm{diag}(f(x)_{l_0,j_0}) - f(x)_{l_0,j_0}f(x)_{l_0,j_0}^\top) - (\mathrm{diag}(f(y)_{l_0,j_0}) - f(y)_{l_0,j_0}f(y)_{l_0,j_0}^\top)\| \\
=& \|(\mathrm{diag}(f(x)_{l_0,j_0}) - \mathrm{diag}(f(y)_{l_0,j_0}) + (f(y)_{l_0,j_0}f(y)_{l_0,j_0}^\top - f(x)_{l_0,j_0}f(x)_{l_0,j_0}^\top)\| \\
\leq& \|\mathrm{diag}(f(x)_{l_0,j_0}) - \mathrm{diag}(f(y)_{l_0,j_0}\| + \|f(y)_{l_0,j_0}f(y)_{l_0,j_0}^\top - f(x)_{l_0,j_0}f(x)_{l_0,j_0}^\top\| \\
\leq& \beta^{-2}n\exp(3R^2)\|x-y\|_2 + 2\beta^{-3}n^2\exp(5R^2)\|x-y\|_2 \\
\leq& 3\beta^{-2}n^2\exp(5R^2)\|x-y\|_2
\end{aligned}$$

where the first step follows from simple algebra, the second step follows from Fact A.4, the third step follows from **Part 6** and **Part 7**, the last step follows from simple algebra.

Proof of Part 10

$$\begin{aligned}
\|c(x,y)_{l_0,j_0,i_0}\| &= \|\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle\| \\
&\leq \|f(x)_{l_0,j_0}\|_2\|h(y)_{l_0,i_0}\|_2 \\
&\leq R\beta^{-1}n\exp(2R^2)
\end{aligned}$$

where the first step follows from the definition of $c(x,y)_{l_0,j_0,i_0}$, the second step follows from Fact A.3, the third step follows from **Part 7**.

Proof of Part 11

Let

$$\begin{aligned}
d(x) :&= c(x,z)_{l_0,j_0,i_0} \\
e(x) :&= \mathrm{diag}(f(x)_{l_0,j_0}) - f(x)_{l_0,j_0}f(x)_{l_0,j_0}^\top
\end{aligned}$$

Then it's obvious that

$$\begin{aligned}
&\|d(x)e(x) - d(y)e(y)\| \\
=& \|c(x,z)_{l_0,j_0,i_0}(\mathrm{diag}(f(x)_{l_0,j_0}) - f(x)_{l_0,j_0}f(x)_{l_0,j_0}^\top) - c(y,z)_{l_0,j_0,i_0}(\mathrm{diag}(f(y)_{l_0,j_0}) - f(y)_{l_0,j_0}f(x)_{l_0,j_0}^\top)\|
\end{aligned}$$

Define

$$\begin{aligned}
C_1 :&= d(x)e(x) - d(x)e(y) \\
C_2 :&= d(x)e(y) - d(y)e(y)
\end{aligned}$$

Thus, it's apparent that

$$\|d(x)e(x) - d(y)e(y)\| = \|C_1 + C_2\|$$

Since C_1 and C_2 are similar, we only need to bound $\|C_1\|$:

$$\begin{aligned}
\|d(x)e(x) - d(x)e(y)\| &= \|d(x)(e(x) - e(y))\| \\
&\leq \|d(x)\|\|e(x) - e(y)\| \\
&\leq R\beta^{-1}n\exp(2R^2)3\beta^{-2}n^2\exp(5R^2)\|x-y\|_2 \\
&= 3R\beta^{-3}\exp(7R^2)\|x-y\|_2
\end{aligned}$$

where the first step follows from simple algebra, the second step follows from Fact A.4, the third step follows from **Part 10** and **Part 9**.

Thus, we have

$$\|d(x)e(x) - d(y)e(y)\| \leq 6R\beta^{-3}\exp(7R^2)\|x-y\|_2$$

□

Lemma C.4. *If the following conditions holds*

- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- $\|x\|_2 \leq R$
- *Let* β *be lower bound on* $\langle \exp(\mathsf{A}_{l_0,j_0} x), \mathbf{1}_n \rangle$

Then we have

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

Proof. We have

$$\begin{aligned} \langle \exp(\mathsf{A}_{l_0,j_0} x), \mathbf{1}_n \rangle \geq & \max_{i \in [n]} \exp(-|(\mathsf{A}_{l_0,j_0} x)_i|) \\ \geq & \exp(-\| \mathsf{A}_{l_0,j_0} x \|_\infty) \\ \geq & \exp(-\| \mathsf{A}_{l_0,j_0} x \|_2) \\ \geq & \exp(-R^2) \end{aligned}$$

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

□

C.4 Lipschitz for $\nabla L(x, :)$

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

- *Let* $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$
- *Let* $b_{l_0,j_0,i_0} \in \mathbb{R}$
- *Let* $\beta \in (0, 0.1)$
- *Let* $R \geq 4$
- *Let* $x, y \in \mathbb{R}^d$ *satisfy* $\|\mathsf{A}_{l_0,j_0} x\| \leq R$ *and* $\|\mathsf{A}_{l_0,j_0} y\| \leq R$
- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- $\langle \exp(\mathsf{A}_{l_0,j_0} x), \mathbf{1}_n \rangle \geq \beta$
- $\langle \exp(\mathsf{A}_{l_0,j_0} y), \mathbf{1}_n \rangle \geq \beta$
- *Let* $R_f := \beta^{-2} n \exp(3R^2)$
- *Let* $\alpha(x)_{l_0,j_0}$ *be defined as Definition 3.2*
- *Let* $c(x)_{l_0,j_0,i_0}$ *be defined as Definition B.1*
- *Let* $f(x)_{l_0,j_0}$ *be defined as Definition 3.3*
- *Let* $\nabla L_{l_0,j_0,i_0}$ *be computed as in Lemma C.1*
- *Let* L *be defined as Definition B.3*

Then we have

$$\|\nabla L(x, y) - \nabla L(\widehat{x}, y)\| \leq 6mndR^2 \exp(10R^2) \|x - \widehat{x}\|_2$$

Proof.

$$\|\nabla L_{l_0,j_0,i_0}(x, y) - \nabla L_{l_0,j_0,i_0}(\widehat{x}, y)\|$$

$$
\begin{aligned}
&= \| c(x,y)_{l_0,j_0,i_0} \mathsf{A}_{l_0,j_0}^\top (\mathrm{diag}(f(x)_{l_0,j_0}) - f(x)_{l_0,j_0} f(x)_{l_0,j_0}^\top) h(y)_{l_0,i_0} \\
&\quad - c(\widehat{x},y)_{l_0,j_0,i_0} \mathsf{A}_{l_0,j_0}^\top (\mathrm{diag}(f(\widehat{x})_{l_0,j_0}) - f(\widehat{x})_{l_0,j_0} f(\widehat{x})_{l_0,j_0}^\top) h(y)_{l_0,i_0} \| \\
&\leq \| \mathsf{A}_{l_0,j_0}^\top \| \| h(y)_{l_0,i_0} \| \\
&\quad \| c(x,y)_{l_0,j_0,i_0} (\mathrm{diag}(f(x)_{l_0,j_0}) - f(x)_{l_0,j_0} f(x)_{l_0,j_0}^\top) - c(\widehat{x},y)_{l_0,j_0,i_0} (\mathrm{diag}(f(\widehat{x})_{l_0,j_0}) - f(\widehat{x})_{l_0,j_0} f(\widehat{x})_{l_0,j_0}^\top) \| \\
&\leq R \| c(x,y)_{l_0,j_0,i_0} (\mathrm{diag}(f(x)_{l_0,j_0}) - f(x)_{l_0,j_0} f(x)_{l_0,j_0}^\top) - c(\widehat{x},y)_{l_0,j_0,i_0} (\mathrm{diag}(f(\widehat{x})_{l_0,j_0}) - f(\widehat{x})_{l_0,j_0} f(\widehat{x})_{l_0,j_0}^\top) \| \\
&\leq 6R^2 \beta^{-3} \exp(7R^2) \|x - \widehat{x}\|_2
\end{aligned}
$$

where the first step follows from the definition of $\nabla L_{l_0,j_0,i_0}(x,y)$, the second step follows from simple algebra, the third step follows from $\| \mathsf{A}_{l_0,j_0}^\top \| \leq R$ and $\|h(y)_{l_0,i_0}\| = \mathsf{A}_{l_0,3}\, y_{i_0} \leq R$, the fourth step follows from **Part 10** of Lemma C.3.

Thus, we have

$$
\begin{aligned}
|\nabla L(x,y) - \nabla L(\widehat{x},y)\| &= \| \sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d (\nabla L_{l_0,j_0,i_0}(x,y) - \nabla L_{l_0,j_0,i_0}(\widehat{x},y)) \| \\
&\leq \sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d \| \nabla L_{l_0,j_0,i_0}(x,y) - \nabla L_{l_0,j_0,i_0}(\widehat{x},y) \| \\
&\leq \sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d 6R^2 \beta^{-3} \exp(7R^2) \|x - \widehat{x}\|_2 \\
&= 6mndR^2 \beta^{-3} \exp(7R^2) \|x - \widehat{x}\|_2 \\
&\leq 6mndR^2 \exp(10R^2) \|x - \widehat{x}\|_2
\end{aligned}
$$

where the first step follows from the definition of L, the second step follows from Fact A.4, the third step follows from the lipschitz of L_{l_0,j_0,i_0}, the fourth step follows from simple algebra, the last step follows from plugging β from Lemma C.4. □

C.5 Lipschitz for $\nabla L(y)$

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

- *Let* $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$
- *Let* $b_{l_0,j_0,i_0} \in \mathbb{R}^n$ *satisfy that* $\|b\|_1 \leq 1$
- *Let* $\beta \in (0, 0.1)$
- *Let* $R \geq 4$
- *Let* $x, y \in \mathbb{R}^d$ *satisfy* $\| \mathsf{A}_{l_0,j_0}\, x\|_2 \leq R$ *and* $\| \mathsf{A}_{l_0,j_0}\, y\|_2 \leq R$
- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- $\langle \exp(\mathsf{A}_{l_0,j_0}\, x), \mathbf{1}_n \rangle \geq \beta$
- $\langle \exp(\mathsf{A}_{l_0,j_0}\, y), \mathbf{1}_n \rangle \geq \beta$
- *Let* $R_f := \beta^{-2} n \exp(3R^2)$
- *Let* $\alpha(x)_{l_0,j_0}$ *be defined as Definition 3.2*
- *Let* $c(x,y)_{l_0,j_0,i_0}$ *be defined as Definition B.1*
- *Let* $f(x)_{l_0,j_0}$ *be defined as Definition 3.3*

Then we have

$$
\|\nabla L(:,y) - \nabla(:,\widehat{y})\| \leq R^2 n^2 md \exp(6R^2) \|y - \widehat{y}\|_2
$$

Proof.

$$\begin{aligned}
\|\nabla L_{l_0,j_0,i_0}(:,y) - \nabla L_{l_0,j_0,i_0}(:,\widehat{y})\| &= \|A_{l_0,3}^\top f(:)_{l_0,j_0} c(:,y)_{l_0,j_0,i_0} - A_{l_0,3}^\top f(:)_{l_0,j_0} c(:,\widehat{y})_{l_0,j_0,i_0}\| \\
&= \|A_{l_0,3}^\top f(:)_{l_0,j_0} (c(:,y)_{l_0,j_0,i_0} - c(:,\widehat{y})_{l_0,j_0,i_0})\| \\
&\leq \|A_{l_0,3}\| \|f(:)_{l_0,j_0}\|_2 |c(:,y)_{l_0,j_0,i_0} - c(:,\widehat{y})_{l_0,j_0,i_0}| \\
&\leq R\beta^{-1} n \exp(2R^2) |c(:,y)_{l_0,j_0,i_0} - c(:,\widehat{y})_{l_0,j_0,i_0}| \\
&= R\beta^{-1} n \exp(2R^2) |\langle f(:)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle - \langle f(:)_{l_0,j_0}, h(\widehat{y})_{l_0,i_0}\rangle| \\
&= R\beta^{-1} n \exp(2R^2) |f(:)_{l_0,j_0}^\top (h(y)_{l_0,i_0} - h(\widehat{y})_{l_0,i_0})| \\
&\leq R\beta^{-1} n \exp(2R^2) \|f(:)_{l_0,j_0}^\top\|_2 \|h(y)_{l_0,i_0} - h(\widehat{y})_{l_0,i_0}\|_2 \\
&\leq R\beta^{-2} n \exp(4R^2) \|h(y)_{l_0,i_0} - h(\widehat{y})_{l_0,i_0}\|_2 \\
&= R\beta^{-2} n \exp(4R^2) \|A_{l_0,3} y - A_{l_0,3}\widehat{y}\|_2 \\
&= R\beta^{-2} n \exp(4R^2) \|A_{l_0,3}(y - \widehat{y})\|_2 \\
&\leq R\beta^{-2} n \exp(4R^2) \|A_{l_0,3}\| \|y - \widehat{y}\|_2 \\
&\leq R^2\beta^{-2} n \exp(4R^2) \|y - \widehat{y}\|_2
\end{aligned}$$

where the first step follows from Lemma C.2, the second step follows from simple algebra, the third step follows from Fact A.3, the fourth step follows from **Part 7** of Lemma C.3 and $\|A_{l_0,3}\| \leq R$, the fifth step follows from the definition of $c(x,y)_{l_0,j_0,i_0}$, the sixth step follows from simple algebra, the seventh step follows from Fact A.3, the eighth step follows from **Part 7** of Lemma C.3, the ninth step follows from the definition of $h(y)_{l_0,i_0}$, the tenth step follows from simple algebra, the eleventh step follows from Fact A.4, the last step follows from $\|A_{l_0,3}\| \leq R$.

Thus, we have

$$\begin{aligned}
\|\nabla L(:,y) - \nabla(:,\widehat{y})\| &= \|\sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d (\nabla L_{l_0,j_0,i_0}(:,y) - \nabla L_{l_0,j_0,i_0}(:,\widehat{y}))\| \\
&\leq \sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d \|\nabla L_{l_0,j_0,i_0}(:,y) - \nabla L_{l_0,j_0,i_0}(:,\widehat{y})\| \\
&\leq \sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d R^2\beta^{-2} n \exp(4R^2) \|y - \widehat{y}\|_2 \\
&= R^2 nmd\beta^{-2} n \exp(4R^2) \|y - \widehat{y}\|_2 \\
&\leq R^2 n^2 md \exp(6R^2) \|y - \widehat{y}\|_2
\end{aligned}$$

where the first step follows from the definition of L, the second step follows from Fact A.3, the third step follows from the lipschitz of $\nabla L_{l_0,j_0,i_0}(:,x)$, the fourth step follows from simple algebra, the last step follows from Lemma C.4. □

D GRADIENT FOR Q

In Section D.1, we define the basic definitions and problems to be used in this section. In Section D.2, we compute the gradient with respect to Q step by step. In Section D.3, we reform the gradient in a way that is easy for us to prove its lipschitz property. In Section D.4, we prove the lipschitz property for several basic terms. In Section D.5, we state some intermediate steps for proving the lipschitz of gradient. In Section D.6, we prove the lipschitz property of the gradient with respect to Q.

D.1 DEFINITIONS

Definition D.1. *Let $A_{l_0,1}, A_{l_0,2} \in \mathbb{R}^{n \times d}$. Let $\mathsf{A}_{l_0} = A_{l_0,1} \otimes A_{l_0,2} \in \mathbb{R}^{n^2 \times d^2}$. Let $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$ denote the j_0-th block of $\mathsf{A}_{l_0} \in \mathbb{R}^{n^2 \times d^2}$.*

For each $l_0 \in [m]$, for each $j_0 \in [n]$.

We define $u(Q)_{l_0,j_0} \in \mathbb{R}^n$ as follows

$$\underbrace{u(Q)_{l_0,j_0}}_{n \times 1} := \exp(\underbrace{\mathsf{A}_{l_0,j_0}}_{n \times d^2} \underbrace{\mathrm{vec}(QK^\top)}_{d^2 \times 1})$$

Definition D.2. *For each $l_0 \in [m]$, for each $j_0 \in [n]$.*

We define $\alpha(Q)_{l_0,j_0} \in \mathbb{R}$ as follows

$$\underbrace{\alpha(Q)_{l_0,j_0}}_{\mathrm{scalar}} := \langle \underbrace{u(Q)_{l_0,j_0}}_{n \times 1}, \underbrace{\mathbf{1}_n}_{n \times 1} \rangle.$$

Definition D.3. *Let $A_{l_0,1}, A_{l_0,2} \in \mathbb{R}^{n \times d}$. Let $\mathsf{A}_{l_0} = A_{l_0,1} \otimes A_{l_0,2} \in \mathbb{R}^{n^2 \times d^2}$. Let $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$ denote the j_0-th block of $\mathsf{A}_{l_0} \in \mathbb{R}^{n^2 \times d^2}$.*

We define $f(Q)_{l_0,j_0} : \mathbb{R}^{d^2} \to \mathbb{R}^n$,

$$\underbrace{f(Q)_{l_0,j_0}}_{n \times 1} := \underbrace{\alpha(Q)_{l_0,j_0}^{-1}}_{\mathrm{scalar}} \cdot \underbrace{u(Q)_{l_0,j_0}}_{n \times 1}$$

Definition D.4. *We define $c(Q,y)_{j_0,i_0} \in \mathbb{R}$ as follows*

$$\underbrace{c(Q,y)_{l_0,j_0,i_0}}_{\mathrm{scalar}} := \langle \underbrace{f(Q)_{l_0,j_0}}_{n \times 1}, \underbrace{h(y)_{l_0,i_0}}_{n \times 1} \rangle - \underbrace{b_{l_0,j_0,i_0}}_{\mathrm{scalar}}$$

Definition D.5. *For each $l_0 \in [m]$, $j_0 \in [n]$, $i_0 \in [d]$. We define L_{l_0,j_0,i_0} as follows*

$$\underbrace{L_{l_0,j_0,i_0}(Q,y)}_{\mathrm{scalar}} := 0.5 \underbrace{c_{l_0,j_0,i_0}(Q,y)^2}_{\mathrm{scalar}}$$

Definition D.6. *The final loss is*

$$\underbrace{L(Q,y)}_{\mathrm{scalar}} := \sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d \underbrace{L_{l_0,j_0,i_0}(Q,y)}_{\mathrm{scalar}}.$$

Definition D.7. *We define the diagonal matrix $D \in \mathbb{R}^{n^2 \times n^2}$ as:*

$$\underbrace{D(Q)}_{n^2 \times n^2} = \mathrm{diag}(\exp(\underbrace{A_1}_{n^2 \times d} \underbrace{Q}_{d \times L} \underbrace{K^\top}_{L \times d} A_2^\top) \mathbf{1}_n)$$

We give our formal definition of the optimization formulation

Definition D.8. *Let $A_1, A_2 \in \mathbb{R}^{n \times d}$. We define the optimization formulation as the following:*

$$\min_{Q \in \mathbb{R}^{d \times d}} L(Q) = \min_{Q \in \mathbb{R}^{d \times d}} \| D(Q)^{-1} \exp(A_1 Q K^\top A_2^\top) A_3 Y - B \|_F^2$$

Definition D.9. *Let $A_1, A_2 \in \mathbb{R}^{n \times d}$. Let $\mathsf{A} = A_1 \otimes A_2 \in \mathbb{R}^{n^2 \times d^2}$. Let $D'(Q) \in \mathbb{R}^{n^2 \times n^2}$ denote the diagonal matrix $D'(Q) := D(Q) \otimes I_n$. We define the vector version of optimization formulation as the following:*

$$\min_{Q \in \mathbb{R}^{d \times d}} L(Q) = \min_{Q \in \mathbb{R}^{d \times d}} \| \mathrm{mat}(D'(Q)^{-1} \exp(\mathsf{A} \cdot \mathrm{vec}(QK^\top))) A_3 Y - B \|_2^2$$

D.2 GRADIENT

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

- *Let Q_{i_2,k_2} denote the i_2-the row and k_2-th column of $Q \in \mathbb{R}^{d \times L}$*

Then, we have

- *Part 1. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}\,\mathrm{vec}(QK^\top)}{\mathrm{d}Q_{i_2,k_2}} = \mathrm{vec}(\underbrace{e_{i_2}}_{d\times 1}\underbrace{e_{k_2}^\top}_{1\times L}\underbrace{K^\top}_{L\times d})$$

- *Part 2. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}\,\mathsf{A}_{l_0,j_0}\,\mathrm{vec}(QK^\top)}{\mathrm{d}Q_{i_2,k_2}} = \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}$$

- *Part 3. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}u(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}} = \underbrace{u(Q)_{l_0,j_0}}_{n\times 1} \circ \underbrace{(\mathsf{A}_{l_0,j_0}\,\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top))}_{n\times 1}$$

- *Part 4. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}\alpha(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}} = \langle \underbrace{u(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\rangle$$

- *Part 5. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}\alpha(Q)_{l_0,j_0}^{-1}}{\mathrm{d}Q_{i_2,k_2}} = -\underbrace{\alpha(Q)_{l_0,j_0}^{-1}}_{scalar}\langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\rangle$$

- *Part 6. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}f(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}} = \underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}) - \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}\langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\rangle$$

- *Part 7. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\begin{aligned}
&\frac{\mathrm{d}c(Q,y)_{l_0,j_0,i_0}}{\mathrm{d}Q_{i_2,k_2}} \\
&= \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}), \underbrace{h(y)_{l_0,i_0}}_{n\times 1}\rangle - \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1}\rangle\langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\rangle
\end{aligned}$$

- *Part 8. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\begin{aligned}
&\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,y)}{\mathrm{d}Q_{i_2,k_2}} \\
&= \underbrace{c_{l_0,j_0,i_0}(Q,y)}_{scalar} \\
&(\langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}), \underbrace{h(y)_{l_0,i_0}}_{n\times 1}\rangle - \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1}\rangle\langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\rangle)
\end{aligned}$$

Proof. **Proof of Part 1.** For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$\begin{aligned}\frac{\mathrm{d}\,\mathrm{vec}(QK^\top)}{\mathrm{d}Q_{i_2,k_2}} =&\ \mathrm{vec}(\frac{\mathrm{d}Q}{\mathrm{d}Q_{i_2,k_2}}K^\top)\\ =&\ \mathrm{vec}(\underbrace{e_{i_2}}_{d\times 1}\underbrace{e_{k_2}^\top}_{1\times L}\underbrace{K^\top}_{L\times d})\end{aligned}$$

where the first step simple algebra.

