Supplementary Materials

A Proofs for the single layer case (Theorem 1)

In this section, we prove our characterization of global minima for the single layer case (Theorem 1). We begin by simplifying the loss into a more concrete form.

A.1 Rewriting the loss function

Recall the in-context loss $f(P, Q)$ defined in (6):

$$f(P, Q) = \mathbf{E}_{Z_0, w_\star}\left[\left[Z_0 + \frac{1}{n}\mathrm{Attn}_{P,Q}(Z_0)\right]_{(d+1),(n+1)} + w_\star^\top x^{(n+1)}\right]^2$$

Using the notation $Z_0 = [z^{(1)}\ z^{(2)}\ \cdots\ z^{(n+1)}]$, one can rewrite Z_1 as follows:

$$\begin{aligned} Z_1 &= Z_0 + \frac{1}{n}\mathrm{Attn}_{P,Q}(Z_0) \\ &= [z^{(1)}\ \cdots\ z^{(n+1)}] + \frac{1}{n}P[z^{(1)}\ \cdots\ z^{(n+1)}]M\left([z^{(1)}\ \cdots\ z^{(n+1)}]^\top Q[z^{(1)}\ \cdots\ z^{(n+1)}]\right). \end{aligned}$$

Thus, the last token of Z_1 can be expressed as

$$z^{(n+1)} + \frac{1}{n}\sum_{i=1}^{n} Pz^{(i)}({z^{(i)}}^\top Q z^{(n+1)}) = \begin{bmatrix} x^{(n+1)} \\ 0 \end{bmatrix} + \frac{1}{n}P\sum_{i=1}^{n} z^{(i)}{z^{(i)}}^\top Q\begin{bmatrix} x^{(n+1)} \\ 0 \end{bmatrix},$$

where note that the summation is for $i = 1, 2, \ldots, n$ due to the mask matrix M. Letting $b^\top$ be the last row of P, and $A \in \mathbb{R}^{d+1,d}$ be the first d columns of Q, then $f(P, Q)$ only depends on b, A and henceforth, we will write $f(P, Q)$ as $f(b, A)$. Then, $f(b, A)$ can be rewritten as

$$\begin{aligned} f(b, A) &= \mathbf{E}_{Z_0, w_\star}\left[b^\top \underbrace{\frac{1}{n}\sum_i z^{(i)}{z^{(i)}}^\top} Ax^{(n+1)} + w_\star^\top x^{(n+1)}\right]^2 \\ &=: \mathbf{E}_{Z_0, w_\star}\left[b^\top \mathcal{M} A x^{(n+1)} + w_\star^\top x^{(n+1)}\right]^2 = \mathbf{E}_{Z_0, w_\star}\left[(b^\top \mathcal{M} A + w_\star^\top)x^{(n+1)}\right]^2, \end{aligned} \quad (13)$$

where we used the notation $\mathcal{M} := \frac{1}{n}\sum_i z^{(i)}{z^{(i)}}^\top$ to simplify. We now analyze the global minima of this loss function.

To illustrate the proof idea clearly, we begin with the proof for the simpler case of isotropic data.

A.2 Warm-up: proof for the isotropic data

As a warm-up, we first prove the result for the special case where $x^{(i)}$ is sampled from $\mathcal{N}(0, I_d)$ and $w_\star$ is sampled from $\mathcal{N}(0, I_d)$.

Step 1: Decomposing the loss function into components

Writing $A = [a_1 \; a_1 \; \cdots \; a_d]$, and use the fact that $\mathbf{E}[x^{(n+1)}[i]x^{(n+1)}[j]] = 0$ for $i \neq j$, we get

$$f\left(b, A\right) = \sum_{j=1}^{d} \mathbf{E}_{Z_0, w_\star}\left[b^\top \mathcal{M} a_j + w_\star[j]\right]^2 \mathbf{E}[x^{(n+1)}[j]^2] = \sum_{j=1}^{d} \mathbf{E}_{Z_0, w_\star}\left[b^\top \mathcal{M} a_j + w_\star[j]\right]^2 .$$

Hence, we first focus on characterizing the global minima of each component in the summation separately. To that end, let us formally define each component in the summation as follows.

$$\begin{aligned} f_j(b, A) &\coloneqq \mathbf{E}_{Z_0, w_\star}\left[b^\top \mathcal{M} a_j + w_\star[j]\right]^2 = \mathbf{E}_{Z_0, w_\star}\left[\mathrm{Tr}(\mathcal{M} a_j b^\top) + w_\star[j]\right]^2 \\ &= \mathbf{E}_{Z_0, w_\star}\left[\left\langle \mathcal{M}, b a_j^\top \right\rangle + w_\star[j]\right]^2 , \end{aligned}$$

where we use the notation $\langle X, Y \rangle \coloneqq \mathrm{Tr}(XY^\top)$ for two matrices X and Y here and below.

Step 2: Characterizing global minima of each component

To characterize the global minima of each objective, we prove the following result.

Lemma 6. *Suppose that $x^{(i)}$ is sampled from $\mathcal{N}(0, I_d)$ and $w_\star$ is sampled from $\mathcal{N}(0, I_d)$. Consider the following objective ($\langle X, Y \rangle \coloneqq \mathrm{Tr}(XY^\top)$ for two matrices X and Y)*

$$f_j(X) = \mathbf{E}_{Z_0, w_\star}\left[\langle \mathcal{M}, X \rangle + w_\star[j]\right]^2 .$$

Then a global minimum is given as

$$X_j = -\frac{1}{\left(\frac{n-1}{n} + (d+2)\frac{1}{n}\right)} E_{d+1, j} ,$$

where E_{i_1, i_2} is the matrix whose (i_1, i_2)-th entry is 1, and the other entries are zero.

Proof of Lemma 6. Note first that f_j is convex in X. Hence, in order to show that a matrix X_0 is the global optimum of f_j, it suffices to show that the gradient vanishes at that point, in other words,

$$\nabla f_j(X_0) = 0 .$$

To verify this, let us compute the gradient of f_j:

$$\nabla f_j(X_0) = 2\mathbf{E}\left[\langle \mathcal{M}, X_0 \rangle \mathcal{M}\right] + 2\mathbf{E}\left[w_\star[j] \mathcal{M}\right] ,$$

where we recall that $\mathcal{M}$ is defined as

$$\mathcal{M} = \frac{1}{n} \sum_i \begin{bmatrix} x^{(i)} x^{(i)\top} & y^{(i)} x^{(i)} \\ y^{(i)} x^{(i)\top} & {y^{(i)}}^2 \end{bmatrix} .$$

To verify that the gradient is equal to zero, let us first compute $\mathbf{E}\left[w_\star[j]\mathcal{M}\right]$. For each $i = 1, \ldots, n$, note that $\mathbf{E}[w_\star[j] x^{(i)} x^{(i)\top}] = O$ because $\mathbf{E}[w_\star] = 0$. Moreover, $\mathbf{E}[w_\star[j] {y^{(i)}}^2] = 0$ because $w_\star$ is symmetric, i.e., $w_\star \overset{d}{=} -w_\star$, and $y^{(i)} = \left\langle w_\star, x^{(i)} \right\rangle$. Lastly, for $k = 1, 2, \ldots, d$, we have

$$\mathbf{E}[w_\star[j] y^{(i)} x^{(i)}[k]] = \mathbf{E}[w_\star[j] \left\langle w_\star, x^{(i)} \right\rangle x^{(i)}[k]] = \mathbf{E}\left[w_\star[j]^2 x^{(i)}[j] x^{(i)}[k]\right] = \mathbb{1}_{[j=k]} \tag{14}$$

because $\mathbf{E}[w_\star[i] w_\star[j]] = 0$ for $i \neq j$. Combining the above calculations, it follows that

$$\mathbf{E}\left[w_\star[j]\mathcal{M}\right] = E_{d+1, j} + E_{j, d+1} . \tag{15}$$

We now compute compute $\mathbf{E}\left[\langle \mathcal{M}, E_{d+1,j} \rangle \mathcal{M}\right]$. Note first that

$$\langle \mathcal{M}, E_{d+1,j} \rangle = \sum_i \left\langle w_\star, x^{(i)} \right\rangle x^{(i)}[j] .$$

Hence, it holds that

$$\mathbf{E}\left[\langle \mathcal{M}, E_{d+1,j}\rangle \left(\sum_i x^{(i)} {x^{(i)}}^\top\right)\right] = \mathbf{E}\left[\left(\sum_i \left\langle w_\star, x^{(i)}\right\rangle x^{(i)}[j]\right)\left(\sum_i x^{(i)} {x^{(i)}}^\top\right)\right] = O\,.$$

because $\mathbf{E}[w_\star] = 0$. Next, we have

$$\mathbf{E}\left[\langle \mathcal{M}, E_{d+1,j}\rangle \left(\sum_i {y^{(i)}}^2\right)\right] = \mathbf{E}\left[\left(\sum_i \left\langle w_\star, x^{(i)}\right\rangle x^{(i)}[j]\right)\left(\sum_i {y^{(i)}}^2\right)\right] = 0$$

because $w_\star \overset{d}{=} -w_\star$. Lastly, we compute

$$\mathbf{E}\left[\langle \mathcal{M}, E_{d+1,j}\rangle \left(\sum_i y^{(i)} {x^{(i)}}^\top\right)\right].$$

To that end, note that for $j \neq j'$,

$$\mathbf{E}\left[\left\langle w_\star, x^{(i)}\right\rangle x^{(i)}[j] \left\langle w_\star, x^{(i')}\right\rangle x^{(i')}[j']\right] = \begin{cases} \mathbf{E}[\left\langle x^{(i)}, x^{(i')}\right\rangle x^{(i)}[j] x^{(i')}[j']] = 0 & \text{if } i \neq i', \\ \mathbf{E}[\left\|x^{(i)}\right\|^2 x^{(i)}[j] x^{(i)}[j']] = 0 & \text{if } i = i', \end{cases}$$

and

$$\mathbf{E}\left[\left\langle w_\star, x^{(i)}\right\rangle x^{(i)}[j] \left\langle w_\star, x^{(i')}\right\rangle x^{(i')}[j]\right] = \begin{cases} \mathbf{E}[x^{(i)}[j]^2 x^{(i')}[j]^2] = 1 & \text{if } i \neq i', \\ \mathbf{E}\left[\left\langle w_\star, x^{(i)}\right\rangle^2 x^{(i)}[j]^2\right] = d+2 & \text{if } i = i', \end{cases} \tag{16}$$

where the last case follows from the fact that the 4th moment of Gaussian is 3 and

$$\mathbf{E}\left[\left\langle w_\star, x^{(i)}\right\rangle^2 x^{(i)}[j]^2\right] = \mathbf{E}\left[\left\|x^{(i)}\right\|^2 x^{(i)}[j]^2\right] = 3 + d - 1 = d + 2.$$