Proof of Part 2. For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$\begin{aligned}\frac{\mathrm{d}\,\mathsf{A}_{l_0,j_0}\,\mathrm{vec}(QK^\top)}{\mathrm{d}Q_{i_2,k_2}} =&\ \mathsf{A}_{l_0,j_0}\,\mathrm{vec}(\frac{\mathrm{d}Q}{\mathrm{d}Q_{i_2,k_2}}K^\top)\\ =&\ \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\end{aligned}$$

where the first step chain rule and the second step follows from **Part 1**.

Proof of Part 3. For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$\begin{aligned}\frac{\mathrm{d}u(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}} =&\ \frac{\mathrm{d}\exp(\mathsf{A}_{l_0,j_0}\,\mathrm{vec}(QK^\top))}{\mathrm{d}Q_{i_2,k_2}}\\ =&\ \underbrace{u(Q)_{l_0,j_0}}_{n\times 1}\circ(\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1})\end{aligned}$$

where the first step follows from the definition of u and the second step follows from chain rule and **Part 2.**

Proof of Part 4 For each $i_2 \in [d]$ and $k_2 \in [L]$

$$\begin{aligned}\frac{\mathrm{d}\alpha(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}} =&\ \frac{\mathrm{d}\langle u(Q)_{l_0,j_0},\mathbf{1}_n\rangle.}{\mathrm{d}Q_{i_2,k_2}}\\ =&\ \langle\frac{\mathrm{d}u(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}},\mathbf{1}_n\rangle\\ =&\ \langle\underbrace{u(Q)_{l_0,j_0}}_{n\times 1}\circ(\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}),\mathbf{1}_n\rangle\\ =&\ \langle\underbrace{u(Q)_{l_0,j_0}}_{n\times 1},\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\rangle\end{aligned}$$

where the first step follows from the definition of $\alpha(Q)_{l_0,j_0}$, the second step follows from simple algebra, the third step follows from **Part 3**.

Proof of Part 5 For each $i_2 \in [d]$ and $k_2 \in [L]$

$$\begin{aligned}\frac{\mathrm{d}\alpha(Q)_{l_0,j_0}^{-1}}{\mathrm{d}Q_{i_2,k_2}} =&\ -\alpha(Q)_{l_0,j_0}^{-2}\frac{\mathrm{d}\alpha(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}}\\ =&\ -\alpha(Q)_{l_0,j_0}^{-2}\langle\underbrace{u(Q)_{l_0,j_0}}_{n\times 1},\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\rangle\\ =&\ -\underbrace{\alpha(Q)_{l_0,j_0}^{-1}}_{scalar}\langle\underbrace{f(Q)_{l_0,j_0}}_{n\times 1},\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top)}_{d^2\times 1}\rangle\end{aligned}$$

where the first step follows from simple algebra, the second step follows from **Part 4**, the third step follows from simple algebra and Fact A.1.

Proof of Part 6 For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$\frac{\mathrm{d}f(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}} = \frac{\mathrm{d}\alpha(Q)_{l_0,j_0}^{-1}\cdot u(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}}$$

$$
\begin{aligned}
&= \frac{\mathrm{d}\alpha(Q)_{l_0,j_0}^{-1}}{\mathrm{d}Q_{i_2,k_2}} u(Q)_{l_0,j_0} + \frac{\mathrm{d}u(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}} \alpha(Q)_{l_0,j_0}^{-1} \\
&= -\underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1} \rangle + \underbrace{u(Q)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1}) \underbrace{\alpha(Q)_{l_0,j_0}^{-1}}_{\textit{scalar}} \\
&= \underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1}) - \underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1} \rangle
\end{aligned}
$$

where the first step follows from the definition of $f(Q)_{l_0,j_0}$, the second step follows from differential chain rule, the third step follows from **Part 3** and **Part 5**, the last step follows from simple algebra.

Proof of Part 7 For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$
\begin{aligned}
&\frac{\mathrm{d}c(Q,y)_{l_0,j_0,i_0}}{\mathrm{d}Q_{i_2,k_2}} \\
&= \frac{\mathrm{d}\langle f(Q)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle - b_{l_0,j_0,i_0}}{\mathrm{d}Q_{i_2,k_2}} \\
&= \frac{\mathrm{d}\langle f(Q)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle}{\mathrm{d}Q_{i_2,k_2}} \\
&= \langle \frac{\mathrm{d}f(Q)_{l_0,j_0}}{\mathrm{d}Q_{i_2,k_2}}, h(y)_{l_0,i_0}\rangle \\
&= \underbrace{h(y)_{l_0,i_0}^\top}_{1\times n} (\underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1}) - \underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1} \rangle) \\
&= \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1}), \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle - \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1} \rangle
\end{aligned}
$$

where the first step follows from the definition of $c(Q,y)_{l_0,j_0,i_0}$, the second step follows from b_{l_0,j_0,i_0} is a constant, the third step follows from only $f(Q)_{l_0,j_0}$ is dependent on Q, the fourth step follows from **Part 6**,the last step follows from simple algebra.

Proof of Part 8 For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$
\begin{aligned}
&\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,y)}{\mathrm{d}Q_{i_2,k_2}} \\
&= \frac{\mathrm{d}0.5 c_{l_0,j_0,i_0}(Q,y)^2}{\mathrm{d}Q_{i_2,k_2}} \\
&= c_{l_0,j_0,i_0}(Q,y) \frac{\mathrm{d}c_{l_0,j_0,i_0}(Q,y)}{\mathrm{d}Q_{i_2,k_2}} \\
&= \underbrace{c_{l_0,j_0,i_0}(Q,y)}_{\text{scalar}} (\langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1}), \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle - \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle \langle \underbrace{f(Q)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)}_{d^2\times 1} \rangle)
\end{aligned}
$$

where the first step follows from the definition of $L_{l_0,j_0,i_0}(Q,y)$, the second step follows from simple algebra, the third step follows from **Part 7**. □

D.3 Reformulating Gradient

Lemma D.11. *If the following conditions hold*

- *Let* $f(Q)_{l_0,j_0}$ *be defined as Definition D.3*
- *Let* $\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}}$ *be compute as* **Part 8** *of Lemma D.10*
- *Let* $v_1 := (\mathsf{A}_{l_0,j_0} \mathrm{vec}(e_{i_2} e_{k_2}^\top K^\top)) \circ h(y)_{l_0,i_0}$
- *Let* $v_2 := h(y)_{l_0,i_0}$

- *Let* $v_3 := \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)$

then $\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,y)}{\mathrm{d}Q_{i_2,k_2}}$ *can be rewrite as*

$$c_{l_0,j_0,i_0}(Q,y)(\langle f(Q)_{l_0,j_0}, v_1\rangle - \langle f(Q)_{l_0,j_0}, v_2\rangle \langle f(Q)_{l_0,j_0}, v_3\rangle)$$

Proof. The proof trivially follows from Fact A.1. □

D.4 Lipschitz of several terms

Lemma D.12. *If the following conditions hold*

- *Let* $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$
- *Let* $b_{l_0,j_0,i_0} \in \mathbb{R}^n$ *satisfy that* $\|b\|_1 \leq 1$
- *Let* $\beta \in (0, 0.1)$
- *Let* $R \geq 4$
- *Let* $\| \operatorname{vec}(QK^\top)\|_2 \leq R$
- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)), \mathbf{1}_n \rangle \geq \beta$
- $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top)), \mathbf{1}_n \rangle \geq \beta$
- *Let* $R_f := \beta^{-2} n \exp(3R^2)$
- *Let* $\alpha(Q)_{l_0,j_0}$ *be defined as Definition D.2*
- *Let* $c(Q)_{l_0,j_0,i_0}$ *be defined as Definition D.4*
- *Let* $f(Q)_{l_0,j_0}$ *be defined as Definition D.3*
- *Let* $v_1 := (\mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top) h(y)_{l_0,i_0}$
- *Let* $v_2 := h(y)_{l_0,i_0}$
- *Let* $v_3 := \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)$

Then we have

- *Part 1.* $\| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top))\|_2 \leq \sqrt{n} \cdot \exp(R^2)$
- *Part 2.* $\|f(Q)_{l_0,j_0}\|_2 \leq \beta^{-1} n \exp(2R^2)$
- *Part 3.* $|c(Q,:)_{l_0,j_0,i_0}| \leq R\beta^{-1} n \exp(2R^2)$
- *Part 4.* $\|v_2\|_2 \leq R^2$
- *Part 5.* $\|v_3\|_2 \leq R^2$
- *Part 6.* $\|v_1\|_2 \leq R^4$
- *Part 7.* $|(\langle f(Q)_{l_0,j_0}, v_1\rangle - \langle f(Q)_{l_0,j_0}, v_2\rangle \langle f(Q)_{l_0,j_0}, v_3\rangle)| \leq \beta^{-1} n^2 R^4 \exp(6R^2)$
- *Part 8.* $\beta \geq \exp(-R^2)$

Proof. **Proof of Part 1**

$$\| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top))\|_2 \leq \sqrt{n} \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)\|_\infty$$

$$\leq \sqrt{n} \cdot \exp(\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)\|_\infty)$$
$$\leq \sqrt{n} \cdot \exp(\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)\|_2)$$
$$\leq \sqrt{n} \cdot \exp(R^2)$$

Proof of Part 2

$$\begin{aligned}\|f(Q)_{l_0,j_0}\|_2 &= \|\alpha(Q)_{l_0,j_0}^{-1} \cdot u(Q)_{l_0,j_0}\|_2 \\ &\leq \|\alpha(Q)_{l_0,j_0}^{-1}\|_2 \|u(Q)_{l_0,j_0}\|_2 \\ &\leq \beta^{-1} \|\exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top))\|_2 \\ &\leq \beta^{-1}\sqrt{n} \cdot \exp(R^2)\end{aligned}$$

where the first step follows from the definition of $f(x)_{l_0,j_0}$, the second step follows from Fact A.3, the third step follows from $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)), \mathbf{1}_n \rangle \geq \beta$,the fourth step follows from **Part 1**.

Proof of Part 3

$$\begin{aligned}\|c(Q,y)_{l_0,j_0,i_0}\| &= \|\langle f(Q)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle\| \\ &\leq \|f(Q)_{l_0,j_0}\|_2 \|h(y)_{l_0,i_0}\|_2 \\ &\leq R\beta^{-1} n \exp(2R^2)\end{aligned}$$

where the first step follows from the definition of $c(Q,y)_{l_0,j_0,i_0}$, the second step follows from Fact A.3, the third step follows from **Part 2**.

Proof of Part 4

$$\begin{aligned}\|h(y)_{l_0,i_0}\|_2 &= \|A_{l_0,3} y_{i_0}\|_2 \\ &\leq \|A_{l_0,3}\| \|y_{i_0}\|_2 \\ &\leq R^2\end{aligned}$$

where the first step follows from the definition of $h(y)_{l_0,i_0}$, the second step follows from Fact A.4, the third step follows from $\|A_{l_0,3}\|$ and $\|y_{i_0}\|_2 \leq R$.

Proof of Part 5

$$\begin{aligned}\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_2 &\leq \| \mathsf{A}_{l_0,j_0} \| \| \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_2 \\ &\leq R^2\end{aligned}$$

where the first step follows from Fact A.4, the second step follows from $\| \mathsf{A}_{l_0,j_0} \| \leq R$ and Fact A.5.

Proof of Part 6

$$\begin{aligned}\|(\mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)) \circ h(y)_{l_0,i_0}\|_2 &\leq \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_\infty \|h(y)_{l_0,i_0}\|_2 \\ &\leq \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_2 \|h(y)_{l_0,i_0}\|_2 \\ &\leq R^4\end{aligned}$$

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

Proof of Part 7

$$\begin{aligned}&|\langle f(Q)_{l_0,j_0}, (\mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top) h(y)_{l_0,i_0}\rangle - \langle f(Q)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle \langle f(Q)_{l_0,j_0}, \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\rangle| \\ \leq &|\langle f(Q)_{l_0,j_0}, (\mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top) h(y)_{l_0,i_0}\rangle| + |\langle f(Q)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle \langle f(Q)_{l_0,j_0}, \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\rangle| \\ \leq &\|f(Q)_{l_0,j_0}\|_2 \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top) h(y)_{l_0,i_0}\|_2 + \|f(Q)_{l_0,j_0}\|_2 \|h(y)_{l_0,i_0}\|_2 \|f(Q)_{l_0,j_0}\|_2 \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_2 \\ \leq &\beta^{-1} n \exp(2R^2) \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top) h(y)_{l_0,i_0}\|_2 + \beta^{-2} n^2 \exp(4R^2) \|h(y)_{l_0,i_0}\|_2 \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_2 \\ \leq &\beta^{-1} n \exp(2R^2) \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_2 \|h(y)_{l_0,i_0}\|_2 + \beta^{-2} n^2 \exp(4R^2) \|h(y)_{l_0,i_0}\|_2 \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)\|_2 \\ \leq &\beta^{-1} n \exp(2R^2) R^4 + \beta^{-2} n^2 \exp(4R^2) R^4 \\ \leq &\beta^{-1} n^2 R^4 \exp(6R^2)\end{aligned}$$

where the first step follows from triangle inequality, the second step follows from Fact A.3, the third step follows from **Part 2**, the fourth step follows from Fact A.3, the fifth step follows from **Part 5** and **Part 4**, the last step follows from simple algebra.

Proof of Part 8 We have

$$\begin{aligned}\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)), \mathbf{1}_n \rangle \geq & \max_{i\in[n]} \exp(-|(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top))_i|) \\ \geq & \exp(-\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)\|_\infty) \\ \geq & \exp(-\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)\|_2) \\ \geq & \exp(-R^2)\end{aligned}$$

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

Lemma D.13. *If the following conditions hold*

- *Let* $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$
- *Let* $b_{l_0,j_0,i_0} \in \mathbb{R}^n$ *satisfy that* $\|b\|_1 \leq 1$
- *Let* $\beta \in (0, 0.1)$
- *Let* $R \geq 4$
- *Let* $\| \operatorname{vec}(QK^\top)\|_2 \leq R$
- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)), \mathbf{1}_n \rangle \geq \beta$
- $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top)), \mathbf{1}_n \rangle \geq \beta$
- *Let* $R_f := \beta^{-2} n \exp(3R^2)$
- *Let* $\alpha(Q)_{l_0,j_0}$ *be defined as Definition D.2*
- *Let* $c(Q)_{l_0,j_0,i_0}$ *be defined as Definition D.4*
- *Let* $f(Q)_{l_0,j_0}$ *be defined as Definition D.3*

Then we have

- *Part 1.* $\| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \leq R^2 \exp(R^2) \|Q - \widehat{Q}\|_F$
- *Part 2.* $|\alpha(Q)_{l_0,j_0} - \alpha(\widehat{Q})_{l_0,j_0}| \leq \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \cdot \sqrt{n}$
- *Part 3.* $|\alpha(Q)^{-1}_{l_0,j_0} - \alpha(\widehat{Q})^{-1}_{l_0,j_0}| \leq \beta^{-2} \cdot |\alpha(Q)_{l_0,j_0} - \alpha(\widehat{Q})_{l_0,j_0}|$
- *Part 4.* $\|f(Q)_{l_0,j_0} - f(Q)_{l_0,j_0}\|_2 \leq \beta^{-2} n \exp(3R^2) \|Q - \widehat{Q}\|_F$
- *Part 5.* $\|c(Q,:)_{l_0,j_0,i_0} - c(\widehat{Q},:)_{l_0,j_0,i_0}\|_2 \leq R^2 \beta^{-2} n \exp(3R^2) \|Q - \widehat{Q}\|_2$

Note that $\|Q\|_F = (\sum_i \sum_j Q_{i,j}^2)^{1/2} = \| \operatorname{vec}(Q)\|_2$

Proof. **Proof of Part 1.**

$$\begin{aligned}\| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \leq & \exp(R^2) \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top) - \mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top)\|_2 \\ \leq & \exp(R^2) \| \mathsf{A}_{l_0,j_0} \| \| \operatorname{vec}(QK^\top) - \operatorname{vec}(\widehat{Q}K^\top)\|_2 \\ = & \exp(R^2) \| \mathsf{A}_{l_0,j_0} \| \|QK^\top - \widehat{Q}K^\top\|_F \\ \leq & \exp(R^2) \| \mathsf{A}_{l_0,j_0} \| \|Q - \widehat{Q}\|_F \|K\|_F\end{aligned}$$

$$\leq R^2 \exp(R^2) \|Q - \widehat{Q}\|_F$$

Proof of Part 2.

$$\begin{aligned} |\alpha(x)_{l_0,j_0} - \alpha(y)_{l_0,j_0}| &= |\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top)), \mathbf{1}_n \rangle| \\ &\leq \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top)) \|_2 \cdot \sqrt{n} \end{aligned}$$

Proof of Part 3.

$$\begin{aligned} |\alpha(Q)^{-1}_{l_0,j_0} - \alpha(\widehat{Q})^{-1}_{l_0,j_0}| &= \alpha(Q)^{-1}_{l_0,j_0} \alpha(\widehat{Q})^{-1}_{l_0,j_0} \cdot |\alpha(Q)_{l_0,j_0} - \alpha(\widehat{Q})_{l_0,j_0}| \\ &\leq \beta^{-2} \cdot |\alpha(Q)_{l_0,j_0} - \alpha(\widehat{Q})_{l_0,j_0}| \end{aligned}$$

Proof of Part 4. We can show that

$$\begin{aligned} \|f(Q)_{l_0,j_0} - f(\widehat{Q})_{l_0,j_0}\|_2 &= \|\alpha(Q)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \alpha(\widehat{Q})^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \\ &\leq \|\alpha(Q)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \alpha(Q)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \\ &\quad + \|\alpha(Q)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top)) - \alpha(\widehat{Q})^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \\ &\leq \alpha(Q)^{-1}_{l_0,j_0} \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \\ &\quad + |\alpha(Q)^{-1}_{l_0,j_0} - \alpha(\widehat{Q})^{-1}_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \end{aligned}$$

where the 1st step follows from the definition of $f(Q)_{l_0,j_0}$ and $\alpha(Q)_{l_0,j_0}$, the 2nd step follows from triangle inequality (**Part 3** of Fact A.3), the 3rd step follows from $\|\alpha A\| \leq |\alpha| \|A\|$(**Part 5** of Fact A.4).

For the first term in the above, we have

$$\begin{aligned} & \alpha(Q)^{-1}_{l_0,j_0} \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \qquad (4) \\ \leq & \beta^{-1} \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \\ \leq & \beta^{-1} \cdot R^2 \exp(R^2) \cdot \|Q - \widehat{Q}\|_F \qquad (5) \end{aligned}$$

where the 1st step follows from $\alpha(x)_{l_0,j_0} \geq \beta$, the 2nd step follows from **Part 1**.

For the second term in the above, we have

$$\begin{aligned} & |\alpha(Q)^{-1}_{l_0,j_0} - \alpha(\widehat{Q})^{-1}_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \\ \leq & \beta^{-2} \cdot |\alpha(Q)_{l_0,j_0} - \alpha(\widehat{Q})_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \\ \leq & \beta^{-2} \cdot |\alpha(Q)_{l_0,j_0} - \alpha(\widehat{Q})_{l_0,j_0}| \cdot \sqrt{n} \exp(R^2) \\ \leq & \beta^{-2} \cdot \sqrt{n} \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top))\|_2 \cdot \sqrt{n} \exp(R^2) \\ \leq & \beta^{-2} \cdot \sqrt{n} \cdot R^2 \exp(R^2) \|Q - \widehat{Q}\|_F \cdot \sqrt{n} \exp(R^2) \\ = & \beta^{-2} \cdot nR^2 \exp(2R^2) \|Q - \widehat{Q}\|_F \qquad (6) \end{aligned}$$

where the 1st step follows from the result of **Part 3**, the 2nd step follows from **Part 1** of Lemma D.12, 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. (4) and Eq. (6) together, we have

$$\begin{aligned} \|f_{l_0,j_0}(Q) - f_{l_0,j_0}(\widehat{Q})\|_2 &\leq \beta^{-1} \cdot R^2 \exp(R^2) \cdot \|Q - \widehat{Q}\|_F + \beta^{-2} \cdot nR^2 \exp(2R^2) \|Q - \widehat{Q}\|_F \\ &\leq 2\beta^{-2} nR^2 \exp(2R^2) \|Q - \widehat{Q}\|_F \\ &\leq \beta^{-2} n \exp(3R^2) \|Q - \widehat{Q}\|_F \end{aligned}$$

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.

$$\|c(Q,:)_{l_0,j_0,i_0} - c(\widehat{Q},:)_{l_0,j_0,i_0}\|_2 = \|\langle f(Q)_{l_0,j_0}, h(:)_{l_0,i_0} \rangle - \langle f(\widehat{Q})_{l_0,j_0}, h(:)_{l_0,i_0} \rangle\|_2$$

$$
\begin{aligned}
&= \|\langle (f(Q)_{l_0,j_0} - f(\widehat{Q})_{l_0,j_0}), h(:)_{l_0,i_0} \rangle\|_2 \\
&\leq \|h(:)_{l_0,i_0}\|_2 \|f(Q)_{l_0,j_0} - f(\widehat{Q})_{l_0,j_0}\|_2 \\
&\leq \|\mathsf{A}_{l_0,3}\, y_{i_0}\|_2 \|f(Q)_{l_0,j_0} - f(\widehat{Q})_{l_0,j_0}\|_2 \\
&\leq \|\mathsf{A}_{l_0,3}\, y_{i_0}\|_2 \cdot \beta^{-2} n \exp(3R^2) \|Q - \widehat{Q}\|_F \\
&\leq \|\mathsf{A}_{l_0,3}\| \|y_{i_0}\|_2 \beta^{-2} n \exp(3R^2) \|Q - \widehat{Q}\|_F \\
&\leq R\beta^{-2} n \exp(3R^2) \|Q - \widehat{Q}\|_F
\end{aligned}
$$

where the first step follows from the definition of $c(x,y)_{l_0,j_0,i_0}$, the second step follows from simple algebra, the third step follows from Fact A.3, the fourth step follows from the definition of $h(y)_{l_0,i_0}$, the fifth step follows from **Part 4**, the sixth step follows from Fact A.4, the last step follows from $\|A_{l_0,3}\| \leq R$ and $\|z_{i_0}\|_2 \leq R$ □

D.5 Summary of 3 steps

Lemma D.14. *If the following conditions hold*

- *Let* $f(Q)_{l_0,j_0}$ *be defined as Definition D.3*
- *Let* $\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}}$ *be compute as* **Part 8** *of Lemma D.10*
- *Let* $v_1 := (\mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)) \circ h(y)_{l_0,i_0}$
- *Let* $v_2 := h(y)_{l_0,i_0}$
- *Let* $v_3 := \mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)$

Then we have

- $|\langle f(Q)_{l_0,j_0}, v_1 \rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_1 \rangle| \leq \beta^{-2} n R^4 \exp(3R^2) \|Q - \widehat{Q}\|_F$
- $|\langle f(Q)_{l_0,j_0}, v_2 \rangle \langle f(Q)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2 \rangle \langle f(\widehat{Q})_{l_0,j_0}, v_3 \rangle| \leq 2\beta^{-3} n \exp(2R^2) R^6 n \exp(3R^2) \|Q - \widehat{Q}\|_F$
- $|\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}} - \frac{\mathrm{d}L_{l_0,j_0,i_0}(\widehat{Q},:)}{\mathrm{d}Q_{i_2,k_2}}| \leq \beta^{-3} n^3 R^7 \exp(19R^2) \|Q - \widehat{Q}\|_F$

Proof. **Proof of Part 1.**

$$
\begin{aligned}
|\langle f(Q)_{l_0,j_0}, v_1 \rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_1 \rangle| &= |\langle f(Q)_{l_0,j_0} - f(\widehat{Q})_{l_0,j_0}, v_1 \rangle| \\
&\leq \|f(Q)_{l_0,j_0} - f(\widehat{Q})_{l_0,j_0}\|_2 \|v_1\|_2 \\
&\leq \beta^{-2} n \exp(3R^2) \|Q - \widehat{Q}\|_F \|(\mathsf{A}_{l_0,j_0} \operatorname{vec}(e_{i_2} e_{k_2}^\top K^\top)) \circ h(y)_{l_0,i_0}\|_2 \\
&\leq \beta^{-2} n R^4 \exp(3R^2) \|Q - \widehat{Q}\|_F
\end{aligned}
$$

where the first step follows from simple algebra, the second step follows from Fact A.3, the third step follows from **Part 4** of Lemma D.13, the fourth step follows from **Part 6** of Lemma D.12.

Proof of Part 2. For convenience, we define

$$
\begin{aligned}
C_1 &:= \langle f(Q)_{l_0,j_0}, v_2 \rangle \langle f(Q)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2 \rangle \langle f(Q)_{l_0,j_0}, v_3 \rangle \\
C_2 &:= \langle f(\widehat{Q})_{l_0,j_0}, v_2 \rangle \langle f(Q)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2 \rangle \langle f(\widehat{Q})_{l_0,j_0}, v_3 \rangle
\end{aligned}
$$

Then it's apparent that

$$
|C_1 + C_2| = |\langle f(Q)_{l_0,j_0}, v_2 \rangle \langle f(Q)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2 \rangle \langle f(\widehat{Q})_{l_0,j_0}, v_3 \rangle|
$$

Since C_1 and C_2 are similar, we only need to bound $|C_1|$:

$$
|C_1| = |\langle f(Q)_{l_0,j_0}, v_2 \rangle \langle f(Q)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2 \rangle \langle f(Q)_{l_0,j_0}, v_3 \rangle|
$$

$$
\begin{aligned}
&= |\langle f(Q)_{l_0,j_0}, v_3\rangle(\langle f(Q)_{l_0,j_0}, v_2\rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2\rangle)| \\
&\leq |\langle f(Q)_{l_0,j_0}, v_3\rangle||\langle f(Q)_{l_0,j_0}, v_2\rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2\rangle| \\
&= |\langle f(Q)_{l_0,j_0}, v_3\rangle||\langle f(Q)_{l_0,j_0} - f(\widehat{Q})_{l_0,j_0}, v_2\rangle| \\
&\leq \|f(Q)_{l_0,j_0}\|_2\|v_3\|_2\|f(Q)_{l_0,j_0} - f(\widehat{Q})_{l_0,j_0}\|_2\|v_2\|_2 \\
&\leq \beta^{-1}n\exp(2R^2)\|v_3\|_2\beta^{-2}n\exp(3R^2)\|Q - \widehat{Q}\|_F\|v_2\|_2 \\
&\leq \beta^{-3}n^2R^6\exp(5R^2)\|Q - \widehat{Q}\|_F
\end{aligned}
$$

where the first step follows from the definition of C_1, the second step follows from simple algebra, the third step follows from triangular inequality, the fourth step follows from simple algebra, the fifth step follows from Fact A.3, the sixth step follows from **Part 4** of Lemma D.13, the last step follows from **Part 4** and **Part 5** of Lemma D.12.

Thus, we obtained the bound for $|\langle f(Q)_{l_0,j_0}, v_2\rangle\langle f(Q)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2\rangle\langle f(\widehat{Q})_{l_0,j_0}, v_3\rangle|$:

$$
|\langle f(Q)_{l_0,j_0}, v_2\rangle\langle f(Q)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2\rangle\langle f(\widehat{Q})_{l_0,j_0}, v_3\rangle| \leq 2\beta^{-3}n^2R^6\exp(5R^2)\|Q - \widehat{Q}\|_F
$$

Proof of Part 3 By Lemma D.11, we know that $\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}}$ can be written as

$$
\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}} = c_{l_0,j_0,i_0}(Q,y)(\langle f(Q)_{l_0,j_0}, v_1\rangle - \langle f(Q)_{l_0,j_0}, v_2\rangle\langle f(Q)_{l_0,j_0}, v_3\rangle)
$$

For convenience, we define

$$
\begin{aligned}
s(Q) &:= c_{l_0,j_0,i_0}(Q,y) \\
t(Q) &:= (\langle f(Q)_{l_0,j_0}, v_1\rangle - \langle f(Q)_{l_0,j_0}, v_2\rangle\langle f(Q)_{l_0,j_0}, v_3\rangle)
\end{aligned}
$$

Thus $\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}}$ can be rewrite as

$$
\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}} = s(Q)t(Q)
$$

Then the lipschitz of $\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}}$ can be expressed as

$$
|\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}} - \frac{\mathrm{d}L_{l_0,j_0,i_0}(\widehat{Q},:)}{\mathrm{d}Q_{i_2,k_2}}| = |s(Q)t(Q) - s(\widehat{Q})t(\widehat{Q})|
$$

Use the same techinque in the proof of **Part 2**, we define

$$
\begin{aligned}
C_1 &:= s(Q)t(Q) - s(Q)t(\widehat{Q}) \\
C_2 &:= s(Q)t(\widehat{Q}) - s(\widehat{Q})t(\widehat{Q})
\end{aligned}
$$

Then it's apparent that

$$
|s(Q)t(Q) - s(\widehat{Q})t(\widehat{Q})| = |C_1 + C_2|
$$

First, we upper bound $|C_1|$ as follows:

$$
\begin{aligned}
|C_1| &= |s(Q)t(Q) - s(Q)t(\widehat{Q})| \\
&= |s(Q)(t(Q) - t(\widehat{Q}))| \\
&\leq |s(Q)||t(Q) - t(\widehat{Q})| \\
&= |s(Q)||(\langle f(Q)_{l_0,j_0}, v_1\rangle - \langle f(Q)_{l_0,j_0}, v_2\rangle\langle f(Q)_{l_0,j_0}, v_3\rangle) - (\langle f(\widehat{Q})_{l_0,j_0}, v_1\rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_2\rangle\langle f(\widehat{Q})_{l_0,j_0}, v_3\rangle)| \\
&= |s(Q)||(\langle f(Q)_{l_0,j_0}, v_1\rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_1\rangle) + (\langle f(\widehat{Q})_{l_0,j_0}, v_2\rangle\langle f(\widehat{Q})_{l_0,j_0}, v_3\rangle - \langle f(Q)_{l_0,j_0}, v_2\rangle\langle f(Q)_{l_0,j_0}, v_3\rangle)|
\end{aligned}
$$

$$
\begin{aligned}
&\leq |s(Q)|(|\langle f(Q)_{l_0,j_0}, v_1\rangle - \langle f(\widehat{Q})_{l_0,j_0}, v_1\rangle| + |\langle f(\widehat{Q})_{l_0,j_0}, v_2\rangle\langle f(\widehat{Q})_{l_0,j_0}, v_3\rangle - \langle f(Q)_{l_0,j_0}, v_2\rangle\langle f(Q)_{l_0,j_0}, v_3\rangle|) \\
&\leq |s(Q)|(\beta^{-2}nR^4\exp(3R^2)\|Q-\widehat{Q}\|_F\| + 2\beta^{-3}n^2R^6\exp(5R^2)\|Q-\widehat{Q}\|_F) \\
&\leq |s(Q)|\cdot\beta^{-2}n^2R^6\exp(8R^2)\|Q-\widehat{Q}\|_F \\
&\leq \beta^{-3}n^3R^7\exp(10R^2)\|Q-\widehat{Q}\|_F
\end{aligned}
$$

where the first step follows from the definition of C_1, the second step follows from simple algebra, the third step follows from Fact A.3, the fourth step follows from the definition of $t(Q)$, the fifth step follows from simple algebra, the sixth step follows from triangular inequality, the seventh step follows from **Part 3** and **Part 4** the last step follows from **Part 3** of Lemma D.12.

Next, we upper bound $|C_2|$:

$$
\begin{aligned}
|C_2| &= |s(Q)t(\widehat{Q}) - s(\widehat{Q})t(\widehat{Q})| \\
&= |(s(Q) - s(\widehat{Q}))t(\widehat{Q})| \\
&\leq |s(Q) - s(\widehat{Q})||t(\widehat{Q}| \\
&\leq R^2\beta^{-2}n\exp(3R^2)\|Q-\widehat{Q}\|_F|t(\widehat{Q}| \\
&\leq R^6\beta^{-3}n^3\exp(9R^2)\|Q-\widehat{Q}\|_F
\end{aligned}
$$

where the first step follows from the definition of C_2, the second step follows from simple algebra, the third step follows from simple algebra, the fourth step follows from **Part 5** of Lemma D.13, the last step follows from **Part 7** of Lemma D.12.

Thus, we can obtain the upper bound for $|s(Q)t(Q) - s(\widehat{Q})t(\widehat{Q})|$:

$$
\begin{aligned}
|s(Q)t(Q) - s(\widehat{Q})t(\widehat{Q})| &= |C_1 + C_2| \\
&\leq \beta^{-3}n^3R^7\exp(10R^2)\|Q-\widehat{Q}\|_F + R^6\beta^{-3}n^3\exp(9R^2)\|Q-\widehat{Q}\|_F \\
&\leq \beta^{-3}n^3R^7\exp(19R^2)\|Q-\widehat{Q}\|_F
\end{aligned}
$$

where the first step follows from simple algebra, the second step follows from the upper bound of $|C_1|$ and $|C_2|$, the last step follows from simple algebra. □

D.6 Lipschitz of $\nabla L_{l_0,j_0,i_0}(Q,:)$

Lemma D.15. *If the following conditions hold*

- *Let $f(Q)_{l_0,j_0}$ be defined as Definition D.3*
- *Let $\frac{\mathrm{d}L_{l_0,j_0,i_0}(Q,:)}{\mathrm{d}Q_{i_2,k_2}}$ be compute as* **Part 8** *of Lemma D.10*
- *Let $v_1 := (\mathsf{A}_{l_0,j_0}\operatorname{vec}(e_{i_2}e_{k_2}^\top K^\top))\circ h(y)_{l_0,i_0}$*
- *Let $v_2 := h(y)_{l_0,i_0}$*
- *Let $v_3 := \mathsf{A}_{l_0,j_0}\operatorname{vec}(e_{i_2}e_{k_2}^\top K^\top)$*

Then we have

$$
\|\frac{\mathrm{d}L(Q,:)}{\mathrm{d}\operatorname{vec}(Q)} - \frac{\mathrm{d}L(\widehat{Q},:)}{\mathrm{d}\operatorname{vec}(Q)}\|_2 \leq dLn^3R^7\exp(22R^2)\|Q-\widehat{Q}\|_F
$$

Proof.

$$
\begin{aligned}
\|\frac{\mathrm{d}L(Q,:)}{\mathrm{d}\operatorname{vec}(Q)} - \frac{\mathrm{d}L(\widehat{Q},:)}{\mathrm{d}\operatorname{vec}(Q)}\|_2 &\leq \sum_{i_2=1}^d\sum_{k_2=1}^L |\frac{\mathrm{d}L(Q,:)}{\mathrm{d}Q_{i_2,k_2}}|_{Q=Q} - \frac{\mathrm{d}L(Q,:)}{\mathrm{d}Q_{i_2,k_2}}|_{Q=\widehat{Q}}| \\
&\leq \sum_{i_2=1}^d\sum_{k_2=1}^L \beta^{-3}n^3R^7\exp(19R^2)\|Q-\widehat{Q}\|_F
\end{aligned}
$$

$$= \beta^{-3} d L n^3 R^7 \exp(19R^2) \|Q - \widehat{Q}\|_F$$
$$\leq d L n^3 R^7 \exp(22R^2) \|Q - \widehat{Q}\|_F$$

where the first step follows from Fact A.3, the second step follows from **Part 3** of Lemma D.14, the fourth step follows from simple algebra, the last step follows from **Part 8** of Lemma D.12. □

E GRADIENT FOR K

In Section E.1, we define the basic definitions and problems to be used in this section. In Section E.2, we compute the gradient with respect to K step by step. In Section E.3, we reform the gradient in a way that is easy for us to prove its lipschitz property. In Section E.5, we prove the lipschitz property for several basic terms. In Section E.6, we state some intermediate steps for proving the lipschitz of gradient. In Section E.7, we prove the lipschitz property of the gradient with respect to K.

E.1 DEFINITIONS

Definition E.1. *Let $A_{l_0,1}, A_{l_0,2} \in \mathbb{R}^{n \times d}$. Let $\mathsf{A}_{l_0} = A_{l_0,1} \otimes A_{l_0,2} \in \mathbb{R}^{n^2 \times d^2}$. Let $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$ denote the j_0-th block of $\mathsf{A}_{l_0} \in \mathbb{R}^{n^2 \times d^2}$.*

For each $l_0 \in [m]$, for each $j_0 \in [n]$.