Combining the above calculations together, we arrive at

$$\begin{aligned} \mathbf{E}\left[\langle \mathcal{M}, E_{d+1,j}\rangle \mathcal{M}\right] &= \frac{1}{n^2} \cdot (n(n-1) + (d+2)n)\left(E_{d+1,j} + E_{j,d+1}\right) \\ &= \left(\frac{n-1}{n} + (d+2)\frac{1}{n}\right)\left(E_{d+1,j} + E_{j,d+1}\right). \end{aligned} \tag{17}$$

Therefore, combining (15) and (17), the results follows. □

Step 3: Combining global minima of each component

Now we finish the proof. From Lemma 6, it follows that

$$X_j = -\frac{1}{\left(\frac{n-1}{n} + (d+2)\frac{1}{n}\right)} E_{d+1,j}\,,$$

is the unique global minimum of f_j. Hence, b and $A = [a_1\ a_1\ \cdots\ a_d]$ achieve the global minimum of $f(b, A) = \sum_{j=1}^d f_j(b, A_j)$ if they satisfy

$$b a_j^\top = -\frac{1}{\left(\frac{n-1}{n} + (d+2)\frac{1}{n}\right)} E_{d+1,j} \quad \text{for all } i = 1, 2, \dots, d.$$

This can be achieve by the following choice:

$$b^\top = \mathbf{e}_{d+1}, \quad a_j = -\frac{1}{\left(\frac{n-1}{n} + (d+2)\frac{1}{n}\right)} \mathbf{e}_j \quad \text{for } i = 1, 2, \dots, d\,,$$

where $\mathbf{e}_j$ is the j-th coordinate vector. This choice precisely corresponds to

$$b = \mathbf{e}_{d+1}, \quad A = -\frac{1}{\left(\frac{n-1}{n} + (d+2)\frac{1}{n}\right)} \begin{bmatrix} I_d \\ 0 \end{bmatrix}.$$

Proof of uniqueness: Suppose X_1 and X_2 are two minimizers of f_j, then $\langle \mathcal{M}, X_1\rangle = \langle \mathcal{M}, X_2\rangle$ almost surely for all $\mathcal{M}$. If $\langle \mathcal{M}, X_1\rangle \neq \langle \mathcal{M}, X_2\rangle$, then $f_j(\frac{1}{2}X_1 + \frac{1}{2}X_2) < \min f_j$ holds since the 1-dimensional quadratic function is strongly convex in its input. This concludes that the minimizer of f_j are a linear combination of $E_{j,d+1}$ with its transpose. Since the constraint $X = ba_j^\top$ ensures X is rank-one, then there are two possible solutions for X: $E_{j,d+1}$ or $E_{d+1,j}$. Given b is shared among all f_j, the only unique solution for X is $E_{d+1,j}$. This ensures the uniqueness of solutions for b and a_j up to scaling.

We next move on to the non-isotropic case.

A.3 Proof for the non-isotropic case

Step 1: Diagonal covariance case

We first consider the case where $x^{(i)}$ is sampled from $\mathcal{N}(0,\Lambda)$ where $\Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_d)$ and $w_\star$ is sampled from $\mathcal{N}(0, I_d)$. We prove the following generalization of Lemma 6.

Lemma 8. *Suppose that $x^{(i)}$ is sampled from $\mathcal{N}(0,\Lambda)$ where $\Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_d)$ and $w_\star$ is sampled from $\mathcal{N}(0, I_d)$. Consider the following objective*

$$f_j(X) = \mathbf{E}_{Z_0, w_\star}\left[\langle \mathcal{M}, X\rangle + w_\star[j]\right]^2 .$$

Then a global minimum is given as

$$X_j = -\frac{1}{\frac{n+1}{n}\lambda_j + \frac{1}{n}\cdot(\sum_k \lambda_k)} E_{d+1,j} ,$$

where E_{i_1,i_2} is the matrix whose (i_1,i_2)-th entry is 1, and the other entries are zero.

Proof of Lemma 8. Similarly to the proof of Lemma 6, it suffices to check that

$$2\mathbf{E}\left[\langle \mathcal{M}, X_0\rangle \mathcal{M}\right] + 2\mathbf{E}\left[w_\star[j]\mathcal{M}\right] = 0 ,$$

where we recall that $\mathcal{M}$ is defined as

$$\mathcal{M} = \frac{1}{n}\sum_i \begin{bmatrix} x^{(i)}{x^{(i)}}^\top & y^{(i)}x^{(i)} \\ y^{(i)}{x^{(i)}}^\top & {y^{(i)}}^2 \end{bmatrix} .$$

A similar calculation as the proof of Lemma 6 yields

$$\mathbf{E}\left[w_\star[j]\mathcal{M}\right] = \lambda_j(E_{d+1,j} + E_{j,d+1}). \tag{18}$$

Here the factor of λ_j comes from the following generalization of (14):

$$\mathbf{E}[w_\star[j]y^{(i)}x^{(i)}[k]] = \mathbf{E}[w_\star[j]\left\langle w_\star, x^{(i)}\right\rangle x^{(i)}[k]] = \mathbf{E}\left[w_\star[j]^2 x^{(i)}[j]x^{(i)}[k]\right] = \lambda_j \mathbb{1}_{[j=k]} .$$

Next, we compute $\mathbf{E}\left[\langle \mathcal{M}, E_{d+1,j}\rangle \mathcal{M}\right]$. Again, we follow a similar calculation to the proof of Lemma 6 except that this time we use the following generalization of (16):

$$\mathbf{E}\left[\left\langle w_\star, x^{(i)}\right\rangle x^{(i)}[j]\left\langle w_\star, x^{(i')}\right\rangle x^{(i')}[j]\right] = \begin{cases} \mathbf{E}[x^{(i)}[j]^2 x^{(i')}[j]^2] = \lambda_j^2 & \text{if } i \neq i', \\ \mathbf{E}\left[\left\langle w_\star, x^{(i)}\right\rangle^2 x^{(i)}[j]^2\right] = \lambda_j\sum_k \lambda_k + 2\lambda_j^2 & \text{if } i = i', \end{cases}$$

where the last line follows since

$$\begin{aligned} \mathbf{E}\left[\left\langle w_\star, x^{(i)}\right\rangle^2 x^{(i)}[j]^2\right] &= \mathbf{E}\left[\left\|x^{(i)}\right\|^2 x^{(i)}[j]^2\right] = \mathbf{E}\left[x^{(i)}[j]^2\sum_k x^{(i)}[k]^2\right] \\ &= \lambda_j\sum_k \lambda_k + 2\lambda_j^2 . \end{aligned}$$

Therefore, we have

$$\begin{aligned} \mathbf{E}\left[\langle \mathcal{M}, E_{d+1,j}\rangle \mathcal{M}\right] &= \frac{1}{n^2}\cdot\left(n(n-1)\lambda_j^2 + n\lambda_j\sum_k \lambda_k + 2n\lambda_j^2\right)(E_{d+1,j} + E_{j,d+1}) \\ &= \left(\frac{n+1}{n}\lambda_j^2 + \frac{1}{n}(\lambda_j\sum_k \lambda_k)\right)(E_{d+1,j} + E_{j,d+1}) . \end{aligned} \tag{19}$$

Therefore, combining (18) and (19), the results follows. □

Now we finish the proof. From Lemma 6, it follows that

$$X_j = -\frac{1}{\frac{n+1}{n}\lambda_j + \frac{1}{n}\cdot\left(\sum_k \lambda_k\right)} E_{d+1,j}$$

is the unique global minimum of f_j. Hence, b and $A = [a_1\ a_1\ \cdots\ a_d]$ achieve the global minimum of $f(b, A) = \sum_{j=1}^{d} f_j(b, A_j)$ if they satisfy

$$ba_j^\top = X_j = -\frac{1}{\frac{n+1}{n}\lambda_j + \frac{1}{n}\cdot\left(\sum_k \lambda_k\right)} E_{d+1,j} \quad \text{for all } i = 1, 2, \ldots, d.$$

This can be achieve by the following choice:

$$b^\top = \mathbf{e}_{d+1}, \quad a_j = -\frac{1}{\frac{n+1}{n}\lambda_j + \frac{1}{n}\cdot\left(\sum_k \lambda_k\right)} \mathbf{e}_j \quad \text{for } i = 1, 2, \ldots, d\,,$$

where $\mathbf{e}_j$ is the j-th coordinate vector. This choice precisely corresponds to

$$b = \mathbf{e}_{d+1}, \quad A = -\begin{bmatrix} \operatorname{diag}\left(\left\{\frac{1}{\frac{n+1}{n}\lambda_j + \frac{1}{n}\cdot\left(\sum_k \lambda_k\right)}\right\}_j\right) \\ 0 \end{bmatrix}.$$

Step 2: Non-diagonal covariance case (the setting of Theorem 1)

We finally prove the general result of Theorem 1, namely $x^{(i)}$ is sampled from a Gaussian with covariance $\Sigma = U\Lambda U^\top$ where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ and $w_\star$ is sampled from $\mathcal{N}(0, I_d)$. The proof works by reducing this case to the previous case. For each i, define $\widetilde{x}^{(i)} := U^T x^{(i)}$. Then $\mathbf{E}[\widetilde{x}^{(i)}(\widetilde{x}^{(i)})^\top] = \mathbf{E}[U^\top(U\Lambda U^\top)U] = \Lambda$. Now let us write the loss function (13) with this new coordinate system: since $x^{(i)} = U\widetilde{x}^{(i)}$, we have

$$f(b, A) = \mathbf{E}_{Z_0, w_\star}\left[(b^\top \mathcal{M} A + w_\star^\top) U \widetilde{x}^{(n+1)}\right]^2 = \sum_{j=1}^{d} \lambda_j \mathbf{E}_{Z_0, w_\star}\left[\left((b^\top \mathcal{M} A + w_\star^\top)U\right)[j]\right]^2 .$$

Hence, let us consider the vector $(b^\top \mathcal{M} A + w_\star^\top)U$. By definition of $\mathcal{M}$, we have

$$\begin{aligned}
(b^\top \mathcal{M} A + w_\star^\top)U &= \frac{1}{n}\sum_i b^\top \begin{bmatrix} x^{(i)} \\ \langle x^{(i)}, w_\star \rangle \end{bmatrix}^{\otimes 2} AU + w_\star^\top U \\
&= \frac{1}{n}\sum_i b^\top \begin{bmatrix} U\tilde{x}_i \\ \langle Ux^{(i)}, w_\star \rangle \end{bmatrix}^{\otimes 2} AU + w_\star^\top U \\
&= \frac{1}{n}\sum_i b^\top \begin{bmatrix} U & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \tilde{x}_i \\ \langle Ux^{(i)}, w_\star \rangle \end{bmatrix}^{\otimes 2} \begin{bmatrix} U^\top & 0 \\ 0 & 1 \end{bmatrix} AU + w_\star^\top U \\
&= \frac{1}{n}\sum_i \widetilde{b}^\top \begin{bmatrix} \tilde{x}_i \\ \langle x^{(i)}, \widetilde{w}_\star \rangle \end{bmatrix}^{\otimes 2} \widetilde{A} + \widetilde{w}_\star^\top
\end{aligned}$$

where we define $\widetilde{b}^\top := b^\top \begin{bmatrix} U & 0 \\ 0 & 1 \end{bmatrix}$, $\widetilde{A} := \begin{bmatrix} U^\top & 0 \\ 0 & 1 \end{bmatrix} AU$, and $\widetilde{w}_\star := U^\top w_\star$. By the rotational symmetry, $\widetilde{w}_\star$ is also distributed as $\mathcal{N}(0, I_d)$. Hence, this reduces to the previous case, and a global minimum is given as

$$\widetilde{b} = \mathbf{e}_{d+1}, \quad \widetilde{A} = -\begin{bmatrix} \operatorname{diag}\left(\left\{\frac{1}{\frac{n+1}{n}\lambda_j + \frac{1}{n}\cdot\left(\sum_k \lambda_k\right)}\right\}_j\right) \\ 0 \end{bmatrix}.$$

From the definition of $\widetilde{b}, \widetilde{A}$, it thus follows that a global minimum is given by

$$b^\top = \mathbf{e}_{d+1}, \quad A = -\begin{bmatrix} U\operatorname{diag}\left(\left\{\frac{1}{\frac{n+1}{n}\lambda_i + \frac{1}{n}\cdot\left(\sum_k \lambda_k\right)}\right\}_i\right) U^\top \\ 0 \end{bmatrix},$$

as desired.