We define $u(K)_{l_0,j_0} \in \mathbb{R}^n$ as follows

$$u(K)_{l_0,j_0} := \exp(\underbrace{\mathsf{A}_{l_0,j_0}}_{n \times d^2} \underbrace{\mathrm{vec}(QK^\top)}_{d^2 \times 1})$$

Definition E.2. *For each $l_0 \in [m]$, for each $j_0 \in [n]$.*

We define $\alpha(K)_{l_0,j_0} \in \mathbb{R}$ as follows

$$\alpha(K)_{l_0,j_0} := \langle u(K)_{l_0,j_0}, \mathbf{1}_n \rangle.$$

Definition E.3. *Let $A_{l_0,1}, A_{l_0,2} \in \mathbb{R}^{n \times d}$. Let $\mathsf{A}_{l_0} = A_{l_0,1} \otimes A_{l_0,2} \in \mathbb{R}^{n^2 \times d^2}$. Let $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$ denote the j_0-th block of $\mathsf{A}_{l_0} \in \mathbb{R}^{n^2 \times d^2}$.*

We define $f(K)_{l_0,j_0} : \mathbb{R}^{d^2} \to \mathbb{R}^n$,

$$f(K)_{l_0,j_0} := \alpha(K)^{-1}_{l_0,j_0} \cdot u(K)_{l_0,j_0}$$

Definition E.4. *We define $c(K,y)_{j_0,i_0} \in \mathbb{R}$ as follows*

$$c(K,y)_{l_0,j_0,i_0} := \langle f(K)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle - b_{l_0,j_0,i_0}$$

Definition E.5. *For each $l_0 \in [m]$, $j_0 \in [n]$, $i_0 \in [d]$. We define L_{l_0,j_0,i_0} as follows*

$$L_{l_0,j_0,i_0}(K,y) := 0.5 c_{l_0,j_0,i_0}(K,y)^2$$

Definition E.6. *The final loss is*

$$L(K,y) := \sum_{l_0=1}^m \sum_{j_0=1}^n \sum_{i_0=1}^d L_{l_0,j_0,i_0}(K,y).$$

Definition E.7. *We define the diagonal matrix $D \in \mathbb{R}^{n^2 \times n^2}$ as:*

$$D(K) = \mathrm{diag}(\exp(A_1 Q K^\top A_2^\top) \mathbf{1}_n)$$

We give our formal definition of the optimization formulation

Definition E.8. *Let $A_1, A_2 \in \mathbb{R}^{n \times d}$. We define the optimization formulation as the following:*

$$\min_{Q \in \mathbb{R}^{d \times d}} L(Q) = \min_{Q \in \mathbb{R}^{d \times d}} \| D(Q)^{-1} \exp(A_1 Q K^\top A_2^\top) A_3 Y - B \|_F^2$$

Definition E.9. *Let $A_1, A_2 \in \mathbb{R}^{n \times d}$. Let $\mathsf{A} = A_1 \otimes A_2 \in \mathbb{R}^{n^2 \times d^2}$. Let $D'(Q) \in \mathbb{R}^{n^2 \times n^2}$ denote the diagonal matrix $D'(Q) := D(Q) \otimes I_n$. We define the vector version of optimization formulation as the following:*

$$\min_{Q \in \mathbb{R}^{d \times d}} L(Q) = \min_{Q \in \mathbb{R}^{d \times d}} \| \mathrm{mat}(D'(Q)^{-1} \exp(\mathsf{A} \cdot \mathrm{vec}(QK^\top))) A_3 Y - B \|_2^2$$

E.2 GRADIENT

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

- *Let K_{i_2,k_2} denote the i_2-the row and k_2-th column of $Q \in \mathbb{R}^{d \times L}$*

Then, we have

- *Part 1. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}\,\mathrm{vec}(QK^\top)}{\mathrm{d}K_{i_2,k_2}} = \mathrm{vec}(\underbrace{Q}_{d\times L}\underbrace{e_{k_2}}_{L\times 1}\underbrace{e_{i_2}^\top}_{1\times d})$$

- *Part 2. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}\,\mathsf{A}_{l_0,j_0}\,\mathrm{vec}(QK^\top)}{\mathrm{d}K_{i_2,k_2}} = \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}$$

- *Part 3. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}u(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}} = \underbrace{u(K)_{l_0,j_0}}_{n\times 1} \circ \underbrace{(\mathsf{A}_{l_0,j_0}\,\mathrm{vec}(e_{i_2}e_{k_2}^\top K^\top))}_{n\times 1}$$

- *Part 4. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}\alpha(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}} = \langle \underbrace{u(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\rangle$$

- *Part 5. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}\alpha(K)_{l_0,j_0}^{-1}}{\mathrm{d}K_{i_2,k_2}} = -\underbrace{\alpha(K)_{l_0,j_0}^{-1}}_{\textit{scalar}}\langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\rangle$$

- *Part 6. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\frac{\mathrm{d}f(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}} = \underbrace{f(K)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}) - \underbrace{f(K)_{l_0,j_0}}_{n\times 1}\langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\rangle$$

- *Part 7. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\begin{aligned}&\frac{\mathrm{d}c(K,y)_{l_0,j_0,i_0}}{\mathrm{d}K_{i_2,k_2}}\\ =\ &\langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}), \underbrace{h(y)_{l_0,i_0}}_{n\times 1}\rangle - \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1}\rangle\langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\rangle\end{aligned}$$

- *Part 8. For each $i_2 \in [d]$ and $k_2 \in [L]$, we have*

$$\begin{aligned}&\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,y)}{\mathrm{d}K_{i_2,k_2}}\\ =\ &\underbrace{c_{l_0,j_0,i_0}(K,y)}_{\textit{scalar}}\\ &(\langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}), \underbrace{h(y)_{l_0,i_0}}_{n\times 1}\rangle - \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1}\rangle\langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\rangle)\end{aligned}$$

Proof. **Proof of Part 1.** For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$\begin{aligned}\frac{\mathrm{d}\,\mathrm{vec}(QK^\top)}{\mathrm{d}K_{i_2,k_2}} =&\ \mathrm{vec}(Q(\frac{\mathrm{d}K}{K_{i_2,k_2}})^\top)\\ =&\ \mathrm{vec}(\underbrace{Q}_{d\times L}\underbrace{e_{k_2}}_{L\times 1}\underbrace{e_{i_2}^\top}_{1\times d})\end{aligned}$$

where the first step simple algebra

Proof of Part 2. For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$\begin{aligned}\frac{\mathrm{d}\,\mathsf{A}_{l_0,j_0}\,\mathrm{vec}(QK^\top)}{\mathrm{d}K_{i_2,k_2}} =&\ \mathsf{A}_{l_0,j_0}\,\mathrm{vec}(\frac{\mathrm{d}Q}{\mathrm{d}K_{i_2,k_2}}K^\top)\\ =&\ \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\end{aligned}$$

where the first step chain rule and the second step follows from **Part 1**.

Proof of Part 3. For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$\begin{aligned}\frac{\mathrm{d}u(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}} =&\ \frac{\mathrm{d}\exp(\mathsf{A}_{l_0,j_0}\,\mathrm{vec}(QK^\top))}{\mathrm{d}K_{i_2,k_2}}\\ =&\ \underbrace{u(K)_{l_0,j_0}}_{n\times 1}\circ(\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1})\end{aligned}$$

where the first step follows from the definition of u and the second step follows from chain rule and **Part 2.**

Proof of Part 4 For each $i_2 \in [d]$ and $k_2 \in [L]$

$$\begin{aligned}\frac{\mathrm{d}\alpha(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}} =&\ \frac{\mathrm{d}\langle u(K)_{l_0,j_0},\mathbf{1}_n\rangle.}{\mathrm{d}K_{i_2,k_2}}\\ =&\ \langle\frac{\mathrm{d}u(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}},\mathbf{1}_n\rangle\\ =&\ \langle\underbrace{u(K)_{l_0,j_0}}_{n\times 1}\circ(\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}),\underbrace{\mathbf{1}_n}_{n\times 1}\rangle\\ =&\ \langle\underbrace{u(K)_{l_0,j_0}}_{n\times 1},\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\rangle\end{aligned}$$

where the first step follows from the definition of $\alpha(K)_{l_0,j_0}$, the second step follows from simple algebra, the third step follows from **Part 3**.

Proof of Part 5 For each $i_2 \in [d]$ and $k_2 \in [L]$

$$\begin{aligned}\frac{\mathrm{d}\alpha(K)_{l_0,j_0}^{-1}}{\mathrm{d}K_{i_2,k_2}} =&\ -\alpha(K)_{l_0,j_0}^{-2}\frac{\mathrm{d}\alpha(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}}\\ =&\ -\underbrace{\alpha(K)_{l_0,j_0}^{-2}}_{\text{scalar}}\langle\underbrace{u(K)_{l_0,j_0}}_{n\times 1},\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\rangle\\ =&\ -\underbrace{\alpha(K)_{l_0,j_0}^{-1}}_{\text{scalar}}\langle\underbrace{f(K)_{l_0,j_0}}_{n\times 1},\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2}\underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}\rangle\end{aligned}$$

where the first step follows from simple algebra, the second step follows from **Part 4**, the third step follows from simple algebra and Fact A.1.

Proof of Part 6 For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$\frac{\mathrm{d}f(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}} = \frac{\mathrm{d}\alpha(K)_{l_0,j_0}^{-1}\cdot u(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}}$$

$$
\begin{aligned}
&= \frac{\mathrm{d}\alpha(K)_{l_0,j_0}^{-1}}{\mathrm{d}K_{i_2,k_2}} u(K)_{l_0,j_0} + \frac{\mathrm{d}u(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}} \alpha(K)_{l_0,j_0}^{-1} \\
&= -\underbrace{f(K)_{l_0,j_0}}_{n\times 1} \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1} \rangle + \underbrace{u(K)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1} \underbrace{\alpha(K)_{l_0,j_0}^{-1}}_{\text{scalar}}) \\
&= \underbrace{f(K)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}) - \underbrace{f(K)_{l_0,j_0}}_{n\times 1} \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1} \rangle
\end{aligned}
$$

where the first step follows from the definition of $f(K)_{l_0,j_0}$, the second step follows from differential chain rule, the third step follows from **Part 3** and **Part 5**, the last step follows from simple algebra.

Proof of Part 7 For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$
\begin{aligned}
&\frac{\mathrm{d}c(K,y)_{l_0,j_0,i_0}}{\mathrm{d}K_{i_2,k_2}} \\
&= \frac{\mathrm{d}\langle f(K)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle - b_{l_0,j_0,i_0}}{\mathrm{d}K_{i_2,k_2}} \\
&= \frac{\mathrm{d}\langle f(K)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle}{\mathrm{d}K_{i_2,k_2}} \\
&= \langle \frac{\mathrm{d}f(K)_{l_0,j_0}}{\mathrm{d}K_{i_2,k_2}}, h(y)_{l_0,i_0}\rangle \\
&= \underbrace{h(y)_{l_0,i_0}^\top}_{1\times n} (\underbrace{f(K)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}) - \underbrace{f(K)_{l_0,j_0}}_{n\times 1} \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1} \rangle) \\
&= \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}), \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle - \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1} \rangle
\end{aligned}
$$

where the first step follows from the definition of $c(K,y)_{l_0,j_0,i_0}$, the second step follows from b_{l_0,j_0,i_0} is a constant, the third step follows from only $f(K)_{l_0,j_0}$ is dependent on Q, the fourth step follows from **Part 6**, the last step follows from simple algebra.

Proof of Part 8 For each $i_2 \in [d]$ and $k_2 \in [L]$,

$$
\begin{aligned}
&\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,y)}{\mathrm{d}K_{i_2,k_2}} \\
&= \frac{\mathrm{d}0.5c_{l_0,j_0,i_0}(K,y)^2}{\mathrm{d}K_{i_2,k_2}} \\
&= c_{l_0,j_0,i_0}(K,y) \frac{\mathrm{d}c_{l_0,j_0,i_0}(K,y)}{\mathrm{d}K_{i_2,k_2}} \\
&= \underbrace{c_{l_0,j_0,i_0}(K,y)}_{\text{scalar}} (\langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1} \circ (\underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1}), \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle - \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{h(y)_{l_0,i_0}}_{n\times 1} \rangle \langle \underbrace{f(K)_{l_0,j_0}}_{n\times 1}, \underbrace{\mathsf{A}_{l_0,j_0}}_{n\times d^2} \underbrace{\mathrm{vec}(Qe_{k_2}e_{i_2}^\top)}_{d^2\times 1} \rangle)
\end{aligned}
$$

where the first step follows from the definition of $L_{l_0,j_0,i_0}(K,y)$, the second step follows from simple algebra, the third step follows from **Part 7**. □

E.3 Reformulating Gradient

Lemma E.11. *If the following conditions hold*

- *Let $f(K)_{l_0,j_0}$ be defined as Definition E.3*
- *Let $\frac{\mathrm{d}L_{l_0,j_0,i_0}(:,K)}{\mathrm{d}Q_{i_2,k_2}}$ be compute as* **Part 8** *of Lemma E.10*
- *Let $v_1 := (\mathsf{A}_{l_0,j_0} \mathrm{vec}(Qe_{k_2}e_{i_2}^\top)) \circ h(y)_{l_0,i_0}$*
- *Let $v_2 := h(y)_{l_0,i_0}$*

- *Let* $v_3 := \mathsf{A}_{l_0,j_0} \operatorname{vec}(Q e_{k_2} e_{i_2}^\top)$

then $\frac{\mathrm{d} L_{l_0,j_0,i_0}(x,K)}{\mathrm{d} Q_{i_2,k_2}}$ *can be rewrite as*

$$c_{l_0,j_0,i_0}(x,K)(\langle f(K)_{l_0,j_0}, v_1 \rangle - \langle f(K)_{l_0,j_0}, v_2 \rangle \langle f(K)_{l_0,j_0}, v_3 \rangle)$$

Proof. The proof trivially follows from Fact A.1. □

E.4 Lipschitz of several terms

Lemma E.12. *If the following conditions hold*

- *Let* $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n \times d^2}$
- *Let* $b_{l_0,j_0,i_0} \in \mathbb{R}^n$ *satisfy that* $\|b\|_1 \leq 1$
- *Let* $\beta \in (0, 0.1)$
- *Let* $R \geq 4$
- *Let* $\| \operatorname{vec}(QK^\top) \|_2 \leq R$
- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)), \mathbf{1}_n \rangle \geq \beta$
- $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q} K^\top)), \mathbf{1}_n \rangle \geq \beta$
- *Let* $R_f := \beta^{-2} n \exp(3R^2)$
- *Let* $\alpha(K)_{l_0,j_0}$ *be defined as Definition E.2*
- *Let* $c(K)_{l_0,j_0,i_0}$ *be defined as Definition E.4*
- *Let* $f(K)_{l_0,j_0}$ *be defined as Definition E.3*
- *Let* $v_1 := (\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q e_{k_2} e_{i_2}^\top) h(y)_{l_0,i_0}$
- *Let* $v_2 := h(y)_{l_0,i_0}$
- *Let* $v_3 := \mathsf{A}_{l_0,j_0} \operatorname{vec}(Q e_{k_2} e_{i_2}^\top)$

Then we have

- *Part 1.* $\| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) \|_2 \leq \sqrt{n} \cdot \exp(R^2)$
- *Part 2.* $\|f(K)_{l_0,j_0}\|_2 \leq \beta^{-1} n \exp(2R^2)$
- *Part 3.* $|c(:,K)_{l_0,j_0,i_0}| \leq R \beta^{-1} n \exp(2R^2)$
- *Part 4.* $\|v_2\|_2 \leq R^2$
- *Part 5.* $\|v_3\|_2 \leq R^2$
- *Part 6.* $\|v_1\|_2 \leq R^4$
- *Part 7.* $|(\langle f(K)_{l_0,j_0}, v_1 \rangle - \langle f(K)_{l_0,j_0}, v_2 \rangle \langle f(K)_{l_0,j_0}, v_3 \rangle)| \leq \beta^{-1} n^2 R^4 \exp(6R^2)$
- *Part 8.* $\beta \geq \exp(-R^2)$

Proof. **Proof of Part 1**

$$\| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) \|_2 \leq \sqrt{n} \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top) \|_\infty$$

$$\leq \sqrt{n} \cdot \exp(\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top) \|_\infty)$$
$$\leq \sqrt{n} \cdot \exp(\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top) \|_2)$$
$$\leq \sqrt{n} \cdot \exp(R^2)$$

Proof of Part 2

$$\begin{aligned}\|f(K)_{l_0,j_0}\|_2 &= \|\alpha(K)_{l_0,j_0}^{-1} \cdot u(K)_{l_0,j_0}\|_2 \\ &\leq \|\alpha(K)_{l_0,j_0}^{-1}\|_2 \|u(K)_{l_0,j_0}\|_2 \\ &\leq \beta^{-1} \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top))\|_2 \\ &\leq \beta^{-1}\sqrt{n} \cdot \exp(R^2)\end{aligned}$$

where the first step follows from the definition of $f(x)_{l_0,j_0}$, the second step follows from Fact A.3, the third step follows from $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)), \mathbf{1}_n \rangle \geq \beta$,the fourth step follows from **Part 1**.

Proof of Part 3

$$\begin{aligned}\|c(K,y)_{l_0,j_0,i_0}\| &= \|\langle f(K)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle\| \\ &\leq \|f(K)_{l_0,j_0}\|_2 \|h(y)_{l_0,i_0}\|_2 \\ &\leq R\beta^{-1} n \exp(2R^2)\end{aligned}$$

where the first step follows from the definition of $c(Q,y)_{l_0,j_0,i_0}$, the second step follows from Fact A.3, the third step follows from **Part 2**.

Proof of Part 4

$$\begin{aligned}\|h(y)_{l_0,i_0}\|_2 &= \|A_{l_0,3} y_{i_0}\|_2 \\ &\leq \|A_{l_0,3}\| \|y_{i_0}\|_2 \\ &\leq R^2\end{aligned}$$

where the first step follows from the definition of $h(y)_{l_0,i_0}$, the second step follows from Fact A.4, the third step follows from $\|A_{l_0,3}\|$ and $\|y_{i_0}\|_2 \leq R$.

Proof of Part 5

$$\begin{aligned}\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)\|_2 &\leq \| \mathsf{A}_{l_0,j_0} \| \| \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)\|_2 \\ &\leq R^2\end{aligned}$$

where the first step follows from Fact A.4, the second step follows from $\| \mathsf{A}_{l_0,j_0} \| \leq R$ and Fact A.5.

Proof of Part 6

$$\begin{aligned}\|(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)) \circ h(y)_{l_0,i_0}\|_2 &\leq \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)\|_\infty \|h(y)_{l_0,i_0}\|_2 \\ &\leq \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)\|_2 \|h(y)_{l_0,i_0}\|_2 \\ &\leq R^4\end{aligned}$$

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

Proof of Part 7

$$\begin{aligned}&|\langle f(K)_{l_0,j_0}, (\mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top) h(y)_{l_0,i_0} \rangle - \langle f(K)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(K)_{l_0,j_0}, \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top) \rangle| \\ \leq &|\langle f(K)_{l_0,j_0}, (\mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top) h(y)_{l_0,i_0} \rangle| + |\langle f(K)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(K)_{l_0,j_0}, \operatorname{vec}(Qe_{k_2} e_{i_2}^\top) \rangle| \\ \leq &\|f(K)_{l_0,j_0}\|_2 \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top) h(y)_{l_0,i_0}\|_2 + \|f(K)_{l_0,j_0}\|_2 \|h(y)_{l_0,i_0}\|_2 \|f(K)_{l_0,j_0}\|_2 \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)\|_2 \\ \leq &\beta^{-1} n \exp(2R^2) \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top) h(y)_{l_0,i_0}\|_2 + \beta^{-2} n^2 \exp(4R^2) \|h(y)_{l_0,i_0}\|_2 \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)\|_2 \\ \leq &\beta^{-1} n \exp(2R^2) \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)\|_2 \|h(y)_{l_0,i_0}\|_2 + \beta^{-2} n^2 \exp(4R^2) \|h(y)_{l_0,i_0}\|_2 \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2} e_{i_2}^\top)\|_2 \\ \leq &\beta^{-1} n \exp(2R^2) R^4 + \beta^{-2} n^2 \exp(4R^2) R^4 \\ \leq &\beta^{-1} n^2 R^4 \exp(6R^2)\end{aligned}$$

where the first step follows from triangle inequality, the second step follows from Fact A.3, the third step follows from **Part 2**, the fourth step follows from Fact A.3, the fifth step follows from **Part 5** and **Part 4**, the last step follows from simple algebra.

Proof of Part 8 We have

$$\begin{aligned}\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)), \mathbf{1}_n \rangle \geq & \max_{i\in[n]} \exp(-|(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top))_i|) \\ \geq & \exp(-\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)\|_\infty) \\ \geq & \exp(-\| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)\|_2) \\ \geq & \exp(-R^2)\end{aligned}$$

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

E.5 Lipschitz for several basic terms

Lemma E.13. *If the following conditions hold*

- *Let* $\mathsf{A}_{l_0,j_0} \in \mathbb{R}^{n\times d^2}$
- *Let* $b_{l_0,j_0,i_0} \in \mathbb{R}^n$ *satisfy that* $\|b\|_1 \leq 1$
- *Let* $\beta \in (0, 0.1)$
- *Let* $R \geq 4$
- *Let* $\| \operatorname{vec}(QK^\top)\|_2 \leq R$
- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)), \mathbf{1}_n \rangle \geq \beta$
- $\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(\widehat{Q}K^\top)), \mathbf{1}_n \rangle \geq \beta$
- *Let* $R_f := \beta^{-2} n \exp(3R^2)$
- *Let* $\alpha(Q)_{l_0,j_0}$ *be defined as Definition D.2*
- *Let* $c(Q)_{l_0,j_0,i_0}$ *be defined as Definition D.4*
- *Let* $f(Q)_{l_0,j_0}$ *be defined as Definition D.3*

Then we have

- *Part 1.* $\| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \leq R^2 \exp(R^2) \|K - \widehat{K}\|_F$
- *Part 2.* $|\alpha(K)_{l_0,j_0} - \alpha(\widehat{K})_{l_0,j_0}| \leq \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \cdot \sqrt{n}$
- *Part 3.* $|\alpha(K)^{-1}_{l_0,j_0} - \alpha(\widehat{K})^{-1}_{l_0,j_0}| \leq \beta^{-2} \cdot |\alpha(K)_{l_0,j_0} - \alpha(\widehat{K})_{l_0,j_0}|$
- *Part 4.* $\|f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}\|_2 \leq \beta^{-2} n \exp(3R^2) \|K - \widehat{K}\|_F$
- *Part 5.* $\|c(K,:)_{l_0,j_0,i_0} - c(\widehat{K},:)_{l_0,j_0,i_0}\|_2 \leq R^2 \beta^{-2} n \exp(3R^2) \|K - \widehat{K}\|_2$

Note that $\|K\|_F = (\sum_i \sum_j K_{i,j}^2)^{1/2} = \| \operatorname{vec}(K)\|_2$

Proof. **Proof of Part 1.**

$$\begin{aligned}\| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \leq & \exp(R^2) \| \mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top) - \mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top)\|_2 \\ \leq & \exp(R^2) \| \mathsf{A}_{l_0,j_0} \| \| \operatorname{vec}(QK^\top) - \operatorname{vec}(Q\widehat{K}^\top)\|_2\end{aligned}$$

$$
\begin{aligned}
&= \exp(R^2) \| \mathsf{A}_{l_0,j_0} \| \|QK^\top - Q\widehat{K}^\top\|_F \\
&\leq \exp(R^2) \| \mathsf{A}_{l_0,j_0} \| \|Q\|_F \|K - \widehat{K}\|_F \\
&\leq R^2 \exp(R^2) \|K - \widehat{K}\|_F
\end{aligned}
$$

Proof of Part 2.

$$
\begin{aligned}
|\alpha(K)_{l_0,j_0} - \alpha(\widehat{K})_{l_0,j_0}| &= |\langle \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top)), \mathbf{1}_n \rangle| \\
&\leq \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \cdot \sqrt{n}
\end{aligned}
$$

Proof of Part 3.

$$
\begin{aligned}
|\alpha(K)^{-1}_{l_0,j_0} - \alpha(\widehat{K})^{-1}_{l_0,j_0}| &= \alpha(K)^{-1}_{l_0,j_0} \alpha(\widehat{K})^{-1}_{l_0,j_0} \cdot |\alpha(K)_{l_0,j_0} - \alpha(\widehat{K})_{l_0,j_0}| \\
&\leq \beta^{-2} \cdot |\alpha(K)_{l_0,j_0} - \alpha(\widehat{K})_{l_0,j_0}|
\end{aligned}
$$

Proof of Part 4. We can show that

$$
\begin{aligned}
\|f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}\|_2 &= \|\alpha(K)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \alpha(\widehat{K})^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \\
&\leq \|\alpha(K)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \alpha(K)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \\
&\quad + \|\alpha(K)^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top)) - \alpha(\widehat{K})^{-1}_{l_0,j_0} \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \\
&\leq \alpha(K)^{-1}_{l_0,j_0} \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \\
&\quad + |\alpha(K)^{-1}_{l_0,j_0} - \alpha(\widehat{K})^{-1}_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2
\end{aligned}
$$

where the 1st step follows from the definition of $f(K)_{l_0,j_0}$ and $\alpha(K)_{l_0,j_0}$, the 2nd step follows from triangle inequality (**Part 3** of Fact A.3), the 3rd step follows from $\|\alpha A\| \leq |\alpha| \|A\|$(**Part 5** of Fact A.4).

For the first term in the above, we have

$$
\begin{aligned}
&\alpha(K)^{-1}_{l_0,j_0} \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \qquad (7) \\
\leq\ & \beta^{-1} \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \\
\leq\ & \beta^{-1} \cdot R^2 \exp(R^2) \cdot \|K - \widehat{K}\|_F \qquad (8)
\end{aligned}
$$

where the 1st step follows from $\alpha(K)_{l_0,j_0} \geq \beta$, the 2nd step follows from **Part 1**.

For the second term in the above, we have

$$
\begin{aligned}
&|\alpha(K)^{-1}_{l_0,j_0} - \alpha(\widehat{K})^{-1}_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \\
\leq\ & \beta^{-2} \cdot |\alpha(K)_{l_0,j_0} - \alpha(\widehat{K})_{l_0,j_0}| \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \\
\leq\ & \beta^{-2} \cdot |\alpha(K)_{l_0,j_0} - \alpha(\widehat{K})_{l_0,j_0}| \cdot \sqrt{n} \exp(R^2) \\
\leq\ & \beta^{-2} \cdot \sqrt{n} \cdot \| \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(QK^\top)) - \exp(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Q\widehat{K}^\top))\|_2 \cdot \sqrt{n} \exp(R^2) \\
\leq\ & \beta^{-2} \cdot \sqrt{n} \cdot R^2 \exp(R^2) \|K - \widehat{K}\|_F \cdot \sqrt{n} \exp(R^2) \\
=\ & \beta^{-2} \cdot nR^2 \exp(2R^2) \|K - \widehat{K}\|_F \qquad (9)
\end{aligned}
$$

where the 1st step follows from the result of **Part 3**, the 2nd step follows from **Part 1** of Lemma D.12, 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. (7) and Eq. (9) together, we have

$$
\begin{aligned}
\|f_{l_0,j_0}(K) - f_{l_0,j_0}(\widehat{K})\|_2 &\leq \beta^{-1} \cdot R^2 \exp(R^2) \cdot \|K - \widehat{K}\|_F + \beta^{-2} \cdot nR^2 \exp(2R^2) \|K - \widehat{K}\|_F \\
&\leq 2\beta^{-2} nR^2 \exp(2R^2) \|K - \widehat{K}\|_F \\
&\leq \beta^{-2} n \exp(3R^2) \|K - \widehat{K}\|_F
\end{aligned}
$$

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.

$$\begin{aligned}
\|c(K,:)_{l_0,j_0,i_0} - c(\widehat{K},:)_{l_0,j_0,i_0}\|_2 &= \|\langle f(K)_{l_0,j_0}, h(:)_{l_0,i_0}\rangle - \langle f(\widehat{K})_{l_0,j_0}, h(:)_{l_0,i_0}\rangle\|_2 \\
&= \|\langle (f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}), h(:)_{l_0,i_0}\rangle\|_2 \\
&\leq \|h(:)_{l_0,i_0}\|_2 \|f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}\|_2 \\
&\leq \|\mathsf{A}_{l_0,3}\, y_{i_0}\|_2 \|f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}\|_2 \\
&\leq \|\mathsf{A}_{l_0,3}\, y_{i_0}\|_2 \cdot \beta^{-2} n \exp(3R^2) \|K - \widehat{K}\|_F \\
&\leq \|\mathsf{A}_{l_0,3}\| \|y_{i_0}\|_2 \beta^{-2} n \exp(3R^2) \|K - \widehat{K}\|_F \\
&\leq R\beta^{-2} n \exp(3R^2) \|K - \widehat{K}\|_F
\end{aligned}$$

where the first step follows from the definition of $c(x,y)_{l_0,j_0,i_0}$, the second step follows from simple algebra, the third step follows from Fact A.3, the fourth step follows from the definition of $h(y)_{l_0,i_0}$, the fifth step follows from **Part 4**, the sixth step follows from Fact A.4, the last step follows from $\|A_{l_0,3}\| \leq R$ and $\|z_{i_0}\|_2 \leq R$ □

E.6 Summary of 3 steps

Lemma E.14. *If the following conditions hold*

- *Let $f(K)_{l_0,j_0}$ be defined as Definition E.3*
- *Let $\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}}$ be compute as* **Part 8** *of Lemma E.10*
- *Let $v_1 := (\mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2}e_{i_2}^\top)) \circ h(y)_{l_0,i_0}$*
- *Let $v_2 := h(y)_{l_0,i_0}$*
- *Let $v_3 := \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2}e_{i_2}^\top)$*

Then we have

- $|\langle f(K)_{l_0,j_0}, v_1\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_1\rangle| \leq \beta^{-2} n R^4 \exp(3R^2) \|K - \widehat{K}\|_F$
- $|\langle f(K)_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle\langle f(\widehat{K})_{l_0,j_0}, v_3\rangle| \leq 2\beta^{-3} n \exp(2R^2) R^6 n \exp(3R^2) \|K - \widehat{K}\|_F$
- $|\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}}|_{K=K} - \frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}}|_{K=\widehat{K}}| \leq \beta^{-3} n^3 R^7 \exp(19R^2) \|K - \widehat{K}\|_F$

Proof. **Proof of Part 1.**

$$\begin{aligned}
|\langle f(K)_{l_0,j_0}, v_1\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_1\rangle| &= |\langle f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}, v_1\rangle| \\
&\leq \|f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}\|_2 \|v_1\|_2 \\
&\leq \beta^{-2} n \exp(3R^2) \|K - \widehat{K}\|_F \|(\mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2}e_{i_2}^\top)) \circ h(y)_{l_0,i_0}\|_2 \\
&\leq \beta^{-2} n R^4 \exp(3R^2) \|K - \widehat{K}\|_F
\end{aligned}$$

where the first step follows from simple algebra, the second step follows from Fact A.3, the third step follows from **Part 4** of Lemma E.13, the fourth step follows from **Part 6** of Lemma D.12.

Proof of Part 2. For convenience, we define

$$\begin{aligned}
C_1 &:= \langle f(K)_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle \\
C_2 &:= \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle\langle f(\widehat{K})_{l_0,j_0}, v_3\rangle
\end{aligned}$$

Then it's apparent that

$$|C_1 + C_2| = |\langle f(K)_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle\langle f(\widehat{K})_{l_0,j_0}, v_3\rangle|$$

Since C_1 and C_2 are similar, we only need to bound $|C_1|$:

$$\begin{aligned}
|C_1| &= |\langle f(K)_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle| \\
&= |\langle f(K)_{l_0,j_0}, v_3\rangle(\langle f(K)_{l_0,j_0}, v_2\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle)| \\
&\leq |\langle f(K)_{l_0,j_0}, v_3\rangle||\langle f(K)_{l_0,j_0}, v_2\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle| \\
&= |\langle f(K)_{l_0,j_0}, v_3\rangle||\langle f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}, v_2\rangle| \\
&\leq \|f(K)_{l_0,j_0}\|_2\|v_3\|_2\|f(K)_{l_0,j_0} - f(\widehat{K})_{l_0,j_0}\|_2\|v_2\|_2 \\
&\leq \beta^{-1}n\exp(2R^2)\|v_3\|_2\beta^{-2}n\exp(3R^2)\|K - \widehat{K}\|_F\|v_2\|_2 \\
&\leq \beta^{-3}n^2R^6\exp(5R^2)\|K - \widehat{K}\|_F
\end{aligned}$$

where the first step follows from the definition of C_1, the second step follows from simple algebra, the third step follows from triangular inequality, the fourth step follows from simple algebra, the fifth step follows from Fact A.3, the sixth step follows from **Part 4** of Lemma E.13, the last step follows from **Part 4** and **Part 5** of Lemma E.12.

Thus, we obtained the bound for $|\langle f(K)_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle\langle f(\widehat{K})_{l_0,j_0}, v_3\rangle|$:

$$|\langle f(K)_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle\langle f(\widehat{K})_{l_0,j_0}, v_3\rangle| \leq 2\beta^{-3}n^2R^6\exp(5R^2)\|K - \widehat{K}\|_F$$

Proof of Part 3 By Lemma E.11, we know that $\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}}$ can be written as

$$\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}Q_{i_2,k_2}} = c_{l_0,j_0,i_0}(K,y)(\langle f(K)_{l_0,j_0}, v_1\rangle - \langle f(K)_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle)$$

For convenience, we define

$$\begin{aligned}
s(K) &:= c_{l_0,j_0,i_0}(K,y) \\
t(K) &:= (\langle f(K)_{l_0,j_0}, v_1\rangle - \langle f(K)_{l_0,j_0}, v_2\rangle\langle f(K)_{l_0,j_0}, v_3\rangle)
\end{aligned}$$

Thus $\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}}$ can be rewrite as

$$\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}} = s(K)t(K)$$

Then the lipschitz of $\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}}$ can be expressed as

$$|\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}} - \frac{\mathrm{d}L_{l_0,j_0,i_0}(\widehat{K},:)}{\mathrm{d}K_{i_2,k_2}}| = |s(K)t(K) - s(\widehat{K})t(\widehat{K})|$$

Use the same techinque in the proof of **Part 2**, we define

$$\begin{aligned}
C_1 &:= s(K)t(K) - s(K)t(\widehat{K}) \\
C_2 &:= s(K)t(\widehat{K}) - s(\widehat{K})t(\widehat{K})
\end{aligned}$$

Then it's apparent that

$$|s(K)t(K) - s(\widehat{K})t(\widehat{K})| = |C_1 + C_2|$$

First, we upper bound $|C_1|$ as follows:

$|C_1|$

$$
\begin{aligned}
&= |s(K)t(K) - s(K)t(\widehat{K})| \\
&= |s(K)(t(K) - t(\widehat{K}))| \\
&\leq |s(K)||t(K) - t(\widehat{K})| \\
&= |s(K)||(\langle f(K)_{l_0,j_0}, v_1\rangle - \langle f(K)_{l_0,j_0}, v_2\rangle \langle f(K)_{l_0,j_0}, v_3\rangle) - (\langle f(\widehat{K})_{l_0,j_0}, v_1\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_2\rangle \langle f(\widehat{K})_{l_0,j_0}, v_3\rangle)| \\
&= |s(K)||(\langle f(K)_{l_0,j_0}, v_1\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_1\rangle) + (\langle f(\widehat{K})_{l_0,j_0}, v_2\rangle \langle f(\widehat{K})_{l_0,j_0}, v_3\rangle - \langle f(K)_{l_0,j_0}, v_2\rangle \langle f(K)_{l_0,j_0}, v_3\rangle)| \\
&\leq |s(K)|(|\langle f(K)_{l_0,j_0}, v_1\rangle - \langle f(\widehat{K})_{l_0,j_0}, v_1\rangle| + |\langle f(\widehat{K})_{l_0,j_0}, v_2\rangle \langle f(\widehat{K})_{l_0,j_0}, v_3\rangle - \langle f(K)_{l_0,j_0}, v_2\rangle \langle f(K)_{l_0,j_0}, v_3\rangle|) \\
&\leq |s(K)|(\beta^{-2} n R^4 \exp(3R^2)\|K - \widehat{K}\|_F\| + 2\beta^{-3} n^2 R^6 \exp(5R^2)\|K - \widehat{K}\|_F) \\
&\leq |s(K)| \cdot \beta^{-2} n^2 R^6 \exp(8R^2)\|K - \widehat{K}\|_F \\
&\leq \beta^{-3} n^3 R^7 \exp(10R^2)\|K - \widehat{K}\|_F
\end{aligned}
$$

where the first step follows from the definition of C_1, the second step follows from simple algebra, the third step follows from Fact A.3, the fourth step follows from the definition of $t(K)$, the fifth step follows from simple algebra, the sixth step follows from triangular inequality, the seventh step follows from **Part 3** and **Part 4** the last step follows from **Part 3** of Lemma D.12.

Next, we upper bound $|C_2|$:

$$
\begin{aligned}
|C_2| &= |s(K)t(\widehat{K}) - s(\widehat{K})t(\widehat{K})| \\
&= |(s(K) - s(\widehat{K}))t(\widehat{K})| \\
&\leq |s(K) - s(\widehat{K})||t(\widehat{K}| \\
&\leq R^2 \beta^{-2} n \exp(3R^2)\|K - \widehat{K}\|_F |t(\widehat{K}| \\
&\leq R^6 \beta^{-3} n^3 \exp(9R^2)\|K - \widehat{K}\|_F
\end{aligned}
$$

where the first step follows from the definition of C_2, the second step follows from simple algebra, the third step follows from simple algebra, the fourth step follows from **Part 5** of Lemma E.13, the last step follows from **Part 7** of Lemma E.12.

Thus, we can obtain the upper bound for $|s(K)t(K) - s(\widehat{K})t(\widehat{K})|$:

$$
\begin{aligned}
|s(K)t(K) - s(\widehat{K})t(\widehat{K})| &= |C_1 + C_2| \\
&\leq \beta^{-3} n^3 R^7 \exp(10R^2)\|K - \widehat{K}\|_F + R^6 \beta^{-3} n^3 \exp(9R^2)\|K - \widehat{K}\|_F \\
&\leq \beta^{-3} n^3 R^7 \exp(19R^2)\|K - \widehat{K}\|_F
\end{aligned}
$$

where the first step follows from simple algebra, the second step follows from the upper bound of $|C_1|$ and $|C_2|$, the last step follows from simple algebra. □

E.7 Lipschitz of $\nabla L_{l_0,j_0,i_0}(K,:)$

Lemma E.15. *If the following conditions hold*

- *Let $f(Q)_{l_0,j_0}$ be defined as Definition E.3*
- *Let $\frac{\mathrm{d}L_{l_0,j_0,i_0}(K,:)}{\mathrm{d}K_{i_2,k_2}}$ be compute as* **Part 8** *of Lemma E.10*
- *Let $v_1 := (\mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2}e_{i_2}^\top)) \circ h(y)_{l_0,i_0}$*
- *Let $v_2 := h(y)_{l_0,i_0}$*
- *Let $v_3 := \mathsf{A}_{l_0,j_0} \operatorname{vec}(Qe_{k_2}e_{i_2}^\top)$*

Then we have

$$
\|\frac{\mathrm{d}L(K,:)}{\mathrm{d}\operatorname{vec}(K)} - \frac{\mathrm{d}L(\widehat{K},:)}{\mathrm{d}\operatorname{vec}(K)}\|_2 \leq dLn^3 R^7 \exp(22R^2)\|K - \widehat{K}\|_F
$$

Proof.

$$\begin{aligned}
\|\frac{\mathrm{d}L(K,:)}{\mathrm{d}\,\mathrm{vec}(K)} - \frac{\mathrm{d}L(\widehat{K},:)}{\mathrm{d}\,\mathrm{vec}(K)}\|_2 \leq & \sum_{i_2=1}^{d}\sum_{k_2=1}^{L} |\frac{\mathrm{d}L(K,:)}{\mathrm{d}K_{i_2,k_2}}\Big|_{K=K} - \frac{\mathrm{d}L(K,:)}{\mathrm{d}K_{i_2,k_2}}\Big|_{K=\widehat{K}}| \\
\leq & \sum_{i_2=1}^{d}\sum_{k_2=1}^{L} \beta^{-3}n^3R^7\exp(19R^2)\|K-\widehat{K}\|_F \\
= & \beta^{-3}dLn^3R^7\exp(19R^2)\|K-\widehat{K}\|_F \\
\leq & dLn^3R^7\exp(22R^2)\|K-\widehat{K}\|_F
\end{aligned}$$

where the first step follows from Fact A.3, the second step follows from **Part 3** of Lemma E.14, the fourth step follows from simple algebra, the last step follows from **Part 8** of Lemma E.12. □

F ANALYSIS ON LOGISTIC FUNCTION

In this section, we provide systematic analysis on logistic function. In Section F.1, we compute the gradient of the loss function based on logistic function. In Section F.2, we prove the lipschitz property of gradient.

F.1 GRADIENT WITH RESPECT TO x

Fact F.1. *If the following conditions hold*

- *Let $g(x) : \mathbb{R} \to \mathbb{R}$ be defined in Definition 3.5*

Then we have

$$\frac{\mathrm{d}g(x)}{\mathrm{d}x} = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

Further more, we have

$$\frac{\mathrm{d}g(x)}{\mathrm{d}x} = g(x)(1-g(x))$$

Lemma F.2 (Formal version of Lemma 3.11)**.** *If the following conditions hold*

- *Let $L(x,y)_{l_0,j_0,i_0}$ be defined as Definition 3.6*
- *Let $f(x)_{l_0,j_0}$ be defined in Definition 3.3*
- *Let $h(y)_{l_0,i_0}$ be defined in Definition 3.4*

Then we have

$$\begin{aligned}
&\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}x_i} \\
=\ & g(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)(1-g(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle))b_{l_0,j_0,i_0} \\
& \cdot (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0}\rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle)
\end{aligned}$$

Proof. For $\forall i \in [d^2]$,

$$\begin{aligned}
\frac{\mathrm{d}L(x,y)_{l_0,j_0,i_0}}{\mathrm{d}x_i} &= \frac{\mathrm{d}}{\mathrm{d}x_i} g(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle b_{l_0,j_0,i_0}) \\
&= \frac{\mathrm{d}g(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle b_{l_0,j_0,i_0})}{\mathrm{d}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle} \frac{\mathrm{d}\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle b_{l_0,j_0,i_0}}{\mathrm{d}x_i} \\
&= g(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)(1-g(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle))b_{l_0,j_0,i_0} \\
&\quad \cdot (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0}\rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i}\rangle)
\end{aligned}$$

where the first step follows from the definition of $L(x,y)_{l_0,j_0,i_0}$, the second step follows from differential chain rule, the last step follows from Lemma F.1 and the computations in **Part 8** of Lemma B.6. □

F.2 GRADIENT LIPSCHITZ WITH RESPECT TO x

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

- *Let $g(x)$ be defined in Definiton 3.5*
- *Let $|x| \leq R$*
- *Let $R \geq 4$*

Then we have

- *Part 1.* $|g(x) - g(\widehat{x})| \leq \exp(R)|x - \widehat{x}|$
- *Part 2.* $|g^2(x) - g^2(\widehat{x})| \leq 2\exp(2R)|x - \widehat{x}|$
- *Part 3.* $|g'(x) - g'(\widehat{x})| \leq 3\exp(2R)|x - \widehat{x}|$

Proof. **Proof of Part 1**

$$\begin{aligned}
|g(x) - g(\widehat{x})| = & |\frac{1}{1+\exp(-x)} - \frac{1}{1+\exp(-\widehat{x})}| \\
= & |\frac{\exp(-\widehat{x}) - \exp(\widehat{-x})}{1+\exp(-\widehat{x}) + \exp(-x) + \exp(-x-\widehat{x})}| \\
\leq & |\exp(-\widehat{x}) - \exp(-x)| \\
\leq & |\exp(-x)||x - \widehat{x}| \\
\leq & \exp(R)|x - \widehat{x}|
\end{aligned}$$

where the first step follows from the definition of $g(x)$, the second step follows from simple algebra, the third step follows from simple algebra, the fourth step follows from Fact A.3, the fifth step follows from $|x| \leq R$.