B Proofs for the multi-layer case

B.1 Proof of Theorem 3

The proof is based on probabilistic methods (Alon and Spencer, 2016). According to Lemma 9, the objective function can be written as (for more details check the derivations in (20))

$$f(A_1, A_2) = \mathbf{E}\,\mathrm{Tr}\left(\mathbf{E}\left[\prod_{i=1}^{2}(I - X_0^\top A_i X_0 M)X_0^\top w_\star w_\star^\top X_0 \prod_{i=1}^{2}(I - M X_0^T A_i X_0)\right]\right)$$
$$= \mathbf{E}\,\mathrm{Tr}\left(\mathbf{E}\left[\prod_{i=2}^{1}(I - X_0^\top A_i X_0 M)X_0^\top X_0 \prod_{j=1}^{2}(I - M X_0^T A_j X_0)\right]\right),$$

where we use the isotropy of $w_\star$ and the linearity of trace to get the last equation. Suppose that A_0^* and A_1^* denote the global minimizer of f over symmetric matrices. Since A_1^* is a symmetric matrix, it admits the spectral decomposition $A_1 = UD_1U^\top$ where D_1 is a diagonal matrix and U is an orthogonal matrix. Remarkably, the distribution of X_0 is invariant to a linear transformation by an orthogonal matrix, i.e, X_0 has the same distribution as $X_0U^\top$. This invariance yields

$$f(UD_1U^\top, A_2^*) = f(D_1, U^\top A_2^* U).$$

Thus, we can assume A_1^* is diagonal without loss of generality. To prove A_2^* is also diagonal, we leverage a probabilistic proof technique. Consider the random diagonal matrix S whose diagonal elements are either 1 or -1 with probability $\frac{1}{2}$. Since the input distribution is invariant to orthogonal transformations, we have

$$f(D_1, A_2^*) = f(SD_1S, SA_2^*S) = f(D_1, SA_2^*S).$$

Note that we use $SD_1S = D_1$ in the last equation, which holds due to D_1 and S are diagonal matrices and S has diagonal elements in $\{+1, -1\}$. Since f is convex in A_2, a straightforward application of Jensen's inequality yields

$$f(D_1, A_2^*) = \mathbf{E}\left[f(D_1, SA_2^*S)\right] \geq f(D_1, \mathbf{E}\left[SA_2^*S\right]) = f(D_1, \mathbf{diag}(A_2^*)).$$

Thus, there are diagonal D_1 and $\mathbf{diag}(A_2^*)$ for which $f(D_1, \mathbf{diag}(A_2^*)) \leq f(A_1^*, A_2^*)$ holds for an optimal A_1^* and A_2^*. This concludes the proof.

B.2 Proof of Theorem 4

Let us drop the factor of $\frac{1}{n}$ which was present in the original update (51). This is because the constant $1/n$ can be absorbed into A_i's. Doing so does not change the theorem statement, but reduces notational clutter.

Let us consider the reformulation of the in-context loss f presented in Lemma 9. Specifically, let $\overline{Z}_0$ be defined as

$$\overline{Z}_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & y^{(n+1)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)},$$

where $y^{(n+1)} = \langle w_\star, x^{(n+1)} \rangle$. Let $\overline{Z}_i$ denote the output of the $(i-1)^{th}$ layer of the linear transformer (as defined in (51), initialized at $\overline{Z}_0$). For the rest of this proof, we will drop the bar, and simply denote $\bar{Z}_i$ by Z_i.[2] Let $X_i \in \mathbb{R}^{d\times n+1}$ denote the first d rows of Z_i and let $Y_i \in \mathbb{R}^{1\times n+1}$ denote the $(d+1)^{th}$ row of Z_k. Under the sparsity pattern enforced in (9), we verify that, for any $i \in \{0...k\}$,

$$X_i = X_0,$$
$$Y_{i+1} = Y_i + Y_i M X_i^\top A_i X_i = Y_0 \prod_{\ell=0}^{i}\left(I + M X_0^T A_\ell X_0\right). \tag{20}$$

[2] This use of Z_i differs the original definition in (1). But we will not refer to the original definition anywhere in this proof.

where $M = \begin{bmatrix} I_{n\times n} & 0 \\ 0 & 0 \end{bmatrix}$. We adopt the shorthand $A = \{A_i\}_{i=0}^k$.

We adopt the shorthand $A = \{A_i\}_{i=0}^k$. Let $\mathcal{S} \subset \mathbb{R}^{(k+1)\times d\times d}$, and $A \in \mathcal{S}$ if and only if for all $i \in \{0...k\}$, there exists scalars $a_i \in \mathbb{R}$ such that $A_i = a_i\Sigma^{-1}$ and $B_i = b_i I$. We use $f(A)$ to refer to the in-context loss of Theorem 4, that is,

$$f(A) := f\left(\left\{Q_i = \begin{bmatrix} A_i & 0 \\ 0 & 0 \end{bmatrix}, P_i = \begin{bmatrix} 0_{d\times d} & 0 \\ 0 & 1 \end{bmatrix}\right\}_{i=0}^k\right).$$

Throughout this proof, we will work with the following formulation of the *in-context loss* from Lemma 9:

$$f(A) = \mathbf{E}_{(X_0,w_\star)}\left[\operatorname{Tr}\left(\left(I-M\right)Y_{k+1}^\top Y_{k+1}\left(I-M\right)\right)\right]. \tag{21}$$

The theorem statement is equivalent to the following:

$$\inf_{A\in\mathcal{S}} \sum_{i=0}^k \|\nabla_{A_i} f(A)\|_F^2 = 0, \tag{22}$$

where $\nabla_{A_i} f$ denotes derivative wrt the Frobenius norm $\|A_i\|_F$. Towards this end, we establish the following intermediate result: if $A \in \mathcal{S}$, then for any $R \in \mathbb{R}^{(k+1)\times d\times d}$, there exists $\tilde{R} \in \mathcal{S}$, such that, at $t = 0$,

$$\frac{d}{dt} f(A + t\tilde{R}) \le \frac{d}{dt} f(A + tR). \tag{23}$$

In fact, we show that $\tilde{R}_i := r_i I$, for $r_i = \frac{1}{d}\operatorname{Tr}\left(\Sigma^{1/2} R_i \Sigma^{1/2}\right)$. This implies (22) via the following simple argument: Consider the "$\mathcal{S}$-constrained gradient flow": let $A(t) : \mathbb{R}^+ \to \mathbb{R}^{(k+1)\times d\times d}$ be defined as

$$\frac{d}{dt} A_i(t) = -r_i(t)\Sigma^{-1}, \quad r_i(t) := \operatorname{Tr}(\Sigma^{1/2}\nabla_{A_i} f(A(t))\Sigma^{1/2})$$

for $i = 0...k$. By (23), we verify that

$$\frac{d}{dt} f(A(t)) \le -\sum_{i=0}^k \|\nabla_{A_i} f(A(t))\|_F^2. \tag{24}$$

We verify from its definition that $f(A) \ge 0$; if the infimum in (22) fails to be zero, then inequality (24) will ensure unbounded descent as $t \to \infty$, contradicting the fact that $f(A)$ is lower-bounded. This concludes the proof.

Step 0: Proof outline

The remainder of the proof will be devoted to showing (23), which we outline as follows:

- In Step 1, we reduce the condition in (24) to a more easily verified *layer-wise* condition. Specifically, we only need to verify (24) when R_i are all zero except for R_j for some fixed j (see (25))

  At the end of Step 1, we set up some additional notation, and introduce an important matrix G, which is roughly "a product of attention layer matrices". In (26), we study the evolution of $f(A(t))$ when $A(t)$ moves in the direction of R, as X_0 is (roughly speaking) randomly transformed.
- In Step 2, we use the results of Step 2 to to study G (see (27)) and $\frac{d}{dt}G(A(t))$ (see (28)) under random transformation of X_0. The idea in (28) is that "randomly transforming X_0" has the same effect as "randomly transforming S" (recall S is the perturbation to B).
- In Step 3, we apply the result from Step 2 to the expression of $\frac{d}{dt}f(A(t))$ in (26). We verify that $\tilde{R}$ in (23) is exactly the expected matrix after "randomly transforming S". This concludes our proof.

Step 1: Reduction to layer-wise condition

To prove (23), it suffices to show the following simpler condition: Let $j \in \{0...k\}$. Let $R_j \in \mathbb{R}^{d\times d}$ be arbitrary matrices. For $C \in \mathbb{R}^{d\times d}$, let $A(tC, j)$ denote the collection of matrices, where

$[A(tC,j)]_j = A_j + tC$, and for $i \neq j$, $A(tC,j)_i = A_i$. We show that for all $j \in \{0...k\}$, $R_j \in \mathbb{R}^{d\times d}$, there exists $\tilde{R}_j = r_j\Sigma^{-1}$, such that, at $t=0$,

$$\frac{d}{dt} f(A(t\tilde{R}_j, j)) \leq \frac{d}{dt} f(A(tR_j, j)) \tag{25}$$

We can verify that (23) is equivalent to (25) by noticing that for any R, at $t=0$, $\frac{d}{dt} f(A+tR) = \sum_{j=0}^{k} \frac{d}{dt} f(A(tR_j, j))$. We will now work towards proving (25) for some index j that is arbitrarily chosen but fixed throughout.