Proof of Part 2

$$\begin{aligned}
|g^2(x) - g^2(\widehat{x})| = & |g(x) - g(\widehat{x})||g(x) + g(\widehat{x})| \\
\leq & \exp(R)|x - \widehat{x}|2\exp(R) \\
= & 2\exp(2R)|x - \widehat{x}|
\end{aligned}$$

where the first step follows from simple algebra, the second step follows from **Part 1** and $|x| \leq R$, the last step follows from simple algebra.

Proof of Part 3

$$\begin{aligned}
|g'(x) - g'(\widehat{x})| = & |(g(x) - g^2(x)) - (g(\widehat{x}) - g^2(\widehat{x}))| \\
= & |(g(x) - g(\widehat{x})) + (g^2(\widehat{x}) - g^2(x))| \\
\leq & |g(x) - g(\widehat{x})| + |g^2(x) - g^2(\widehat{x})| \\
\leq & 3\exp(2R)|x - \widehat{x}|
\end{aligned}$$

where the first step follows from simple algebra, the second step follows from simple algebra, the third step follows from triangular inequality, the last step follows from **Part 1** and **Part 2**. □

Lemma F.4 (Formal version of Lemma 3.12)**.** *If the following conditions hold*

- *Let $g(x)$ be defined in Definition 3.5*

Then we have

$$|g'(x) - g'(\widehat{x})| \leq |x - \widehat{x}|$$

Proof. We can easily bound $g''(x)$ as follows:

$$\begin{aligned}
g''(x) = & (g(x) - g(x)^2)' \\
= & g'(x) - 2g(x)g'(x) \\
= & g(x) - g^2(x) - 2g(x)(g(x) - g^2(x)) \\
= & g(x) - g^2(x) - (2g^2(x) - 2g^3(x)) \\
= & g(x) - 3g^2(x) + 2g^3(x) \\
\leq & 1
\end{aligned}$$

Then by Lagrange's mean value theorem, we have

$$|g'(x) - g'(\widehat{x})| \leq 1 \cdot |x - \widehat{x}|$$

□

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

- *Let $f(x)_{l_0,j_0}$ be defined in Definition 3.3*
- *Let $h(y)_{l_0,i_0}$ be defined in Definition 3.4*
- *Let $d(x) := \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle$*
- *Let $R \geq 4$*
- *Let $x, y \in \mathbb{R}^d$ satisfy $\| \mathsf{A}_{l_0,j_0} x \|_2 \leq R$ and $\| \mathsf{A}_{l_0,j_0} y \|_2 \leq R$*
- *$\| \mathsf{A}_{l_0,j_0} \| \leq R$*

Then we have

$$|d(x) - d(\widehat{x})| \leq nR^2 \exp(5R^2) \|x - \widehat{x}\|_2$$

Proof.

$$\begin{aligned}
|d(x) - d(\widehat{x})| &= |\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle - \langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0} \rangle| \\
&= |\langle f(x)_{l_0,j_0} - f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0} \rangle| \\
&\leq \|h(y)_{l_0,i_0}\|_2 \|f(x)_{l_0,j_0} - f(\widehat{x})_{l_0,j_0}\|_2 \\
&\leq \|A_{l_0,3} y_{i_0}\|_2 \beta^{-2} n \exp(3R^2) \|x - \widehat{x}\|_2 \\
&\leq R^2 \beta^{-2} n \exp(3R^2) \|x - \widehat{x}\|_2 \\
&\leq R^2 n \exp(5R^2) \|x - \widehat{x}\|_2
\end{aligned}$$

where the first step follows from the definition of $d(x)$, the second step follows from simple algebra, the third step follows from Fact A.3, the fourth step follows from the definition of $h(y)_{l_0,i_0}$ and **Part 4** of Lemma C.3, the fifth step follows from $\|A_{l_0,3}\| \leq R$ and $\|y\|_2 \leq R$, the last step follows from Lemma C.4. □

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

- *Let $f(x)_{l_0,j_0}$ be defined in Definition 3.3*
- *Let $h(y)_{l_0,i_0}$ be defined in Definition 3.4*
- *Let $L(x,y)_{l_0,j_0,i_0}$ be defined as Definition 3.6*
- *Let ∇L be computed as Lemma F.2*

Then we can rewrite ∇L as

$$g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) b_{l_0,j_0,i_0} \cdot (\langle f(x)_{l_0,j_0}, v_1 \rangle - \langle f(x)_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle)$$

where

$$\begin{aligned}
v_1 &:= h(y)_{l_0,i_0} \circ \mathsf{A}_{l_0,j_0,i} \\
v_2 &:= h(y)_{l_0,i_0} \\
v_3 &:= \mathsf{A}_{l_0,j_0,i}
\end{aligned}$$

Proof.

$$\begin{aligned}
& g(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle)(1 - g(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle)) b_{l_0,j_0,i_0} \\
& \quad \cdot (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \\
= & \, g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) b_{l_0,j_0,i_0} \\
& \quad \cdot (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \\
= & \, g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) b_{l_0,j_0,i_0} \\
& \quad \cdot (\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \circ \mathsf{A}_{l_0,j_0,i} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \\
= & \, g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) b_{l_0,j_0,i_0} \cdot (\langle f(x)_{l_0,j_0}, v_1 \rangle - \langle f(x)_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle)
\end{aligned}$$

where the first step follows from simple derivative, the second step follows from simple algebra, the third step follows from the definition of v_1, v_2 and v_3. □

Lemma F.7. *If the following conditions hold*

- *Let* $f(x)_{l_0,j_0}$ *be defined in Definition 3.3*
- *Let* $h(y)_{l_0,i_0}$ *be defined in Definition 3.4*
- *Let* $R \geq 4$
- *Let* $x, y \in \mathbb{R}^d$ *satisfy* $\| \mathsf{A}_{l_0,j_0} x\|_2 \leq R$ *and* $\| \mathsf{A}_{l_0,j_0} y\|_2 \leq R$
- $\| \mathsf{A}_{l_0,j_0} \| \leq R$
- *Let* $v_1 := h(y)_{l_0,i_0} \circ \mathsf{A}_{l_0,j_0,i}$
- *Let* $v_2 := h(y)_{l_0,i_0}$
- *Let* $v_3 := \mathsf{A}_{l_0,j_0,i}$

Then we have

- *Part 1.* $\|v_2\|_2 \leq R^2$
- *Part 2.* $\|v_3\|_2 \leq R$
- *Part 3.* $\|v_1\|_2 \leq R^3$
- *Part 4.* $\| \exp(\mathsf{A}_{l_0,j_0} x)\|_2 \leq \sqrt{n} \exp(R^2)$
- *Part 5.* $\|f(x)_{l_0,j_0}\|_2 \leq \beta^{-1} n \exp(2R^2)$

Proof. **Proof of Part 1**

$$\begin{aligned} \|v_2\| &= \|h(y)_{l_0,i_0}\|_2 \\ &= \|A_{l_0,3} y_{i_0}\| \\ &\leq \|A_{l_0,3}\| \|y_{i_0}\|_2 \\ &\leq R^2 \end{aligned}$$

where the first step folllows from the definition of v_2, the second step follows from the definition of $h(y)_{l_0,i_0}$, the third step follows from Fact A.4, the last step follows from $\|A_{l_0,3}\| \leq R$ and $\|y\|_2 \leq R$.

Proof of Part 2 This trivially follows from $\|A_{l_0,3}\| \leq R$.

Proof of Part 3

$$\begin{aligned} \|v_1\|_2 &= \|h(y)_{l_0,i_0} \circ \mathsf{A}_{l_0,j_0,i} \|_2 \\ &\leq \|h(y)_{l_0,i_0}\|_2 \| \mathsf{A}_{l_0,j_0,i} \|_2 \\ &\leq R^3 \end{aligned}$$

where the first step follows from the definition of v_1, the second step follows from Fact A.3, the third step follows from **Part 1** and $\| \mathsf{A}_{l_0,j_0,i} \| \leq R$.

Proof of Part 4 We can show that

$$\begin{aligned} \| \exp(\mathsf{A}_{l_0,j_0} x)\|_2 &\leq \sqrt{n} \cdot \| \exp(\mathsf{A}_{l_0,j_0} x)\|_\infty \\ &\leq \sqrt{n} \cdot \exp(\| \mathsf{A}_{l_0,j_0} x\|_\infty) \\ &\leq \sqrt{n} \cdot \exp(\| \mathsf{A}_{l_0,j_0} x\|_2) \\ &\leq \sqrt{n} \cdot \exp(R^2), \end{aligned}$$

where the first step follows from **Part 4** of Fact A.3, the second step follows from **Part6** of Fact A.3, the third step follows from Fact A.3, and the last step follows from $\| \mathsf{A}_{l_0,j_0} \| \leq R$ and $\|x\|_2 \leq R$.

Proof of Part 5

$$\begin{aligned}
\|f(x)_{l_0,j_0}\|_2 &= \|\alpha(x)_{l_0,j_0}^{-1} \cdot u(x)_{l_0,j_0}\|_2 \\
&\leq \|\alpha(x)_{l_0,j_0}^{-1}\|_2 \|u(x)_{l_0,j_0}\|_2 \\
&\leq \beta \|\alpha(x)_{l_0,j_0}\| \exp(\mathsf{A}_{l_0,j_0} x)\|_2 \\
&\leq \beta^{-1} \|\langle \exp(\mathsf{A}_{l_0,j_0} x), \mathbf{1}_n \rangle\|_2 \sqrt{n} \cdot \exp(R^2) \\
&\leq \beta^{-1} \|\exp(\mathsf{A}_{l_0,j_0} x)\|_2 \|\mathbf{1}_n\|_2 \sqrt{n} \cdot \exp(R^2) \\
&\leq \beta^{-1} \sqrt{n} \cdot \exp(R^2) \sqrt{n} \cdot \exp(R^2) \\
&= \beta^{-1} n \exp(2R^2)
\end{aligned}$$

where the first step follows from the definition of $f(x)_{l_0,j_0}$, the second step follows from Fact A.3, the third step follows from $\langle \exp(\mathsf{A}_{l_0,j_0} x), \mathbf{1}_n \rangle \geq \beta$, the fourth step follows from **Part 4**, the fifth step follows from Fact A.3, the sixth step follows from **Part 4**, the last step follows from simple algebra. □

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

- *Let $f(x)_{l_0,j_0}$ be defined in Definition 3.3*
- *Let $h(y)_{l_0,i_0}$ be defined in Definition 3.4*
- *Let $R \geq 4$*
- *Let $x, y \in \mathbb{R}^d$ satisfy $\| \mathsf{A}_{l_0,j_0} x\|_2 \leq R$ and $\| \mathsf{A}_{l_0,j_0} y\|_2 \leq R$*
- *$\| \mathsf{A}_{l_0,j_0} \| \leq R$*
- *Let $v_1 := h(y)_{l_0,i_0} \circ \mathsf{A}_{l_0,j_0,i}$*
- *Let $v_2 := h(y)_{l_0,i_0}$*
- *Let $v_3 := \mathsf{A}_{l_0,j_0,i}$*
- *Let $s(x) := \langle f(x)_{l_0,j_0}, v_1 \rangle$*
- *Let $t(x) := \langle f(x)_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle$*

Then we have

- *Part 1. $|s(x) - s(\widehat{x})| \leq nR^2 \exp(5R^2) \|x - y\|_2$*
- *Part 2. $|t(x) - t(\widehat{x})| \leq 2n^2 R^4 \exp(8R^2) \|x - y\|_2$*
- *Part 3. $|(s(x) - t(x)) - (s(\widehat{x}) - t(\widehat{x}))| \leq n^2 R^4 \exp(13R^2) \|x - y\|_2$*

Proof. **Proof of Part 1**

$$\begin{aligned}
|s(x) - s(\widehat{x})| &= |\langle f(x)_{l_0,j_0}, v_1 \rangle - \langle f(\widehat{x})_{l_0,j_0}, v_1 \rangle| \\
&= |\langle f(x)_{l_0,j_0} - f(\widehat{x})_{l_0,j_0}, v_1 \rangle| \\
&\leq \|f(x)_{l_0,j_0} - f(\widehat{x})_{l_0,j_0}\|_2 \|v_1\|_2 \\
&\leq nR^2 \exp(5R^2) \|x - y\|_2
\end{aligned}$$

where the first step follows from the definition of $s(x)$, the second step follows from simple algebra, the third step follows from Fact A.3, the last step follows from combining **Part 4** of Lemma C.3, Lemma C.4 and **Part 1** of Lemma F.7.

Proof of Part 2 First, note that

$$|t(x) - t(\widehat{x})| = |\langle f(x)_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{x})_{l_0,j_0}, v_2 \rangle \langle f(\widehat{x})_{l_0,j_0}, v_3 \rangle|$$

For convenience, we define

$$C_1 := \langle f(x)_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{x})_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle$$

$$C_2 := \langle f(\widehat{x})_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{x})_{l_0,j_0}, v_2 \rangle \langle f(\widehat{x})_{l_0,j_0}, v_3 \rangle$$

Then, it's easy to know

$$|t(x) - t(\widehat{x})| = |C_1 + C_2|$$

Since C_1 and C_2 are symmetry, we only need to upper bound $|C_1|$:

$$\begin{aligned} |C_1| &= |\langle f(x)_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle - \langle f(\widehat{x})_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle| \\ &= |\langle f(x)_{l_0,j_0} - f(\widehat{x})_{l_0,j_0}, v_2 \rangle| |\langle f(x)_{l_0,j_0}, v_3 \rangle| \\ &\leq \|f(x)_{l_0,j_0} - f(\widehat{x})_{l_0,j_0}\|_2 \|v_2\|_2 \|f(x)_{l_0,j_0}\|_2 \|v_3\|_2 \\ &\leq n^2 R^4 \exp(8R^2) \|x - y\|_2 \end{aligned}$$

where the first step follows from the definition of C_1, the second step follows from simple algebra, the third step follows from Fact A.3, the fourth step follows from combining **Part 4** of Lemma C.3, **Part 5**, **Part 1** and **Part 3** of Lemma F.7.

Thus, we obtained the bound:

$$\begin{aligned} |t(x) - t(\widehat{x})| &= |C_1 + C_2| \\ &\leq 2n^2 R^4 \exp(8R^2) \|x - y\|_2 \end{aligned}$$

Proof of Part 3

$$\begin{aligned} |(s(x) - t(x)) - (s(\widehat{x}) - t(\widehat{x}))| &= |(s(x) - s(\widehat{x}) + (t(\widehat{x}) - t(x)))| \\ &\leq |s(x) - s(\widehat{x})| + |t(\widehat{x}) - t(x)| \\ &\leq n^2 R^4 \exp(13R^2) \|x - y\|_2 \end{aligned}$$

where the first step follows from simple algebra, the second step follows from triangular inequality, the last step follows from **Part 2** and **Part 3**. □

Lemma F.9 (Formal version of Lemma 3.13)**.** *If the following conditions hold*

- *Let $f(x)_{l_0,j_0}$ be defined in Definition 3.3*
- *Let $h(y)_{l_0,i_0}$ be defined in Definition 3.4*
- *Let $d(x) := \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle$*
- *Let $R \geq 4$*
- *Let $x, y \in \mathbb{R}^d$ satisfy $\| \mathsf{A}_{l_0,j_0} x \|_2 \leq R$ and $\| \mathsf{A}_{l_0,j_0} y \|_2 \leq R$*
- *$\| \mathsf{A}_{l_0,j_0} \| \leq R$*
- *Let $L(x,y)_{l_0,j_0,i_0}$ be defined in Definition 3.5*
- *Let $w(x) := \langle f(x)_{l_0,j_0}, v_1 \rangle - \langle f(x)_{l_0,j_0}, v_2 \rangle \langle f(x)_{l_0,j_0}, v_3 \rangle$*

Then we have

$$|\nabla L(x,:)_{l_0,j_0,i_0} - \nabla L(\widehat{x},:)_{l_0,j_0,i_0}| \leq 3n^3 R^7 \exp(13R^2) \|x - \widehat{x}\|_2$$

Proof.

$$\begin{aligned} & |\nabla L(x,:)_{l_0,j_0,i_0} - \nabla L(\widehat{x},:)_{l_0,j_0,i_0}| \\ = & |g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) b_{l_0,j_0,i_0} \cdot w(x) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) b_{l_0,j_0,i_0} \cdot w(\widehat{x})| \\ \leq & b_{l_0,j_0,i_0} |g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) w(x) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) w(\widehat{x})| \end{aligned}$$

where the first step follows from Lemma F.6 and the definition of $w(x)$, the second step follows from simple algebra.

For convenience, we define

$$C_1 := g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(x) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(x)$$
$$C_2 := g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(x) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(\widehat{x})$$

First, we upper bound $|C_1|$ as follows:

$$\begin{aligned}
|C_1| &= |g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(x) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(x)| \\
&\leq |g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)||w(x)| \\
&\leq |\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle - \langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle||w(x)| \\
&\leq nR^2\exp(5R^2)\|x-\widehat{x}\|_2|\langle f(x)_{l_0,j_0}, v_1\rangle - \langle f(x)_{l_0,j_0}, v_2\rangle\langle f(x)_{l_0,j_0}, v_3\rangle| \\
&\leq nR^2\exp(5R^2)\|x-\widehat{x}\|_2(|\langle f(x)_{l_0,j_0}, v_1\rangle| + |\langle f(x)_{l_0,j_0}, v_2\rangle\langle f(x)_{l_0,j_0}, v_3\rangle|) \\
&\leq nR^2\exp(5R^2)\|x-\widehat{x}\|_2(\|f(x)_{l_0,j_0}\|_2\|v_1|_2 + \|f(x)_{l_0,j_0}\|_2\|v_2\|_2\|f(x)_{l_0,j_0}\|_2\|v_3\|_2) \\
&\leq nR^2\exp(5R^2)\|x-\widehat{x}\|_2 2n^2R^4\exp(6R^2) \\
&= 2n^3R^6\exp(11R^2)\|x-\widehat{x}\|_2
\end{aligned}$$

where the first step follows from the definition of C_1, the second step follows from Fact A.3, the third step follows from Lemma F.4, the fourth step follows from the definition of $w(x)$, the fifth step follows from triangular inequality, the sixth step follows from Fact A.3, the seventh step follows from Lemma F.7, the last step follows from simple algebra.

Then, we upper bound $|C_2|$ as follows:

$$\begin{aligned}
|C_2| &= |g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(x) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(\widehat{x})| \\
&\leq |g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)||w(x) - w(\widehat{x})| \\
&\leq n^2R^4\exp(13R^2)\|x-y\|_2
\end{aligned}$$

where the first step follows from the definition of C_2, the second step follows from simple algebra, the third step follows from **Part 3** of Lemma F.8 and $g(x)(1-g(x)) \leq 1$.