By (20) and (21),

$$\begin{aligned}
& f(A(tR_j, j)) \\
=& \mathbf{E}\left[\mathrm{Tr}\left((I-M)\, Y_{k+1}^\top Y_{k+1}\,(I-M)\right)\right] \\
=& \mathbf{E}\left[\mathrm{Tr}\left((I-M)\, G(X_0, A_j + tR_j)^\top w_\star^\top w_\star G(X_0, A_j + tR_j)\,(I-M)\right)\right] \\
=& \mathbf{E}\left[\mathrm{Tr}\left((I-M)\, G(X_0, A_j + tR_j)^\top \Sigma^{-1} G(X_0, A_j + tR_j)\,(I-M)\right)\right]
\end{aligned}$$

where $G(X, A_j + C) := X \prod_{i=0}^{k}\left(I - MX_0^\top [A(tC,j)]_i X\right)$. The second equality follows from plugging in (20). For the rest of this proof, let U denote a uniformly randomly sampled orthogonal matrix. Let $U_\Sigma := \Sigma^{1/2} U \Sigma^{-1/2}$. Using the fact that $X_0 \overset{d}{=} U_\Sigma X_0$, we can verify

$$\begin{aligned}
& \frac{d}{dt} f(A(tR_j, j))\Big|_{t=0} \\
=& \frac{d}{dt}\mathbf{E}\left[\mathrm{Tr}\left((I-M)\, G(X_0, A_j + tR_j)^\top \Sigma^{-1} G(X_0, A_j + tR_j)\,(I-M)\right)\right]\Big|_{t=0} \\
=& \frac{d}{dt}\mathbf{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\, G(U_\Sigma X_0, A_j + tR_j)^\top \Sigma^{-1} G(U_\Sigma X_0, A_j + tR_j)\,(I-M)\right)\right]\Big|_{t=0} \\
=& 2\mathbf{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\, G(U_\Sigma X_0, A_j)^\top \Sigma^{-1}\, \frac{d}{dt} G(U_\Sigma X_0, A_j + tR_j)\Big|_{t=0} (I-M)\right)\right].
\end{aligned} \tag{26}$$

Step 2: G and $\frac{d}{dt}G$ under random transformation of X_0

We will now verify that $G(U_\Sigma X_0, A_j) = U_\Sigma G(X_0, A_j)$:

$$\begin{aligned}
& G(U_\Sigma X_0, A_j) \\
=& U_\Sigma X_0 \prod_{i=0}^{k}\left(I + MX_0^T U_\Sigma^\top A_i U_\Sigma X_0\right) \\
=& U_\Sigma G(X_0, A_j),
\end{aligned} \tag{27}$$

where we use the fact that $U_\Sigma^\top A_i U_\Sigma = U_\Sigma^\top (a_i \Sigma^{-1}) U_\Sigma = A_i$. Next, we verify that

$$\begin{aligned}
\frac{d}{dt} G(U_\Sigma X_0, R_j) =& U_\Sigma X_0 \left(\prod_{i=0}^{j-1}(I + MX_0^T A_i X_0)\right) MX_0^T U_\Sigma^\top R_j U_\Sigma X_0 \prod_{i=j+1}^{k}(I + MX_0^T A_i X_0) \\
=& U_\Sigma \frac{d}{dt} G(X_0, U_\Sigma^\top R_j U_\Sigma)
\end{aligned} \tag{28}$$

where the first equality again uses the fact that $U_\Sigma^\top A_i U_\Sigma = A_i$.

Step 3: Putting everything together

Let us continue from (26). Plugging (27) and (28) into (26),

$$\begin{aligned}
& \frac{d}{dt} f(A(tR_j, j))\Big|_{t=0} \\
=& 2\mathbf{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\, G(U_\Sigma X_0, A_j)^\top \Sigma^{-1}\, \frac{d}{dt} G(U_\Sigma X_0, A_j + tR_j)\Big|_{t=0} (I-M)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(i)}{=}2\mathbf{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G(X_0,A_j)^\top\Sigma^{-1}\,\frac{d}{dt}G(X_0,A_j+tU_\Sigma^\top R_jU_\Sigma)\Big|_{t=0}(I-M)\right)\right]\\
&=2\mathbf{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0,A_j)^\top\Sigma^{-1}\mathbf{E}_U\left[\frac{d}{dt}G(X_0,A_j+tU_\Sigma^\top R_jU_\Sigma)\Big|_{t=0}\right](I-M)\right)\right]\\
&\overset{(ii)}{=}2\mathbf{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0,A_j)^\top\Sigma^{-1}\,\frac{d}{dt}G(X_0,A_j+t\mathbf{E}_U\left[U_\Sigma^\top R_jU_\Sigma\right])\Big|_{t=0}(I-M)\right)\right]\\
&=2\mathbf{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0,A_j)^\top\Sigma^{-1}\,\frac{d}{dt}G(X_0,A_j+t\cdot r_j\Sigma^{-1})\Big|_{t=0}(I-M)\right)\right]\\
&=\frac{d}{dt}f(A(t\cdot r_j\Sigma^{-1},j))\Big|_{t=0},
\end{aligned}
$$

where $r_j := \frac{1}{d}\mathrm{Tr}\left(\Sigma^{1/2}R_j\Sigma^{1/2}\right)$. In the above, (i) uses 1. (27) and (28), as well as the fact that $U_\Sigma^\top\Sigma^{-1}U_\Sigma = \Sigma^{-1}$. (ii) uses the fact that $\frac{d}{dt}G(X_0,A_j+tC)\big|_{t=0}$ is affine in C. To see this, one can verify from the definition of G, e.g. using similar algebra as (28), that $\frac{d}{dt}G(X_0,A_j+C)$ is affine in C. Thus $\mathbf{E}_U\left[G(X_0,A_j+tU_\Sigma^\top R_jU_\Sigma)\right] = G(X_0,A_j+t\mathbf{E}_U\left[U_\Sigma^\top R_jU_\Sigma\right])$.

B.3 Proof of Theorem 5

The proof of Theorem 5 is similar to that of Theorem 4, and with a similar setup. However to keep the proof self-contained, we will restate the setup. Once again, we drop the factor of $\frac{1}{n}$ which was present in the original update (51). This is because the constant $1/n$ can be absorbed into A_i's. Doing so does not change the theorem statement, but reduces notational clutter.

Let us consider the reformulation of the in-context loss f presented in Lemma 9. Specifically, let $\overline{Z}_0$ be defined as

$$
\overline{Z}_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & y^{(n+1)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)},
$$

where $y^{(n+1)} = \langle w_\star, x^{(n+1)}\rangle$. Let $\overline{Z}_i$ denote the output of the $(i-1)^{th}$ layer of the linear transformer (as defined in (51), initialized at $\overline{Z}_0$). For the rest of this proof, we will drop the bar, and simply denote $\bar{Z}_i$ by Z_i.[3] Let $X_i \in \mathbb{R}^{d\times n+1}$ denote the first d rows of Z_i and let $Y_i \in \mathbb{R}^{1\times n+1}$ denote the $(d+1)^{th}$ row of Z_k. Under the sparsity pattern enforced in (11), we verify that, for any $i \in \{0...k\}$,

$$
\begin{aligned}
&X_{i+1} = X_i + B_iX_iMX_i^\top A_iX_i\\
&Y_{i+1} = Y_i + Y_iMX_i^\top A_iX_i = Y_0\prod_{\ell=0}^{i}\left(I+MX_\ell^TA_\ell X_\ell\right).
\end{aligned}
\tag{29}
$$

We adopt the shorthand $A = \{A_i\}_{i=0}^k$ and $B = \{B_i\}_{i=0}^k$. Let $\mathcal{S} \subset \mathbb{R}^{2\times(k+1)\times d\times d}$, and $(A,B)\in\mathcal{S}$ if and only if for all $i\in\{0...k\}$, there exists scalars $a_i, b_i \in \mathbb{R}$ such that $A_i = a_i\Sigma^{-1}$ and $B_i = b_iI$. Throughout this proof, we will work with the following formulation of the *in-context loss* from Lemma 9:

$$
f(A,B) := \mathbf{E}_{(X_0,w_\star)}\left[\mathrm{Tr}\left((I-M)\,Y_{k+1}^\top Y_{k+1}\,(I-M)\right)\right]. \tag{30}
$$

(note that the only randomness in Z_0 comes from X_0 as Y_0 is a deterministic function of X_0). The theorem statement is equivalent to the following:

$$
\inf_{(A,B)\in\mathcal{S}}\sum_{i=0}^{k}\|\nabla_{A_i}f(A,B)\|_F^2 + \|\nabla_{B_i}f(A,B)\|_F^2 = 0 \tag{31}
$$

where $\nabla_{A_i}f$ denotes derivative wrt the Frobenius norm $\|A_i\|_F$.

[3]This use of Z_i differs the original definition in (1). But we will not refer to the original definition anywhere in this proof.

Our goal is to show that, if $(A, B) \in \mathcal{S}$, then for any $(R, S) \in \mathbb{R}^{2\times(k+1)\times d\times d}$, there exists $(\tilde{R}, \tilde{S}) \in \mathcal{S}$, such that, at $t = 0$,

$$\frac{d}{dt} f(A + t\tilde{R}, B + t\tilde{S}) \leq \frac{d}{dt} f(A + tR, B + tS). \tag{32}$$

In fact, we show that $\tilde{R}_i := r_i I$, for $r_i = \frac{1}{d}\mathrm{Tr}\left(\Sigma^{1/2} R_i \Sigma^{1/2}\right)$ and $\tilde{S}_i = s_i I$, for $s_i = \frac{1}{d}\mathrm{Tr}\left(\Sigma^{-1/2} S_i \Sigma^{1/2}\right)$. This implies (31) via the following simple argument: Consider the "$\mathcal{S}$-constrained gradient flow": let $A(t) : \mathbb{R}^+ \to \mathbb{R}^{(k+1)\times d\times d}$ and $B(t) : \mathbb{R}^+ \to \mathbb{R}^{(k+1)\times d\times d}$ be defined as

$$\begin{aligned}
\frac{d}{dt} A_i(t) &= -r_i(t)\Sigma^{-1}, \quad r_i(t) := \mathrm{Tr}(\Sigma^{1/2} \nabla_{A_i} f(A(t), B(t)) \Sigma^{1/2}) \\
\frac{d}{dt} B_i(t) &= -s_i(t)\Sigma^{-1}, \quad s_i(t) := \mathrm{Tr}(\Sigma^{-1/2} \nabla_{B_i} f(A(t), B(t)) \Sigma^{1/2}),
\end{aligned}$$

for $i = 0...k$. By (32), we verify that

$$\frac{d}{dt} f(A(t), B(t)) \leq -\left(\sum_{i=0}^{k} \|\nabla_{A_i} f(A(t), B(t))\|_F^2 + \|\nabla_{B_i} f(A(t), B(t))\|_F^2\right). \tag{33}$$

We verify from its definition that $f(A, B) \geq 0$; if (31) does not hold then (33) will ensure unbounded descent as $t \to \infty$, contradicting the fact that $f(A, B)$ is lower-bounded. This concludes the proof.