Thus, we obtained the bound:

$$\begin{aligned}
|g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(x) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(\widehat{x})| &= |C_1 + C_2| \\
&\leq |C_1| + |C_2| \\
&\leq 3n^3R^6\exp(13R^2)\|x-\widehat{x}\|_2
\end{aligned}$$

Finally, we have

$$b_{l_0,j_0,i_0}|g'(\langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(x) - g'(\langle f(\widehat{x})_{l_0,j_0}, h(y)_{l_0,i_0}\rangle)w(\widehat{x})| \leq R\cdot 3n^3R^6\exp(13R^2)\|x-\widehat{x}\|_2$$

□

G MAIN RESULTS

Lemma G.1 (Lemma 8 of Tarzanagh et al. (2023a)). *By Lemma F.9, if $\eta \leq 1/L_x$, then for any initialization $x(0)$, Algorithm W-GD (Definition 3.7) satisfies*

$$L(x(k+1)) - L(x(k)) \leq -\frac{\eta}{2}\|\nabla L(x(k))\|_F^2$$

for $\forall k \geq 0$. Additionally, it holds that

$$\sum_{k=0}^{\infty} \|\nabla L(x(k))\|_F^2 < \infty$$
$$\lim_{k\to\infty} \|\nabla L(x(k))\|_F^2 = 0$$

Proof. The proof is similar to Lemma 5 of Tarzanagh et al. (2023b). □

Lemma G.2 (Lemma 9 of Tarzanagh et al. (2023a)). *Let W^{mm} be the SVM solution of ATT-SVM(Tarzanagh et al. (2023a)). Assumption 3.10 hold. Then, for $\forall W \in \mathbb{R}^{d\times d}$, the training loss W-ERM(Tarzanagh et al. (2023a)) obyes $\langle \nabla L(W), W^{mm}\rangle \leq -c < 0$, for some constant $c > 0$ (see Eq. (16)) depending on the data, the head v, and a loss derivative bound.*

Proof. Let

$$\overline{h}_i := U_i X^{mm} z_i$$
$$\gamma_i := V_i \cdot U_i v$$
$$h_i := U_i X z_i$$

which implies that

$$\begin{aligned}
\langle \nabla L(X), X^{mm}\rangle &= \frac{1}{m}\sum_{i=1}^{m} l'(\gamma_i^\top \mathbb{S}(h_i)) \cdot \langle U_i^\top \mathbb{S}'(h_i)\gamma_i z_i^\top, X^{mm}\rangle \\
&= \frac{1}{m}\sum_{i=1}^{m} l_i' \cdot \mathrm{tr}[(X^{mm})^\top U_i^\top \mathbb{S}(h_i)\gamma_i z_i^\top] \\
&= \frac{1}{m}\sum_{i=1}^{m} l_i' \cdot \overline{h}_i^\top \mathbb{S}'(h_i)\gamma_i \\
&= \frac{1}{m}\sum_{i=1}^{m} l_i' \cdot (\overline{h}_i^\top \,\mathrm{diag}(s_i)\gamma_i - \overline{h}_i^\top s_i s_i^\top \gamma_i) \qquad (10)
\end{aligned}$$

Here, let $l_i' := l'(\gamma_i^\top \mathbb{S}(h_i))$, $s_i = \mathbb{S}(h_i)$, the third step follows from $\mathrm{tr}[ba^\top] = a^\top b$

In order to move forward, we will establish the following result, with a focus on the equal score condition (the second assumption in Assumption 3.10): Let $\gamma = \gamma_{t\geq 2}$ be a constant, and let γ_1 and $\overline{h}_1$ represent the largest indices of vectors γ and $\overline{h}$ respectively. For $\forall s$ that satisfies $\sum_{t\in[T]} cs_t = 1$ and $s_t > 0$, we aim to prove that $\overline{h}^\top \,\mathrm{diag}(s)\gamma - \overline{h}^\top ss^\top \gamma > 0$. To demonstrate this, we proceed by writing the following:

$$\begin{aligned}
\overline{h}^\top \,\mathrm{diag}(s)\gamma - \overline{h}^\top ss^\top \gamma &= \sum_{t=1}^{n} \overline{h}_t\gamma_t s_t - \sum_{t=1}^{n}\overline{h}_t s_t \sum_{t=1}^{n}\gamma_t s_t \\
&= (\overline{h}_1(\gamma_1-\gamma)s_1(1-s_1)) - (\gamma_1-\gamma)s_1\sum_{t\geq 2}^{n}\overline{h}_t s_t \\
&= (\gamma_1-\gamma)(1-s_1)s_1[\overline{h}_1 - \frac{\sum_{t\geq 2}^{n}\overline{h}_t s_t}{\sum_{t\geq 2}^{n} s_t}] \\
&\geq (\gamma_1-\gamma)(1-s_1)s_1(\overline{h}_1 - \max_{t\geq 2}\overline{h}_t) \qquad (11)
\end{aligned}$$

To proceed, we define

$$\gamma_{gap}^{i} := \gamma_{i\mathtt{opt}_i} - \max_{t \neq \mathtt{opt}_i} \gamma_{it}$$
$$\overline{h}_{gap}^{i} := \overline{h}_{i\mathtt{opt}_i} - \max_{t \neq \mathtt{opt}_i} \overline{h}_{it}$$

With these, we obtain

$$\overline{h}_i^\top \operatorname{diag}(s_i)\gamma_i - \overline{h}_i^\top s_i s_i^\top \gamma_i \geq \gamma_{gap}^{i} \overline{h}_{gap}^{i} (1 - s_{i\mathtt{opt}_i}) s_{i\mathtt{opt}_i} \tag{12}$$

Note that

$$\overline{h}_{gap}^{i} = \min_{i \neq \mathtt{opt}_i} (x_{i\mathtt{opt}_i} - x_{it})^\top W^{mm} z_i > 1$$
$$\gamma_{gap}^{i} = \min_{i \neq \mathtt{opt}_i} \gamma_{i\mathtt{opt}_i} - \gamma_{it} > 0$$
$$s_{i\mathtt{opt}_i}(1 - s_{i\mathtt{opt}_i}) > 0$$

Hence,

$$c_0 := \min_{i \in [n]} \{ (\min_{i \neq \mathtt{opt}_i} (x_{i\mathtt{opt}_i} - x_{it})^\top W^{mm} z_i) \cdot (\min_{i \neq \mathtt{opt}_i} \gamma_{i\mathtt{opt}_i} - \gamma_{it}) \cdot s_{i\mathtt{opt}_i}(1 - s_{i\mathtt{opt}_i}) \} > 0 \tag{13}$$

It follows from Eq. (12) and Eq. (13) that

$$\min_{i \in [n]} \{ \overline{h}_i^\top \operatorname{diag}(s_i)\gamma_i - \overline{h}_i^\top s_i s_i^\top \gamma_i \} \geq c_0 \geq 0 \tag{14}$$

Since $l_i' < 0$, l' is continuous and the domain is bounded, the maximum is attained and negative, and thus

$$-c_1 = \max_x l'(x), for\ some\ c_1 > 0 \tag{15}$$

Hence, using Eq. (15) and Eq. (14) in Eq. (10), we obtain

$$\langle \nabla L(X), X^{mm} \rangle \leq -c < 0 \tag{16}$$

where

$$c = c_1 \cdot c_0$$

In the scenario that the second assumption in Assumption 3.10 holds (all tokens are support), $\overline{h}_t = x_{it}^\top W^{mm} z_i$ is constant for $\forall t \geq 2$. Hence, following similar steps as in Eq. (11) completes the proof. □

Lemma G.3 (Lemma 10 of Tarzanagh et al. (2023a)). *Let x^{mm} be the SVM solution of the problem Att-SVM (Tarzanagh et al. (2023a)). Suppose $L(\cdot)$ is strictly decreasing and differentiable. For any choice of $\pi > 0$, there exists $R := R_\pi$ such that, for any x with $\|x\|_F \geq R$, we have*

$$\langle \nabla L(x), \frac{x}{\|x\|_F} \rangle \geq (1 + \pi) \langle \nabla L(x), \frac{x^{mm}}{\|x^{mm}\|_F} \rangle$$

Proof. We define

$$\overline{x} := \frac{\|x^{mm}\|_F x}{\|x\|_F}$$
$$M := \sup_{i,t} \|u_{it} z_i^\top\|$$
$$\Theta := \frac{1}{\|x^{mm}\|_F}$$
$$s_i := \mathbb{S}(U_i X z_i)$$
$$h_i := U_i \overline{x} z_i$$
$$\overline{h_i} := U_i x^{mm} z_i$$

without loss of generality, assume

$$\alpha_i = \texttt{opt}_i = 1$$

for $\forall i \in [n]$.

Repeating the proof for Lemma 9 of Tarzanagh et al. (2023a) yields

$$\langle \nabla L(x), x^{mm} \rangle = \frac{1}{m} \sum_{i=1}^{m} L_i'(\gamma_{i1} - \gamma)(1 - s_{i1}) s_{i1} \left[\overline{h}_{i1} - \frac{\sum_{t\geq 2}^{n} \overline{h}_{it} s_{it}}{\sum_{t\geq 2}^{n} s_{it}} \right]$$

$$\langle \nabla L(x), \overline{x} \rangle = \frac{1}{m} \sum_{i=1}^{m} L_i'(\gamma_{i1} - \gamma)(1 - s_{i1}) s_{i1} \left[h_{i1} - \frac{\sum_{t\geq 2}^{n} h_{it} s_{it}}{\sum_{t\geq 2}^{n} s_{it}} \right]$$

Focusing on a single example $i \in [n]$ with $s, h, \overline{h}$ vectors (dropping subscript i), given π, for suffciently large R, we wish to show that

$$\left[h_1 - \frac{\sum_{t\geq 2}^{n} h_t s_t}{\sum_{t\geq 2}^{n} s_t} \right] \leq (1 + \pi) \left[\overline{h}_1 - \frac{\sum_{t\geq 2}^{n} \overline{h}_t s_t}{\sum_{t\geq 2}^{n} s_t} \right] \tag{17}$$

We consider two scenarios.

Scenarios 1: $\|\overline{x} - x^{mm}\|_F \leq \epsilon := \pi/(2M)$. In this scenario, for any token, we find that

$$\begin{aligned} |h_t - \overline{h}_t| &= |s_t^\top (\overline{x} - x^{mm}) z_t| \\ &\leq M \|\overline{x} - x^{mm}\|_F \\ &\leq M\epsilon \end{aligned}$$

Consequently, we obtain

$$\begin{aligned} \overline{h}_1 - \frac{\sum_{t\geq 2}^{n} \overline{h}_t s_t}{\sum_{t\geq 2}^{n} s_t} &\geq h_1 - \frac{\sum_{t\geq 2}^{n} h_t s_t}{\sum_{t\geq 2}^{n} s_t} - 2M\epsilon \\ &= h_1 - \frac{\sum_{t\geq 2}^{n} h_t s_t}{\sum_{t\geq 2}^{n} s_t} - \pi \end{aligned}$$

Also noticing

$$\overline{h}_1 - \frac{\sum_{t\geq 2}^{n} \overline{h}_t s_t}{\sum_{t\geq 2}^{n} s_t} \geq 1$$

This implies Eq.(17).

Scenario 2: $\|\overline{x} - x^{mm}\|_F \geq \epsilon := \pi/(2M)$. In this scenario, for some $\delta = \delta(\epsilon)$ and $\tau \geq 2$, we have

$$h_1 - h_\tau \leq 1 - 2\delta$$

Recall that $s = \mathbb{S}(\overline{R}h)$ where $\overline{R} = \|x\|_F / \|x^{mm}\|_F$. To proceed, split the tokens into two groups: Let $\mathcal{N}$ be the group of tokens obeying $(u_1 - u_t)^\top \overline{x} z \geq 1 - \delta$ for $t \in \mathcal{N}$ and $[n] - \mathcal{N}$ be the rest. Observe that

$$\begin{aligned} \frac{\sum_{t\in\mathcal{N}} s_t}{\sum_{t\geq 2}^{n} s_t} &\leq \frac{\sum_{t\in\mathcal{N}} s_t}{s_\tau} \\ &\leq n \frac{e^{\delta \overline{R}}}{e^{2\delta \overline{R}}} \\ &= n e^{-\delta \overline{R}} \end{aligned}$$

Set $\overline{M} = M/\Theta$ and note that

$$\|h_t\| \leq \|x^{mm}\|_F \cdot \|u_t z^\top\| \leq M$$

Using

$$(u_1 - u_t)^\top \overline{x} z < 1 - \delta$$

over $t \in [n] - \mathcal{N}$ and plugging in the bound above, we obtain

$$\begin{aligned}\frac{\sum_{t\geq 2}^{n}(h_1 - h_t)s_t}{\sum_{t\geq 2}^{n} s_t} &= \frac{\sum_{t\in[n]-\mathcal{N}}(h_1 - h_t)s_t}{\sum_{t\geq 2}^{n} s_t} + \frac{\sum_{t\in\mathcal{N}}(h_1 - h_t)s_t}{\sum_{t\geq 2}^{n} s_t} \\ &\leq (1-\delta) + 2\overline{M}ne^{-\delta\overline{R}}\end{aligned}$$

Using the fact that

$$\overline{h}_1 - \frac{\sum_{t\geq 2}^{n}\overline{h}_t s_t}{\sum_{t\geq 2}^{n} s_t} \geq 1$$

the above implies Eq.(17) with $\pi' = 2\overline{M}ne^{-\delta\overline{R}} - \delta$. To proceed, choose

$$R_\pi = \delta^{-1}\Theta^{-1}\log(\frac{2\overline{M}n}{\pi})$$

to ensure $\pi' < \pi$ □

Theorem G.4 (A variation of Theorem 4 in page 8 in Tarzanagh et al. (2023a), formal version of Theorem 3.14)**.** *Suppose Assumption 3.10 on the tokens's score hold. Then, Algorithm W-GD (Definition 3.7) with the step size $\eta \leq 1/L_X$ and any starting point $X(0)$ satisfies*

$$\lim_{k\to\infty} \frac{X(k)}{\|X(k)\|_F} = \frac{X^{mm}}{\|X^{mm}\|_F}$$

Proof. Given $\forall \epsilon \in (0,1)$, we define

$$\pi := \frac{\epsilon}{1-\epsilon}$$

By Theorem 3 of Tarzanagh et al. (2023a), we have

$$\lim_{k\to\infty} \|X(k)\|_F = \infty$$

Hence, we can choose k_ϵ such that for any $k \geq k_\epsilon$, for some parameter R_ϵ, it holds that

$$\|X(k)\|_F > R_\epsilon \vee \frac{1}{2}$$

Now, for any $k \geq k_\epsilon$, by Lemma 10 of Tarzanagh et al. (2023a), we have

$$\langle -\nabla L(X(k)), \frac{X^{mm}}{\|X^{mm}\|_F}\rangle \geq (1-\epsilon)\langle -\nabla L(X(k)), \frac{X(k)}{\|X(k)\|_F}\rangle$$

Multiplying both sides by the stepsize η and using the gradient descent update, we have

$$\begin{aligned}\langle X(k+1) - X(k), \frac{X^{mm}}{\|X^{mm}\|_F}\rangle &\geq (1-\epsilon)\langle X(k+1) - X(k), \frac{X(k)}{\|X(k)\|_F}\rangle \\ &= \frac{1-\epsilon}{2\|X(k)\|_F}(\|X(k+1)\|_F^2 - \|X(k)\|_F^2 - \|X(k+1) - X(k)\|_F^2) \\ &\geq (1-\epsilon)\left(\frac{\|X(k+1)_F^2\| - \|X(k)\|_F^2}{2\|X(k)\|_F} - \|X(k+1) - X(k)\|_F^2\right) \\ &\geq (1-\epsilon)(\|X(k+1)\|_F^2 - \|X(k)\|_F^2 - \|X(k+1) - X(k)\|_F^2) \\ &\geq (1-\epsilon)(\|X(k+1)\|_F - \|X(k)\|_F - 2\eta(L(X(k)) - L(X(k+1))))\end{aligned}$$

where the first step follows from simple algebra, the second step follows from $\|x(k)\|_F \geq 1/2$, the third step follows from $(a^2 - b^2)/2b - (a - b) \geq 0$ holds for $\forall a, b > 0$, the last step follows from Lemma 8 of Tarzanagh et al. (2023a).

By summing the inequality over $k \geq k_\epsilon$, we have

$$\langle \frac{X^{mm}}{\|X^{mm}\|_F}, \frac{X(k)}{\|X(k)\|_F} \rangle \geq 1 - \epsilon + \frac{C(\epsilon, \eta)}{\|X(k)\|_F}$$

where the finite constant $C(\epsilon, \eta)$ is defined as

$$C(\epsilon, \eta) := \langle X(k_\epsilon), \frac{X^{mm}}{\|X^{mm}\|_F} \rangle - (1 - \epsilon)\|X(k_\epsilon)\|_F - 2\eta(1 - \epsilon)(L(X(k_\epsilon)) - L_*)$$

where $L_* \leq L(x(k_\epsilon))$ for $\forall k \geq 0$.

Since $\|x(k)\|_F \to \infty$, we have

$$\lim_{k \to \infty} \inf \langle \frac{X^{mm}}{\|X^{mm}\|_F}, \frac{X(k)}{\|X(k)\|_F} \rangle \geq 1 - \epsilon$$

Since ϵ is arbitary, we can consider the limit as $\epsilon \to 0$. Thus, we have

$$\frac{X(k)}{\|X(k)\|_F} \to \frac{X^{mm}}{\|X^{mm}\|_F}$$

□

Theorem G.5 (A variation of Theorem 5 in page 8 in Tarzanagh et al. (2023a), formal version of Theorem 3.15)**.** *For any initialization $X(0)$, there exists a dataset dependent sufficiently small $\delta > 0$ such that the following holds: Suppose non-optimal scores obey $|\gamma_{it} - \gamma_{i\tau}| \leq \delta$ for all $t, \tau \neq opt_i, i \in [m]$. Then, Algorithm x-GD, with $\eta \leq 1/(2L_x)$ obeys $\lim_{k\to\infty} \|X(k)\|_F = \infty$ and $\lim_{k\to\infty} \frac{X_{(k)}}{\|X(k)\|_F} = \frac{X^{mm}}{\|X^{mm}\|_F}$*

Proof. We provide the proof in three steps.

Step 1: Defining the original and equally-scored problems. Given the original dataset (U_i, z_i, V_i) with scores $\gamma_i t$, define an approximate dataset $(\widetilde{U}_i, \widetilde{z}_i, \widetilde{V}_i)$ as follows. Let $P_v^\perp$ denote the projection onto the subspace orthogonal to the linear head v. For a given input i, we define an index $s, \texttt{opt}_i$ as follows:

- If the setting is cross-attention, then, $s, \texttt{opt}_i$ is arbitrary.
- If the setting is self-attention, then $s = 1$ whenever $\texttt{opt}_i \neq 1$ and $s \neq \texttt{opt}_i$ is arbitrary otherwise.