Step 0: Proof outline

The remainder of the proof will be devoted to showing (32), which we outline as follows:

- In Step 1, we reduce the condition in (32) to a more easily verified *layer-wise* condition. Specifically, we only need to verify (32) in one of the two cases: (I) when R_i, S_i are all zero except for R_j for some fixed j (see (35)), or (II) when R_i, S_i are all zero except for S_j for some fixed j (see (34)).
  We focus on the proof of (II), as the proof of (I) is almost identical. At the end of Step 1, we set up some additional notation, and introduce an important matrix G, which is roughly "a product of attention layer matrices". In (36), we study the evolution of $f(A, B(t))$ when $B(t)$ moves in the direction of S, as X_0 is (roughly speaking) randomly transformed. This motivates the subsequent analysis in Steps 2 and 3 below.
- In Step 2, we study how outputs of each layer (29) changes when X_0 is randomly transformed. There are two main results here: First we provide the expression for X_i in (37). Second, we provide the expression for $\frac{d}{dt} X_i(B(t))$ in (38).
- In Step 3, we use the results of Step 2 to to study G (see (42)) and $\frac{d}{dt} G(B(t))$ (see (43)) under random transformation of X_0.
  The idea in (43) is that "randomly transforming X_0" has the same effect as "randomly transforming S" (recall S is the perturbation to B).
- In Step 4, we use the results from Steps 2 and 3 to the expression of $\frac{d}{dt} f(A, B(t))$ in (36). We verify that $\tilde{S}$ in (32) is exactly the expected matrix after "randomly transforming S". This concludes our proof of (II).
- In Step 5, we sketch the proof of (I), which is almost identical to Steps 2-4.

Step 1: Reduction to layer-wise condition

To prove (32), it suffices to show the following simpler condition: Let $j \in \{0...k\}$. Let $R_j, S_j \in \mathbb{R}^{d\times d}$ be arbitrary matrices. For $C \in \mathbb{R}^{\mathbf{d}\times d}$, let $A(tC, j)$ denote the collection of matrices, where $A(tC, j)_j = A_j + tC$, and for $i \neq j$, $A(tC, j)_i = A_i$. Define $B(tC, j)$ analogously. We show that for all $j \in \{0...k\}$ and all $R_j, S_j \in \mathbb{R}^{d\times d}$, there exists $\tilde{R}_j = r_j\Sigma^{-1}$ and $\tilde{S}_j = s_j\Sigma^{-1}$, such that, at $t = 0$,

$$\frac{d}{dt} f(A(t\tilde{R}_j, j), B) \leq \frac{d}{dt} f(A(tR_j, j), B) \tag{34}$$

$$\text{and} \qquad \frac{d}{dt} f(A, B(t\tilde{S}_j, j)) \leq \frac{d}{dt} f(A, B(tS_j, j)). \tag{35}$$

We can verify that (32) is equivalent to (34)+(35) by noticing that for any $(R, S) \in \mathbb{R}^{2\times(k+1)\times d\times d}$, at $t=0$, $\frac{d}{dt} f(A+tR, B+tS) = \sum_{j=0}^{k} \left(\frac{d}{dt} f(A(tR_j, j), B) + \frac{d}{dt} f(A, B(tS_j, j))\right)$.

We will first focus on proving (35) (the proof of (34) is similar, and we present it in Step 5 at the end), for some index j that is arbitrarily chosen but fixed throughout. Notice that X_i and Y_i in (29) are in fact functions of A, B and X_0. For most of our subsequent discussion, A_i (for all i) and B_i (for all $i \neq j$) can be treated as constant matrices. We will however make the dependence on X_0 and B_j explicit (as we consider the curve $B_j + tS$), i.e. we use $X_i(X, C)$ (resp $Y_i(X, C)$) to denote the value of X_i (resp Y_i) from (29), with $X_0 = X$, and $B_j = C$.

By (30) and (29),

$$\begin{aligned}
& f(A, B(tS_j, j)) \\
=& \mathbf{E}\left[\operatorname{Tr}\left((I-M)\, Y_{k+1}(X_0, B_j + tS)^\top Y_{k+1}(X_0, B_j + tS_j)\,(I-M)\right)\right] \\
=& \mathbf{E}\left[\operatorname{Tr}\left((I-M)\, G(X_0, B_j + tS_j)^\top w_\star^\top w_\star G(X_0, B_j + tS_j)\,(I-M)\right)\right] \\
=& \mathbf{E}\left[\operatorname{Tr}\left((I-M)\, G(X_0, B_j + tS_j)^\top \Sigma^{-1} G(X_0, B_j + tS_j)\,(I-M)\right)\right]
\end{aligned}$$

where $G(X, C) := X \prod_{i=0}^{k}\left(I - M X_i(X, C)^T A_i X_i(X, C)\right)$. The second equality follows from plugging in (29).

For the rest of this proof, let U denote a uniformly randomly sampled orthogonal matrix. Let $U_\Sigma := \Sigma^{1/2} U \Sigma^{-1/2}$. Using the fact that $X_0 \overset{d}{=} U_\Sigma X_0$, we can verify

$$\begin{aligned}
& \frac{d}{dt} f(A, B(tS_j, j))\Big|_{t=0} \\
=& \frac{d}{dt}\mathbf{E}_{X_0}\left[\operatorname{Tr}\left((I-M)\, G(X_0, B_j + tS_j)^\top \Sigma^{-1} G(X_0, B_j + tS_j)\,(I-M)\right)\right]\Big|_{t=0} \\
=& \frac{d}{dt}\mathbf{E}_{X_0, U}\left[\operatorname{Tr}\left((I-M)\, G(U_\Sigma X_0, B_j + tS_j)^\top \Sigma^{-1} G(U_\Sigma X_0, B_j + tS_j)\,(I-M)\right)\right]\Big|_{t=0} \\
=& 2\mathbf{E}_{X_0, U}\left[\operatorname{Tr}\left((I-M)\, G(U_\Sigma X_0, B_j)^\top \Sigma^{-1} \frac{d}{dt} G(U_\Sigma X_0, B_j + tS_j)\Big|_{t=0} (I-M)\right)\right]. \qquad (36)
\end{aligned}$$

Step 2: X_i and $\frac{d}{dt} X_i$ under random transformation of X_0

In this step, we prove that when X_0 is transformed by U_Σ, X_i for $i \geq 1$ are likewise transformed in a simple manner. The first goal of this step is to show

$$X_i(U_\Sigma X_0, B_j) = U_\Sigma X_i(X_0, B_j). \qquad (37)$$

We will prove this by induction. When $i = 0$, this clearly holds by definition. Suppose that (37) holds for some i. Then

$$\begin{aligned}
& X_{i+1}(U_\Sigma X_0, B_j) \\
=& X_i(U_\Sigma X_0, B_j) + B_i X_i(U_\Sigma X_0, B_j) M X_i(U_\Sigma X_0, B_j)^T A_i X_i(U_\Sigma X_0, B_j) \\
=& U_\Sigma X_i(X_0, B_j) + U_\Sigma B_i X_i(X_0, B_j) M X_i(X_0, B_j)^T A_i X_i(X_0, B_j) \\
=& U_\Sigma X_{i+1}(X_0, B_j)
\end{aligned}$$

where the second equality uses the inductive hypothesis, and the fact that $A_i = a_i \Sigma^{-1}$, so that $U_\Sigma^T A_i U_\Sigma = A_i$, and the fact that $B_i = b_i I$, from the definition of $\mathcal{S}$ and our assumption that $(A, B) \in \mathcal{S}$. This concludes the proof of (37).

We now present the second main result of this step. Let $U_\Sigma^{-1} := \Sigma^{1/2} U^T \Sigma^{-1/2}$, so that it satisfies $U_\Sigma U_\Sigma^{-1} = U_\Sigma^{-1} U_\Sigma = I$. For all i,

$$U_\Sigma^{-1} \frac{d}{dt} X_i(U_\Sigma X_0, B_j + tS_j)\Big|_{t=0} = \frac{d}{dt} X_i(X_0, B_j + t U_\Sigma^{-1} S_j U_\Sigma)\Big|_{t=0}. \qquad (38)$$

To reduce notation, we will not write $\cdot|_{t=0}$ explicitly in the subsequent proof. We first write down the dynamics for the right-hand-side term of (38): From (29), for any $\ell \leq j$, and for any $i \geq j+1$, and for any $C \in \mathbb{R}^{d\times d}$,

$$
\begin{aligned}
&\frac{d}{dt}X_\ell\left(X_0, B_j + tC\right) = 0\\
&\frac{d}{dt}X_{j+1}\left(X_0, B_j + tC\right) = CX_j\left(X_0, B_j\right)MX_j\left(X_0, B_j\right)^\top A_j X_j\left(X_0, B_j\right)\\
&\frac{d}{dt}X_{i+1}\left(X_0, B_j + tC\right) = \frac{d}{dt}X_i\left(X_0, B_j + tC\right)\\
&\qquad\qquad + B_i\left(\frac{d}{dt}X_i\left(X_0, B_j + tC\right)\right)MX_i\left(X_0, B_j\right)^\top A_i X_i\left(X_0, B_j\right)\\
&\qquad\qquad + B_i X_i\left(X_0, B_j\right)M\left(\frac{d}{dt}X_i\left(X_0, B_j + tC\right)\right)^\top A_i X_i\left(X_0, B_j\right)\\
&\qquad\qquad + B_i X_i\left(X_0, B_j\right)MX_i\left(X_0, B_j\right)^\top A_i\left(\frac{d}{dt}X_i\left(X_0, B_j + tC\right)\right)
\end{aligned}
\tag{39}
$$

We are now ready to prove (38) using induction. For the base case, we verify that for $\ell \le j$, $U_\Sigma^{-1}\frac{d}{dt}X_\ell\left(U_\Sigma X_0, B_k + tS_j\right) = 0 = \frac{d}{dt}X_\ell\left(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma\right)$ (see first equation in (39)). For index $j+1$, we verify that

$$
\begin{aligned}
U_\Sigma^{-1}\frac{d}{dt}X_{j+1}\left(U_\Sigma X_0, B_j + tS_j\right) =& U_\Sigma^{-1}S_jU_\Sigma X_j(X_0, B_j)MX_j(U_\Sigma X_0, B_j)^\top A_j\\
=& \frac{d}{dt}X_{j+1}\left(U_\Sigma X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma X_j\right)
\end{aligned}
\tag{40}
$$

where we use two facts: 1. $X_i(U_\Sigma X_0, B_j) = U_\Sigma X_i(X_0, B_j)$ from (37), 2. $A_i = a_i\Sigma^{-1}$, so that $U_\Sigma^\top A_i U_\Sigma = A_i$. We verify by comparison to the second equation in (39) that $U_\Sigma^{-1}\frac{d}{dt}X_j\left(U_\Sigma X_0, B_j + tS_j\right) = 0 = \frac{d}{dt}X_j\left(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma\right)$. These conclude the proof of the base case.

Now suppose that (38) holds for some i. We will now prove (38) holds for $i+1$. From (29),

$$
\begin{aligned}
&U_\Sigma^{-1}\frac{d}{dt}X_{i+1}\left(U_\Sigma X_0, B_j + tS_j\right)\\
=&U_\Sigma^{-1}\frac{d}{dt}\left(X_i\left(U_\Sigma X_0, B_j + tS_j\right)\right)\\
&\quad + U_\Sigma^{-1}\frac{d}{dt}\Big(B_i X_i\left(U_\Sigma X_0, B_j + tS_j\right)MX_i\left(U_\Sigma X_0, B_j + tS_j\right)^\top A_i X_i\left(U_\Sigma X_0, B_j + tS_j\right)\Big)\\
=&U_\Sigma^{-1}\frac{d}{dt}\left(X_i\left(U_\Sigma X_0, B_j + tS_j\right)\right)\\
&\quad + U_\Sigma^{-1}B_i\left(\frac{d}{dt}X_i\left(U_\Sigma X_0, B_j + tS_j\right)\right)MX_i\left(U_\Sigma X_0, B_j\right)^\top A_i X_i\left(U_\Sigma X_0, B_j\right)\\
&\quad + U_\Sigma^{-1}B_i X_i\left(U_\Sigma X_0, B_j\right)M\left(\frac{d}{dt}X_i\left(U_\Sigma X_0, B_j + tS_j\right)\right)^\top A_i X_i\left(U_\Sigma X_0, B_j\right)\\
&\quad + U_\Sigma^{-1}B_i X_i\left(U_\Sigma X_0, B_j\right)MX_i\left(U_\Sigma X_0, B_j\right)^\top A_i\left(\frac{d}{dt}X_i\left(U_\Sigma X_0, B_j + tS_j\right)\right)\\
\overset{(i)}{=}&U_\Sigma^{-1}\frac{d}{dt}X_i\left(U_\Sigma X_0, B_j + tS_j\right)\\
&\quad + B_i\left(U_\Sigma^{-1}\frac{d}{dt}X_i\left(U_\Sigma X_0, B_j + tS_j\right)\right)MX_i\left(X_0, B_j\right)^\top A_i X_i\left(X_0, B_j\right)\\
&\quad + B_i X_i\left(X_0, B_j\right)M\left(U_\Sigma^{-1}\frac{d}{dt}X_i\left(U_\Sigma X_0, B_j + tS_j\right)\right)^\top A_i X_i\left(X_0, B_j\right)\\
&\quad - B_i X_i\left(X_0, B_j\right)MX_i\left(X_0, B_j\right)^\top A_i\left(U_\Sigma^{-1}\frac{d}{dt}X_i\left(U_\Sigma X_0, B_j + tS_j\right)\right)\\
\overset{(ii)}{=}&\frac{d}{dt}X_i\left(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ B_i \left(\frac{d}{dt} X_i \left(X_0, B_j + tU_\Sigma^{-1} S_j U_\Sigma \right) \right) M X_i \left(X_0, B_j \right)^\top A_i X_i \left(X_0, B_j \right) \\
&+ B_i X_i \left(X_0, B_j \right) M \left(\frac{d}{dt} X_i \left(X_0, B_j + tU_\Sigma^{-1} S_j U_\Sigma \right) \right)^\top A_i X_i \left(X_0, B_j \right) \\
&+ B_i X_i \left(X_0, B_j \right) M X_i \left(X_0, B_j \right)^\top A_i \left(\frac{d}{dt} X_i \left(X_0, B_j + tU_\Sigma^{-1} S_j U_\Sigma \right) \right)
\end{aligned}
\tag{41}
$$

In (i) above, we crucially use the following facts: 1. $B_i = b_i I$ so that $U_\Sigma^{-1} B_i = B_i U_\Sigma^{-1}$, 2. $X_i(U_\Sigma X_0, B_j) = U_\Sigma X_i(X_0, B_j)$ from (37), 3. $A_i = a_i \Sigma^{-1}$, so that $U_\Sigma^\top A_i U_\Sigma = A_i$, 4. $U_\Sigma U_\Sigma^{-1} = U_\Sigma^{-1} U_\Sigma = I$. (ii) follows from our inductive hypothesis. The inductive proof is complete by verifying that (41) exactly matches the third equation of (39) when $C = U_\Sigma^{-1} S U_\Sigma$.

Step 3: G and $\frac{d}{dt}G$ under random transformation of X_0

We now verify that $G(U_\Sigma X_0, B_j) = U_\Sigma G(X_0, B_j)$. This is a straightforward consequence of (37) as

$$
\begin{aligned}
&G(U_\Sigma X_0, B_j) \\
=&U_\Sigma X_0 \prod_{i=0}^{k} \left(I + M X_i(U_\Sigma X_0, B_j)^T A_i X_i(U_\Sigma X_0, B_j) \right) \\
=&U_\Sigma X_0 \prod_{i=0}^{k} \left(I + M X_i(X_0, B_j)^T A_i X_i(X_0, B_j) \right) \\
=&U_\Sigma G(X_0, B_j),
\end{aligned}
\tag{42}
$$

where the second equality uses (37), as well as the fact that $U_\Sigma^\top A_i U_\Sigma = A_i$. Next, we will show that

$$
U_\Sigma^{-1} \left. \frac{d}{dt} G(U_\Sigma X_0, B_j + tS_j) \right|_{t=0} = \left. \frac{d}{dt} G(X_0, B_j + tU_\Sigma^{-1} S_j U_\Sigma) \right|_{t=0}. \tag{43}
$$

To see this, we can expand

$$
\begin{aligned}
&U_\Sigma^{-1} \frac{d}{dt} G(U_\Sigma X_0, B_j + tS_j) \\
=&U_\Sigma^{-1} \frac{d}{dt} \left(U_\Sigma X_0 \prod_{i=0}^{k} \left(I + M X_i(U_\Sigma X_0, B_j + tS_j)^T A_i X_i(U_\Sigma X_0, B_j + tS_j) \right) \right) \\
=&X_0 \sum_{i=0}^{k} \left(\prod_{\ell=0}^{i-1} \left(I + M X_\ell(U_\Sigma X_0, B_j)^T A_\ell X_i(U_\Sigma X_0, B_\ell) \right) \right) \\
&\cdot M \frac{d}{dt} \left(X_i(U_\Sigma X_0, B_j + tS_j)^T A_i X_i(U_\Sigma X_0, B_j) \right) \\
&\cdot \left(\prod_{\ell=i+1}^{k} \left(I + M X_\ell(U_\Sigma X_0, B_j)^T A_\ell X_i(U_\Sigma X_0, B_\ell) \right) \right) \\
\overset{(i)}{=}&X_0 \sum_{i=0}^{k} \left(\prod_{\ell=0}^{i-1} \left(I + M X_\ell(X_0, B_j)^T A_\ell X_\ell(X_0, B_\ell) \right) \right) \\
&\cdot M \left(\left(U_\Sigma^{-1} \frac{d}{dt} X_i(U_\Sigma X_0, B_j + tS_j) \right)^T A_i X_i(X_0, B_j) + M X_i(X_0, B_j)^T A_i \left(U_\Sigma^{-1} \frac{d}{dt} X_i(U_\Sigma X_0, B_j + tS_j) \right) \right) \\
&\cdot \left(\prod_{\ell=i+1}^{k} \left(I + M X_\ell(X_0, B_j)^T A_\ell X_\ell(X_0, B_\ell) \right) \right) \\
\overset{(ii)}{=}&X_0 \sum_{i=0}^{k} \left(\prod_{\ell=0}^{i-1} \left(I + M X_\ell(X_0, B_j)^T A_\ell X_\ell(X_0, B_\ell) \right) \right)
\end{aligned}
$$

$$\cdot M\left(\left(\frac{d}{dt}X_i(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma)\right)^T A_iX_i(X_0, B_j) + MX_i(X_0, B_j)^T A_i\left(\frac{d}{dt}X_i(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma)\right)\right)$$
$$\cdot\left(\prod_{\ell=i+1}^{k}\left(I + MX_\ell(X_0, B_j)^T A_\ell X_\ell(X_0, B_\ell)\right)\right)$$
$$\overset{(iii)}{=}\frac{d}{dt}G(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma)$$

In (i) above, we the following facts: 1. $X_i(U_\Sigma X_0, B_j) = U_\Sigma X_i(X_0, B_j)$ from (37), 2. $A_i = a_i\Sigma^{-1}$, so that $U_\Sigma^\top A_iU_\Sigma = A_i$, 3. $U_\Sigma U_\Sigma^{-1} = U_\Sigma^{-1}U_\Sigma = I$. (ii) follows from (38). (iii) is by definition of G.

Step 4: Putting everything together

Let us now continue from (36). We can now plug (42) and (43) into (36):

$$\begin{aligned}
&\frac{d}{dt}f(A, B(tS_j, j))\Big|_{t=0}\\
=&2\mathbf{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G(U_\Sigma X_0, B_j)^\top\Sigma^{-1}\ \frac{d}{dt}G(U_\Sigma X_0, B_j + tS_j)\Big|_{t=0}(I-M)\right)\right]\\
\overset{(i)}{=}&2\mathbf{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G(X_0, B_j)^\top\Sigma^{-1}\ \frac{d}{dt}G(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma)\Big|_{t=0}(I-M)\right)\right]\\
=&2\mathbf{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0, B_j)^\top\Sigma^{-1}\mathbf{E}_U\left[\frac{d}{dt}G(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma)\Big|_{t=0}\right](I-M)\right)\right]\\
\overset{(ii)}{=}&2\mathbf{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0, B_j)^\top\Sigma^{-1}\ \frac{d}{dt}G(X_0, B_j + t\mathbf{E}_U\left[U_\Sigma^{-1}S_jU_\Sigma\right])\Big|_{t=0}(I-M)\right)\right]\\
=&2\mathbf{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0, B_j)^\top\Sigma^{-1}\ \frac{d}{dt}G(X_0, B_j + ts_jI)\Big|_{t=0}(I-M)\right)\right]\\
=&\frac{d}{dt}f(A, B(ts_jI, j))\Big|_{t=0}
\end{aligned}$$

where $s_j := \frac{1}{d}\mathrm{Tr}\left(\Sigma^{-1/2}S_j\Sigma^{1/2}\right)$. In the above, (i) uses 1. (42) and (43), as well as the fact that $U_\Sigma^\top\Sigma^{-1}U_\Sigma = \Sigma^{-1}$. (ii) uses the fact that $\frac{d}{dt}G(X_0, B_j + tC)\big|_{t=0}$ is affine in C. To see this, one can verify from (39), using a simple induction argument, that $\frac{d}{dt}X_i(X_0, B_j + tC)$ is affine in C for all i. We can then verify from the definition of G, e.g. using similar algebra as the proof of (43), that $\frac{d}{dt}G(X_0, B_j + C)$ is affine in $\frac{d}{dt}X_i(X_0, B_j + tC)$. Thus $\mathbf{E}_U\left[G(X_0, B_j + tU_\Sigma^{-1}S_jU_\Sigma)\right] = G(X_0, B_j + t\mathbf{E}_U\left[U_\Sigma^{-1}S_jU_\Sigma)\right]$.

With this, we conclude our proof of (35).