Note this construction does not touch u_{is} and guarantees for equal scores $\gamma_{it} = \gamma_{is}$ for all $t, \texttt{opt}_i$. Observe that by construction $\|\widetilde{u}_{it} - u_{it}\| \leq \delta/\|v\|$ since non-optimal score differences are at most δ. Additionally, we always set $\widetilde{z}_i = z_i$. This is clear for Cross-Attention. For Self-Attention, we use the fact that xi1 is unchanged thanks to our choice of s and we set $\widetilde{z}_i = u_{it1} = z_i$. Following this setting, we define $L(W)$ and$\widetilde{L}(\widetilde{x})$ as the ERM objectives of the original and equally-scored problems, respectively, as follows:

$$L(X) = \frac{1}{n} \sum_{i=1}^{n} L(V_i v^\top U_i^\top \mathbb{S}(U_i X z_i)) \tag{18}$$

$$\widetilde{L}(\widetilde{X}) = \frac{1}{m} \sum_{i=1}^{m} L(V_i v^\top \widetilde{U}_i^\top \mathbb{S}(\widetilde{U}_i \widetilde{X} z_i)) \tag{19}$$

Let X^{mm} and $\widetilde{X}^{mm}$ denote the solution of ATT-SVM(Tarzanagh et al. (2023a)) for the original and equally-scored SVM problems, respectively:

$$X^{mm} = \arg\min_{x} \|x\|_F \qquad \text{subj.to} \quad (u_{i\texttt{opt}_i} - u_{it})^\top x z_i \geq 1 \quad \text{for} \quad \forall t \neq \texttt{opt}_i, i \in [n] \tag{20}$$

$$\widetilde{X}^{mm} = \arg\min_{\widetilde{X}} \|\widetilde{X}\|_F \qquad \text{subj.to} \quad (u_{i\texttt{opt}_i} - \widetilde{u}_{it})^\top \widetilde{X} z_i \geq 1 \quad \text{for} \quad \forall t \neq \texttt{opt}_i, i \in [n] \tag{21}$$

Recall that we assume a solution to the original problem Wmm exists. Additionally, $\widetilde{x}^{mm}$ is guaranteed to exist by making δ smaller than a dataset-dependent constant. Also note that there exists $\Delta_0(\delta) > 0$ that depends solely on the original problem and δ, and can be made arbitrarily small by decreasing δ, such that

$$\frac{\|\widehat{x}^{mm} - x^{mm}\|_F}{\|x^{mm}\|_F} \leq \Delta_0(\delta) \tag{22}$$

To proceed, let $\gamma_i = V_i \cdot U_i v, \widetilde{\gamma}_i = V_i \cdot \widetilde{U}_i v$ and $\widetilde{h}_i = \widehat{U}_i \widehat{x}_i z_i$. Following Lemma F.9, we have

$$\begin{aligned}
\|\nabla L(X) - \nabla \widetilde{L}(X)\|_F &\leq \frac{1}{m}\sum_{i=1}^{m} \|l'(\gamma_i^\top \mathbb{S}(h_i)) \cdot z_i \gamma_i^\top \mathbb{S}'(h_i) U_i - l'(\widetilde{\gamma}_i^\top \mathbb{S}(\widetilde{h}_i)) \cdot z_i \widetilde{\gamma}_i^\top \mathbb{S}'(\widetilde{h}_i) \widetilde{U}_i\|_F \\
&\leq \frac{1}{m}\sum_{i=1}^{m} M_0 \|z_i \widetilde{\gamma}_i^\top \mathbb{S}'(\widetilde{h}_i)\widetilde{U}_i\|_F \|\gamma_i^\top \mathbb{S}(h_i) - \widetilde{\gamma}_i^\top \mathbb{S}(\widetilde{h}_i)\| \\
&\quad + \frac{1}{m}\sum_{i=1}^{m} M_1 \|z_i \gamma_i^\top \mathbb{S}'(h_i) U_i - z_i \widetilde{\gamma}_i^\top \mathbb{S}'(\widetilde{h}_i)\widetilde{U}_i\|_F
\end{aligned} \tag{23}$$

where by Lemma F.4 we have $M_1 = 3n^3 R^7 \exp(13R^2)$ and $M_0 = 2n^2 R^2 \exp(6R^2)$

Step 2: Monitoring the fluctuations during iterations until $x(k)$ enters the local cone around x^{mm} Fix $X(0) = \widetilde{X}(0)$. Algorithm W-GD(Definition 3.7) applied to $L(X)$ and $\widetilde{L}(\widetilde{X})$ defined in Eq. 20 and Eq. (19) implies that

$$\widetilde{X}(k+1) = \widetilde{X}(k) - \eta \nabla \widetilde{L}(\widetilde{X}(k)) \tag{24}$$

$$X(k+1) = X(k) - \eta \nabla L(X(k)) \tag{25}$$

For the original problem with Objective Eq.(20), it follows from Theorem 7 of Tarzanagh et al. (2023a) that there exist parameters $\mu = \mu(\texttt{opt}) \in (0,1)$ and $R = R\mu > 0$ and a conic set $C_{\mu,R}(X^{mm})$ such that gradient descent converges to the max-margin direction X^{mm} when initialized anywhere within $C_{\mu,R}(X^{mm})$, where

$$C_{\mu,R}(X^{mm}) = \{X \in \mathbb{R}^{d\times d} : \langle \frac{X}{\|X\|_F}, \frac{X^{mm}}{\|X^{mm}\|_F} \rangle \geq 1 - \mu, \|X\|_F \geq R\}$$

We will prove the following claim.

- **Claim 1** For a sufficiently large data-dependent δ (see (22),(26),(28),(34), there exists $k \geq 1$ such that $X(k) \in C_{\mu,R}(X^{mm})$, where $X(k)$ is defined in Eq. (25))

Let L_X and $L_{\widetilde{X}}$ denote the lipschitz constants of gradients of objectives $L(X)$ and $\widetilde{L}(\widetilde{X})$ defined in Eq. (18) and Eq. (19) respectively. From Lemma F.9, we obtain

$$2\|v\|\|z_i\|^2\|U_i\|^3(M_0\|v\|\|X_i\| + 3M_1) - \|v\|\|z_i\|^2\|\widetilde{X}_i\|^3(M_0\|v\|\|\widetilde{X}_i\| + 3M_1) = 2L_X - L_{\widetilde{X}}$$

Hence, there exists $\Delta_1(\delta) > 0$ that depends solely on the original problem and δ, and can be made arbitarily small by decreasing δ, such that

$$2L_X - L_{\widetilde{X}} \geq \Delta_1(\delta) L_{\widetilde{X}} \tag{26}$$

which implies that

$$\frac{1}{2L_X} \leq \frac{1}{(1+\Delta_1(\delta))L_{\widetilde{X}}} \leq \frac{1}{L_{\widetilde{X}}}$$

Hence, any stepsize $\eta \leq 1/(2L_X)$ satisfies the requirements of Lemma 8 for the original and auxiliary objectives $L(X)$ and $\widetilde{L}(\widetilde{X})$, respectively. As a result, the gradient descent updates (24) and (25) with any stepsize $\eta \leq 1/(2L_X)$ guarantees the descent of Objectives (18) and (19), respectively. To proceed, let $\widetilde{\mu} = \mu/2$. For the equally-scored problem with Objective (19), Theorem 4 in Tarzanagh et al. (2023b) assures the existence of $k = k_{\widetilde{\mu}}$ such that when we run gradient descent in (24) with the step size η for k iterations,then $\|\widetilde{X}\|_F \geq R_{\widetilde{\mu}}$ for some $R_{\widetilde{\mu}} \geq 2R_\mu$,and

$$\langle \frac{\widetilde{X}(k)}{\|\widetilde{X}(k)\|_F}, \frac{\widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} \rangle \geq 1 - \widetilde{\mu} \tag{27}$$

Using Eq. (23), let $\Delta_{2,\tau} := \|\nabla L(X(\tau)) - \nabla \widetilde{L}(\widetilde{X}(\tau))\|_F$. Since $X(0) = \widetilde{X}(0)$, it follows from Eq. (23) that there exists $\Delta_2(\delta)$ that depends on the original problem (due to Eq. (23) and $\mu = 2\widetilde{\mu}$) and δ, and $\Delta_{2,\tau}(\delta)$ and $\Delta_2(\delta)$ can be made arbitrarily small by decreasing δ such that

$$\frac{\|\widetilde{X}(k) - X(k)\|_F}{\|\widetilde{X}(k)\|_F} \leq \frac{\eta}{R_{\widetilde{\mu}}} \sum_{\tau=0}^{k_{\widetilde{\mu}}-1} \|\nabla L(X(\tau)) - \nabla \widetilde{L}(\widetilde{X}(\tau))\|_F \leq \frac{\eta}{R_{\widetilde{\mu}}} \sum_{\tau=0}^{k_{\widetilde{\mu}}-1} \Delta_{2,\tau}(\delta) \leq \Delta_2(\delta) \tag{28}$$

Let

$$\Delta(\delta) := \Delta_2(\delta) + \Delta_0(\delta) + \Delta_2(\delta)\Delta_0(\delta) \tag{29}$$

where $\Delta_0(\delta)$ is given in Eq. (22) and $\Delta_2(\delta)$ is given in Eq. (28)

It follows from Eq. (22), Eq. (28), Eq. (27) and Eq. (29) that

$$\begin{aligned}
\langle \frac{X(k)}{\|X(k)\|_F}, \frac{\widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} \rangle &= \langle \frac{\widetilde{X}(k)}{\|\widetilde{X}(k)\|_F} + \frac{X(k) - \widetilde{X}(k)}{\|\widetilde{X}(k)\|_F}, \frac{\widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} + \frac{X^{mm} - \widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} \rangle \\
&= \langle \frac{\widetilde{X}(k)}{\|\widetilde{X}(k)\|_F}, \frac{\widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} \rangle + \langle \frac{X(k) - \widetilde{X}(k)}{\|\widetilde{X}(k)\|_F}, \frac{\widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} \rangle \\
&\quad + \langle \frac{\widetilde{X}(k)}{\|\widetilde{X}(k)\|_F}, \frac{X^{mm} - \widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} \rangle + \langle \frac{X(k) - \widetilde{X}(k)}{\|\widetilde{X}(k)\|_F}, \frac{X^{mm} - \widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} \rangle \\
&\geq 1 - \widetilde{\mu} + (-\Delta_2(\delta) - \Delta_0(\delta) - \Delta_2(\delta)\Delta_0(\delta)) \\
&\geq 1 - 2\widetilde{\mu} + \widetilde{\mu} - \Delta(\delta) \\
&\geq 1 - \mu + \widetilde{\mu} - \Delta(\delta)
\end{aligned} \tag{30}$$

$$\frac{\|X(k)\|_F}{\|\widetilde{X}(k)\|} \leq 1 + \frac{X(k) - \widetilde{X}(k)}{\|\widetilde{X}(k)\|_F} \leq 1 + \Delta_2(\delta) \tag{31}$$

$$\frac{\|X^{mm}\|_F}{\|\widetilde{X}^{mm}\|} \leq 1 + \frac{X^{mm} - \widetilde{X}^{mm}}{\|\widetilde{X}^{mm}\|_F} \leq 1 + \Delta_0(\delta) \tag{32}$$

Now, it follows from Eq. (30) - Eq. (32) for $k = k_{\widetilde{\mu}}$,

$$\begin{aligned}
\langle \frac{X(k)}{\|X(k)\|_F}, \frac{X^{mm}}{\|X^{mm}\|_F} \rangle &\geq \frac{1}{1 + \Delta_2(\delta)} \cdot \frac{1}{1 + \Delta_0(\delta)} (1 - \mu + \widetilde{\mu} - \Delta(\delta)) \\
&= \frac{1}{1 + \Delta(\delta)} (1 - \mu + \widetilde{\mu} - \Delta(\delta)) \\
&\geq 1 - \mu + \frac{1}{1 + \Delta(\delta)} (\widetilde{\mu} - (2 - \mu)\Delta(\delta)) \\
&\geq 1 - \mu
\end{aligned} \tag{33}$$

Here, the last inequality is obtained by choosing $\delta > 0$ to ensure that (26) holds, and both $\Delta_2(\delta)$ and $\Delta_0(\delta)$ are sufficiently small such that (22), (28), and the following condition are satisfied:

$$\widetilde{\mu} - (2 - \mu)\Delta(\delta) \geq 0 \tag{34}$$

We can similarly guarantee $\|X(k)\|_F \geq R_\mu$ by using $R_{\widetilde{\mu}} \geq 2R_\mu$. Hence, we have shown that, for sufficiently small data-dependent δ (see Conditions (22),(26),(28),(34)), Claim 1 holds, and for $k = k_{\widetilde{\mu}}$, $X(k) \in C_{\mu,R}(X^{mm})$, where $X(k)$ is defined in (24), (25).

Step 3: The proof now follows by applying Theorem 7 on the original problem. This is because gradient descent iterations starting at $X(k)$ for $k = k_{\widetilde{\mu}}$ which lies within the cone provably converges to the max-margin direction. □

H HESSIAN

In this section, we provide a brief analysis on the hessian of our loss function. In Section H.1, we compute the hessian with respect to x. In Section H.2, we reform the hessian for the ease of the analysis afterwards. In Section H.3, we are able to show that the hessian can be decomposed into several diagnal matrices and low rank matrices.

H.1 HESSIAN COMPUTATION WITH RESPECT TO x

Lemma H.1. *If the following conditions hold*

- *Let* $\gamma(x)_{l_0,j_0,i_0} := \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle$
- *Let* $\frac{\mathrm{d}L_{l_0,i_0,j_0}(x,y)}{\mathrm{d}x_i}$ *be computed as Lemma B.6*

Then we have

- *Part 1.*

$$\begin{aligned}
&\frac{\mathrm{d}L_{l_0,i_0,j_0}(x,y)}{\mathrm{d}x_i \mathrm{d}x_i} \\
=& (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \gamma(x)_{l_0,j_0,i_0} \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle)^2 \\
&+ c(x,y)_{l_0,j_0,i_0} \\
&(\\
&+ \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle (1 - \gamma(x)_{l_0,j_0,i_0}) \\
&- 2 \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \\
&+ 2 \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle^2 \gamma(x)_{l_0,j_0,i_0} \\
&)
\end{aligned}$$

- *Part 2.*

$$\begin{aligned}
&\frac{\mathrm{d}L_{l_0,i_0,j_0}(x,y)}{\mathrm{d}x_i \mathrm{d}x_j} \\
=& (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \gamma(x)_{l_0,j_0,i_0} \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \cdot \\
&(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle - \gamma(x)_{l_0,j_0,i_0} \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle) \\
&+ c(x,y)_{l_0,j_0,i_0} \\
&(\\
&+ \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle (1 - \gamma(x)_{l_0,j_0,i_0}) \\
&- \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle - \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle \\
&+ 2 \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle \gamma(x)_{l_0,j_0,i_0} \\
&)
\end{aligned}$$

Proof. **Proof of Part 1** First, we compute

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x_i} (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \\
=& \frac{\mathrm{d}}{\mathrm{d}x_i} \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \\
&- (\frac{\mathrm{d}}{\mathrm{d}x_i} \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \\
&- (\frac{\mathrm{d}}{\mathrm{d}x_i} \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \\
=& \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \\
&- (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle
\end{aligned}$$

$$
\begin{aligned}
&- (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, \mathsf{A}_{l_0,j_0,i} \rangle - \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \\
= &\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - 2 \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0} \mathsf{A}_{l_0,j_0,i} \rangle \\
&+ 2 \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle^2 \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle
\end{aligned}
$$

where the first step follows from simple derivative, the second step follows from simple algebra, the third step follows from simple algebra.

Then, we have

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x_i} \frac{\mathrm{d}}{\mathrm{d}x_i} L_{l_0,i_0,j_0}(x,y) \\
= &\frac{\mathrm{d}}{\mathrm{d}x_i} (c(x,y)_{l_0,j_0,i_0} \mathsf{A}_{l_0,j_0,i}^\top (f(x)_{l_0,j_0} - f(x)_{l_0,j_0} f(x)_{l_0,j_0}^\top) h(y)_{l_0,i_0}) \\
= &(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \gamma(x)_{l_0,j_0,i_0} \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle)^2 \\
&+ c(x,y)_{l_0,j_0,i_0} \frac{\mathrm{d}}{\mathrm{d}x_i} (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle)
\end{aligned}
$$

where the first step follows from differential chain rule, the second step follows from **Part 8** of Lemma B.6 and differential chain rule.

By combining the two equations, we completes the proof.

Proof of Part 2 First, we compute

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x_j} (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \\
= &\frac{\mathrm{d}}{\mathrm{d}x_j} \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \\
&- (\frac{\mathrm{d}}{\mathrm{d}x_j} \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle) \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \\
&- (\frac{\mathrm{d}}{\mathrm{d}x_j} \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \\
= &\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \\
&- (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle) \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \\
&- (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, \mathsf{A}_{l_0,j_0,j} \rangle - \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle) \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \\
= &\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \\
&- \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0} \mathsf{A}_{l_0,j_0,j} \rangle - \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0} \mathsf{A}_{l_0,j_0,i} \rangle \\
&+ 2 \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \\
&- \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle
\end{aligned}
$$

where the first step follows from simple derivative, the second step follows from simple algebra, the third step follows from simple algebra.

Then, we have

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x_j} \frac{\mathrm{d}}{\mathrm{d}x_i} L_{l_0,i_0,j_0}(x,y) \\
= &\frac{\mathrm{d}}{\mathrm{d}x_j} (c(x,y)_{l_0,j_0,i_0} \mathsf{A}_{l_0,j_0,i}^\top (f(x)_{l_0,j_0} - f(x)_{l_0,j_0} f(x)_{l_0,j_0}^\top) h(y)_{l_0,i_0}) \\
= &(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \gamma(x)_{l_0,j_0,i_0} \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle) \cdot \\
&(\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle - \gamma(x)_{l_0,j_0,i_0} \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle) \\
&+ c(x,y)_{l_0,j_0,i_0} \frac{\mathrm{d}}{\mathrm{d}x_j} (\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle - \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle)
\end{aligned}
$$

where the first step follows from differential chain rule, the second step follows from **Part 8** of Lemma B.6 and differential chain rule.

By combining the two equations, we completes the proof. □

H.2 Reformulating Several Terms

Lemma H.2. *We have*

- *Part 1.*
$$\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle = \mathsf{A}_{l_0,j_0,i}^\top \operatorname{diag}(f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}) \mathsf{A}_{l_0,j_0,j}$$
- *Part 2.*
$$\begin{aligned} &\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle + \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle \\ =\ & \mathsf{A}_{l_0,j_0,i}^\top ((f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}) f(x)_{l_0,j_0}^\top + f(x)_{l_0,j_0} (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0})^\top) \mathsf{A}_{l_0,j_0,j} \end{aligned}$$
- *Part 3.*
$$\begin{aligned} &\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \\ =\ & \mathsf{A}_{l_0,j_0,i}^\top (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}) (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0})^\top \mathsf{A}_{l_0,j_0,j} \end{aligned}$$
- *Part 4.*
$$\langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle = \mathsf{A}_{l_0,j_0,i}^\top f(x)_{l_0,j_0} f(x)_{l_0,j_0}^\top \mathsf{A}_{l_0,j_0,j}$$

Proof. **Proof of Part 1.** This trivially follows from Fact A.1

Proof of Part 2.

$$\begin{aligned} &\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,i} \rangle + \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0}, \mathsf{A}_{l_0,j_0,j} \rangle \\ =\ & \langle f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}, \mathsf{A}_{l_0,j_0,i} \rangle f(x)_{l_0,j_0}^\top \mathsf{A}_{l_0,j_0,j} + \langle f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}, \mathsf{A}_{l_0,j_0,j} \rangle \mathsf{A}_{l_0,j_0,i}^\top f(x)_{l_0,j_0} \\ =\ & \mathsf{A}_{l_0,j_0,i}^\top ((f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}) f(x)_{l_0,j_0}^\top + f(x)_{l_0,j_0} (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0})^\top) \mathsf{A}_{l_0,j_0,j} \end{aligned}$$

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

Proof of Part 3

$$\begin{aligned} &\langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,i}, h(y)_{l_0,i_0} \rangle \langle f(x)_{l_0,j_0} \circ \mathsf{A}_{l_0,j_0,j}, h(y)_{l_0,i_0} \rangle \\ =\ & \langle f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}, \mathsf{A}_{l_0,j_0,i} \rangle \langle f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}, \mathsf{A}_{l_0,j_0,j} \rangle \\ =\ & \mathsf{A}_{l_0,j_0,i}^\top (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}) (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0})^\top \mathsf{A}_{l_0,j_0,j} \end{aligned}$$

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

Proof of Part 4 This trivially follows from Fact A.1. □

H.3 Decomposing $\nabla^2 L_{l_0,i_0,j_0}(x,y)$

Definition H.3. *Let $\gamma(x)_{l_0,j_0,i_0} := \langle f(x)_{l_0,j_0}, h(y)_{l_0,i_0} \rangle$ for convenience, then we define $B(x)$ as follows:*

$$B(x) := B_{\mathrm{diag}} + B_{\mathrm{rank}}^1 + B_{\mathrm{rank}}^2 + B_{\mathrm{rank}}^3$$

where

- $B_{\mathrm{diag}} = (1 - \gamma(x)_{l_0,j_0,i_0}) c(x,y)_{l_0,j_0,i_0} \operatorname{diag}(f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0})$
- $B_{\mathrm{rank}}^1 = -(2\gamma(x)_{l_0,j_0,i_0} + c(x,y)_{l_0,j_0,i_0})((f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}) f(x)_{l_0,j_0}^\top + f(x)_{l_0,j_0} (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0})^\top)$
- $B_{\mathrm{rank}}^2 = (2\gamma(x)_{l_0,j_0,i_0} c(x,y)_{l_0,j_0,i_0} + \gamma(x)_{l_0,j_0,i_0}^2) f(x)_{l_0,j_0} f(x)_{l_0,j_0}^\top$
- $B_{\mathrm{rank}}^3 = (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0}) (f(x)_{l_0,j_0} \circ h(y)_{l_0,i_0})^\top$