Step 5: Proof of (34)

We will now prove (34) for fixed but arbitrary j, i.e. there is some r_j such that

$$\frac{d}{dt}f(A(t\cdot r_j\Sigma^{-1}, j), B) \le \frac{d}{dt}f(A(tR_j, j), B).$$

The proof is very similar to the proof of (35) that we just saw, and we will essentially repeat the same steps from Step 2-4 above.

Let us introduce a redefinition: let $X_i(X, C)$ (resp $Y_i(X, C)$) to denote the value of X_i (resp Y_i) from (29), with $X_0 = X$, and $A_j = C$ (previously it was with $B_j = C$). Once again, let $G(X, C) := X\prod_{i=0}^{i}\left(I + MX_i(X, C)^T\bar{A}_iX_i(X, C)\right)$, where $\bar{A}_j = A_j + tC$, and $\bar{A}_\ell = A_\ell$ for all $\ell \in \{0...k\}\setminus\{j\}$.

We first verify that

$$\begin{aligned}
X_i(U_\Sigma X_0, B_j) &= U_\Sigma X_i(X_0, B_j)\\
G(U_\Sigma X_0, B_j) &= U_\Sigma G(X_0, B_j).
\end{aligned} \tag{44}$$

The proofs are identical to the proofs of (37) and (42) so we omit them. Next, we show that for all i,

$$U_\Sigma^{-1} \frac{d}{dt} X_i(U_\Sigma X_0, A_j + tR_j)\bigg|_{t=0} = \frac{d}{dt} X_i(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma)\bigg|_{t=0}. \tag{45}$$

We establish the dynamics for the right-hand-side of (45):

$$\begin{aligned}
&\frac{d}{dt} X_\ell \left(X_0, A_j + tC\right) = 0\\
&\frac{d}{dt} X_{j+1} \left(X_0, A_j + tC\right) = B_j X_j \left(X_0, A_j\right) M X_j \left(X_0, A_j\right)^\top C X_j \left(X_0, A_j\right)\\
&\frac{d}{dt} X_{i+1} \left(X_0, A_j + tC\right) = \frac{d}{dt} X_i \left(X_0, A_j + tC\right)\\
&\qquad\qquad + B_i \left(\frac{d}{dt} X_i \left(X_0, A_j + tC\right)\right) M X_i \left(X_0, A_j\right)^\top A_i X_i \left(X_0, A_j\right)\\
&\qquad\qquad + B_i X_i \left(X_0, A_j\right) M \left(\frac{d}{dt} X_i \left(X_0, A_j + tC\right)\right)^\top A_i X_i \left(X_0, A_j\right)\\
&\qquad\qquad + B_i X_i \left(X_0, A_j\right) M X_i \left(X_0, A_j\right)^\top A_i \left(\frac{d}{dt} X_i \left(X_0, A_j + tC\right)\right)
\end{aligned} \tag{46}$$

Similar to (40), we show that for $i \leq j$,

$$\begin{aligned}
U_\Sigma^{-1} \frac{d}{dt} X_i \left(U_\Sigma X_0, A_j + tR_j\right) =& 0 = U_\Sigma^{-1} \frac{d}{dt} X_i \left(U_\Sigma X_0, A_j + tU_\Sigma R_j U_\Sigma\right)\\
U_\Sigma^{-1} \frac{d}{dt} X_{j+1} \left(U_\Sigma X_0, A_j + tR_j\right) =& U_\Sigma^{-1} B_j U_\Sigma X_j(X_0, A_j) M X_j(U_\Sigma X_0, A_j)^\top A_j\\
=& \frac{d}{dt} X_{j+1} \left(U_\Sigma X_0, A_j + tU_\Sigma^\top R_j U_\Sigma X_j\right).
\end{aligned}$$

Finally, for the inductive step, we follow identical steps leading up to (41) to show that

$$\begin{aligned}
&U_\Sigma^{-1} \frac{d}{dt} X_{i+1} \left(U_\Sigma X_0, A_j + tR_j\right)\\
=& \frac{d}{dt} X_i \left(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma\right)\\
&\quad + B_i \left(\frac{d}{dt} X_i \left(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma\right)\right) M X_i \left(X_0, A_j\right)^\top A_i X_i \left(X_0, A_j\right)\\
&\quad + B_i X_i \left(X_0, A_j\right) M \left(\frac{d}{dt} X_i \left(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma\right)\right)^\top A_i X_i \left(X_0, A_j\right)\\
&\quad + B_i X_i \left(X_0, A_j\right) M X_i \left(X_0, A_j\right)^\top A_i \left(\frac{d}{dt} X_i \left(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma\right)\right)
\end{aligned} \tag{47}$$

The inductive proof is complete by verifying that (47) exactly matches the third equation of (46) when $C = U_\Sigma^{-1} S U_\Sigma$. This concludes the proof of (45).

Next, we study the time derivative of $G(U_\Sigma X_0, A_j + tR_j)$ and show that

$$U_\Sigma^{-1} \frac{d}{dt} G(U_\Sigma X_0, A_j + tR_j) = \frac{d}{dt} G(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma). \tag{48}$$

This proof differs significantly from that of (43) in a few places, so we provide the whole derivation below. By chain-rule, we can write

$$U_\Sigma^{-1} \frac{d}{dt} G(U_\Sigma X_0, A_j + tR_j) = \spadesuit + \heartsuit$$

where

$$\spadesuit := U_\Sigma^{-1} \frac{d}{dt} \left(U_\Sigma X_0 \prod_{i=0}^{k} \left(I + M X_i(U_\Sigma X_0, A_j + tR_j)^T A_i X_i(U_\Sigma X_0, A_j + tR_j)\right)\right)$$

and

$$\heartsuit := U_\Sigma^{-1} U_\Sigma X_0 \left(\prod_{i=0}^{j-1} \left(I + M X_i(U_\Sigma X_0, A_j)^T A_i X_i(U_\Sigma X_0, A_j)\right) \right) \cdot M X_j(U_\Sigma X_0, A_j)^T R_j X_j(U_\Sigma X_0, A_j) \cdot \left(\prod_{i=j+1}^{k} \left(I + M X_i(U_\Sigma X_0, A_j)^T A_i X_i(U_\Sigma X_0, A_j)\right) \right).$$

We will separately simplify ♠ and ♡, and verify at the end that summing them recovers the right-hand-side of (48). We begin with ♠, and the steps are almost identical to the proof of (43).

$$\begin{aligned}
&\spadesuit \\
=& U_\Sigma^{-1} \frac{d}{dt} \left(U_\Sigma X_0 \prod_{i=0}^{k} \left(I + M X_i(U_\Sigma X_0, A_j + tR_j)^T A_i X_i(U_\Sigma X_0, A_j + tR_j)\right) \right) \\
=& X_0 \sum_{i=0}^{k} \left(\prod_{\ell=0}^{i-1} \left(I + M X_\ell(U_\Sigma X_0, A_j)^T A_\ell X_i(U_\Sigma X_0, A_\ell)\right) \right) \\
&\cdot M \frac{d}{dt} \left(X_i(U_\Sigma X_0, A_j + tR_j)^T A_i X_i(U_\Sigma X_0, A_j + tR_j)\right) \\
&\cdot \left(\prod_{\ell=i+1}^{k} \left(I + M X_\ell(U_\Sigma X_0, A_j)^T A_\ell X_i(U_\Sigma X_0, A_\ell)\right) \right) \\
\overset{(i)}{=}& X_0 \sum_{i=0}^{k} \left(\prod_{\ell=0}^{i-1} \left(I + M X_\ell(X_0, A_j)^T A_\ell X_\ell(X_0, A_\ell)\right) \right) \\
&\cdot M \left(\left(U_\Sigma^{-1} \frac{d}{dt} X_i(U_\Sigma X_0, A_j + tR_j) \right)^T A_i X_i(X_0, A_j) + M X_i(X_0, A_j)^T A_i \left(U_\Sigma^{-1} \frac{d}{dt} X_i(U_\Sigma X_0, A_j + tR_j) \right) \right) \\
&\cdot \left(\prod_{\ell=i+1}^{k} \left(I + M X_\ell(X_0, A_j)^T A_\ell X_\ell(X_0, A_\ell)\right) \right) \\
\overset{(ii)}{=}& X_0 \sum_{i=0}^{k} \left(\prod_{\ell=0}^{i-1} \left(I + M X_\ell(X_0, A_j)^T A_\ell X_\ell(X_0, A_\ell)\right) \right) \\
&\cdot M \left(\left(\frac{d}{dt} X_i(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma) \right)^T A_i X_i(X_0, A_j) + M X_i(X_0, A_j)^T A_i \left(\frac{d}{dt} X_i(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma) \right) \right) \\
&\cdot \left(\prod_{\ell=i+1}^{k} \left(I + M X_\ell(X_0, A_j)^T A_\ell X_\ell(X_0, A_\ell)\right) \right) \\
=& X_0 \sum_{i=0}^{k} \left(\prod_{\ell=0}^{i-1} \left(I + M X_\ell(X_0, A_j)^T A_\ell X_\ell(X_0, A_\ell)\right) \right) \\
&\cdot M \frac{d}{dt} \left(X_i(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma)^T A_i X_i(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma)\right) \\
&\cdot \left(\prod_{\ell=i+1}^{k} \left(I + M X_\ell(X_0, A_j)^T A_\ell X_\ell(X_0, A_\ell)\right) \right)
\end{aligned} \tag{49}$$

In (i) above, we the following facts: 1. $X_i(U_\Sigma X_0, B_j) = U_\Sigma X_i(X_0, B_j)$ from (44), 2. $A_i = a_i \Sigma^{-1}$, so that $U_\Sigma^\top A_i U_\Sigma = A_i$, 3. $U_\Sigma U_\Sigma^{-1} = U_\Sigma^{-1} U_\Sigma = I$. (ii) follows from (45).

We will now simplify $\heartsuit$.

$$\begin{aligned}
&\heartsuit\\
=&U_\Sigma^{-1}U_\Sigma X_0 \left(\prod_{i=0}^{j-1}\left(I + MX_i(U_\Sigma X_0, A_j)^T A_i X_i(U_\Sigma X_0, A_j)\right)\right)\\
&\cdot MX_j(U_\Sigma X_0, A_j)^T R_j X_j(U_\Sigma X_0, A_j)\\
&\cdot \left(\prod_{i=j+1}^{k}\left(I + MX_i(U_\Sigma X_0, A_j)^T A_i X_i(U_\Sigma X_0, A_j)\right)\right)\\
\overset{(i)}{=}&X_0 \left(\prod_{i=0}^{j-1}\left(I + MX_i(X_0, A_j)^T A_i X_i(X_0, A_j)\right)\right) MX_j(X_0, A_j)^\top U_\Sigma^\top R_j U_\Sigma X_j(X_0, A_j)\\
&\cdot \left(\prod_{i=j+1}^{k}\left(I + MX_i(X_0, A_j)^T A_i X_i(X_0, A_j)\right)\right), \qquad (50)
\end{aligned}$$

where (i) uses the fact that $X_i(U_\Sigma X_0, B_j) = U_\Sigma X_i(X_0, B_j)$ from (44) and the fact that $A_i = a_i\Sigma^{-1}$.

By expanding $\frac{d}{dt}G(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma)$, we verify that

$$\frac{d}{dt}G(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma) = (49) + (50) = \spadesuit + \heartsuit = U_\Sigma^{-1}\frac{d}{dt}G(U_\Sigma X_0, A_j + tR_j),$$

this concludes the proof of (48).

The remainder of the proof is similar to what was done in (36) in Step 4:

$$\begin{aligned}
&\frac{d}{dt}f(A(tR_j, j), B\Big|_{t=0}\\
=&2\mathbf{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G(U_\Sigma X_0, A_j)^\top\Sigma^{-1}\,\frac{d}{dt}G(U_\Sigma X_0, A_j + tR_j)\Big|_{t=0}(I-M)\right)\right]\\
\overset{(i)}{=}&2\mathbf{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G(X_0, A_j)^\top\Sigma^{-1}\,\frac{d}{dt}G(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma)\Big|_{t=0}(I-M)\right)\right]\\
\overset{(ii)}{=}&2\mathbf{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0, A_j)^\top\Sigma^{-1}\,\frac{d}{dt}G(X_0, A_j + t\mathbf{E}_U\left[U_\Sigma^\top R_j U_\Sigma\right])\Big|_{t=0}(I-M)\right)\right]\\
=&2\mathbf{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0, A_j)^\top\Sigma^{-1}\,\frac{d}{dt}G(X_0, A_j + t\cdot r_j\Sigma^{-1})\Big|_{t=0}(I-M)\right)\right]\\
=&\frac{d}{dt}f(A(t\cdot r_j\Sigma^{-1}, j), B)\Big|_{t=0},
\end{aligned}$$

where $r_j := \frac{1}{d}\mathrm{Tr}\left(\Sigma^{1/2}R_j\Sigma^{1/2}\right)$. In the above, (i) uses 1. (44) and (48), as well as the fact that $U_\Sigma^\top\Sigma^{-1}U_\Sigma = \Sigma^{-1}$. (ii) uses the fact that $\frac{d}{dt}G(X_0, A_j + tC)\big|_{t=0}$ is affine in C. To see this, one can verify using a simple induction argument, that $\frac{d}{dt}X_i(X_0, A_j + tC)$ is affine in C for all i. We can then verify from the definition of G, e.g. using similar algebra as the proof of (48), that $\frac{d}{dt}G(X_0, A_j + C)$ is affine in $\frac{d}{dt}X_i(X_0, A_j + tC)$ and C. Thus $\mathbf{E}_U\left[G(X_0, A_j + tU_\Sigma^\top R_j U_\Sigma)\right] = G(X_0, A_j + t\mathbf{E}_U\left[U_\Sigma^\top R_j U_\Sigma\right])$.

This concludes the proof of (34), and hence of the whole theorem.

B.4 Equivalence under permutation

Lemma 7. *Consider the same setup as Theorem 4. Let* $A = \{A_i\}_{i=0}^k$, *with* $A_i = a_i\Sigma^{-1}$. *Let* $f(A) := f\left(\left\{Q_i = \begin{bmatrix} A_i & 0\\ 0 & 0\end{bmatrix}, P_i = \begin{bmatrix} 0_{d\times d} & 0\\ 0 & 1\end{bmatrix}\right\}_{i=0}^k\right)$. *Let* $i, j \in \{0...k\}$ *be any two arbitrary indices, and let* $\tilde{A}_i = A_j$, $\tilde{A}_j = A_i$, *and let* $\tilde{A}_\ell = A_\ell$ *for all* $\ell \in \{0...k\}\setminus\{i,j\}$. *Then* $f(A) = f(\tilde{A})$

Proof. Following the same setup leading up to (21) in the proof of Theorem 4, we verify that the in-context loss is

$$f(A) = \mathbf{E}\left[\mathrm{Tr}\left((I-M)\, G(X_0,A)^\top \Sigma^{-1} G(X_0,A)\,(I-M)\right)\right]$$

where $G(X_0,A) := X_0 \prod_{\ell=0}^{k}\left(I + MX_0^T A_\ell X_0\right)$.

Consider any fixed index ℓ. We will show that

$$\left(I + MX_0^T A_\ell X_0\right)\left(I + MX_0^T A_{\ell+1} X_0\right) = \left(I + MX_0^T A_{\ell+1} X_0\right)\left(I + MX_0^T A_\ell X_0\right).$$

The lemma can then be proven by repeatedly applying the above, so that indices of A_i and A_j are swapped.

To prove the above equality,

$$\begin{aligned}
&\left(I + MX_0^T A_\ell X_0\right)\left(I + MX_0^T A_{\ell+1} X_0\right)\\
=&I + MX_0^T A_\ell X_0 + MX_0^T A_{\ell+1} X_0 + MX_0^T A_\ell X_0 M X_0^T A_{\ell+1} X_0\\
=&I + MX_0^T A_\ell X_0 + MX_0^T A_{\ell+1} X_0 + MX_0^T a_\ell \Sigma^{-1} X_0 M X_0^T a_{\ell+1}\Sigma^{-1} X_0\\
=&I + MX_0^T A_\ell X_0 + MX_0^T A_{\ell+1} X_0 + MX_0^T a_{\ell+1} \Sigma^{-1} X_0 M X_0^T a_{\ell}\Sigma^{-1} X_0\\
=&\left(I + MX_0^T A_{\ell+1} X_0\right)\left(I + MX_0^T A_\ell X_0\right).
\end{aligned}$$

This concludes the proof. Notice that we crucially used the fact that A_ℓ and $A_{\ell+1}$ are the same matrix up to scaling. □

C Auxiliary Lemmas

C.1 Reformulating the in-context loss

In this section, we will develop a re-formulation in-context loss, defined in (5), in a more convenient form (see Lemma 9).

For the entirety of this section, we assume that the transformer parameters $\{P_i, Q_i\}_{i=0}^k$ are of the form defined in (11), which we reproduce below for ease of reference:

$$P_i = \begin{bmatrix} B_i & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_i = \begin{bmatrix} A_i & 0 \\ 0 & 0 \end{bmatrix}.$$

Recall the update dynamics in (4), which we reproduce below:

$$Z_{i+1} = Z_i + \frac{1}{n} P Z_i M Z_i^\top Q Z_i, \tag{51}$$

where M is a mask matrix given by $M := \begin{bmatrix} I_{n\times n} & 0 \\ 0 & 0 \end{bmatrix}$. Let $X_k \in \mathbb{R}^{d\times n+1}$ denote the first d rows of Z_k and let $Y_k \in \mathbb{R}^{1\times n+1}$ denote the $(d+1)^{th}$ (last) row of Z_k. Then the dynamics in (51) is equivalent to

$$\begin{aligned}
X_{i+1} &= X_i + \frac{1}{n} B_i X_i M X_i^T A_i X_i\\
Y_{i+1} &= Y_i + \frac{1}{n} Y_i M X_i^T A_i X_i.
\end{aligned} \tag{52}$$

We present below an equivalent form for the in-context loss from (5):

Lemma 9. *Let p_x and p_w denote distributions over $\mathbb{R}^d$. Let $x^{(1)} \ldots x^{(n+1)} \overset{iid}{\sim} p_x$ and $w_\star \sim p_w$. Let $Z_0 \in \mathbb{R}^{d+1\times n+1}$ is specifically defined in (1) as*

$$Z_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)}.$$

Let Z_k denote the output of the $(k-1)^{th}$ layer of the linear transformer (as defined in (51), initialized at Z_0). Let $f\left(\{P_i, Q_i\}_{i=0}^k\right)$ denote the in-context loss defined in (5), i.e.

$$f\left(\{P_i, Q_i\}_{i=0}^k\right) = \mathbf{E}_{(Z_0,w_\star)}\left[\left([Z_k]_{(d+1),(n+1)} + w_\star^\top x^{(n+1)}\right)^2\right]. \tag{53}$$

Let $\overline{Z}_0$ be defined as

$$\overline{Z}_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & y^{(n+1)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)},$$

where $y^{(n+1)} = \langle w_\star, x^{(n+1)}\rangle$. Let $\overline{Z}_k$ denote the output of the $(k-1)^{th}$ layer of the linear transformer (as defined in (51), initialized at $\overline{Z}_0$). Assume $\{P_i, Q_i\}_{i=0}^k$ be of the form in (11). Then the loss in (5) has the equivalent form

$$f\left(\{A_i, B_i\}_{i=0}^k\right) = f\left(\{P_i, Q_i\}_{i=0}^k\right) = \mathbf{E}_{(\overline{Z}_0,w_\star)}\left[\mathit{Tr}\left((I-M)\,\overline{Y}_k^\top \overline{Y}_k\,(I-M)\right)\right],$$

where $\bar{Y}_k \in \mathbb{R}^{1\times n+1}$ is the $(d+1)^{th}$ row of $\bar{Z}_k$.

Before proving Lemma 9, we first establish an intermediate result (Lemma 10 below). To facilitate discussion, let us define a function $F_X\left(\{A_i, B_i\}_{i=0}^k, X_0, Y_0\right)$ and $F_Y\left(\{A_i, B_i\}_{i=0}^k, X_0, Y_0\right)$ to be the outputs, after k layers of linear transformers respectively. I.e.

$$F_X\left(\{A_i, B_i\}_{i=0}^k, X_0, Y_0\right) = X_{k+1}$$
$$F_Y\left(\{A_i, B_i\}_{i=0}^k, X_0, Y_0\right) = Y_{k+1},$$

as defined in (52), given initialization X_0, Y_0.

We now prove a useful lemma showing that $[Y_0]_{n+1} = y^{(n+1)}$ influences X_i, Y_i in a very simple manner:

Lemma 10. *Let X_i, Y_i follow the dynamics in (52). Then*

1. *$[X_i]$ is are independent of $[Y_0]_{n+1}$.*
2. *For $j \neq n+1$, $[Y_i]_j$ is independent of $[Y_0]_{n+1}$.*
3. *$[Y_i]_{n+1}$ depends additively on $[Y_0]_{n+1}$.*

In other words, for $C := [0, 0, 0..., 0, c] \in \mathbb{R}^{d+1\times 1}$,

$$1: F_X\left(\{A_i, B_i\}_{i=0}^k, X_0, Y_0 + C\right) = F_X\left(\{A_i, B_i\}_{i=0}^k, X_0, Y_0\right)$$
$$2+3: F_Y\left(\{A_i, B_i\}_{i=0}^k, X_0, Y_0 + C\right) = F_Y\left(\{A_i, B_i\}_{i=0}^k, X_0, Y_0\right) + C$$

Proof of Lemma 10. The first and second items follows directly from observing that the dynamics for X_i and Y_i in (52) do not involve $[Y_i]_{n+1}$, due to the effect of M.

The third item again uses the fact that $Y_{i+1} - Y_i$ does not depend on $[Y_i]_{n+1}$. □

We are now ready to prove Lemma 9

Proof of Lemma 9. Let Z_0, Z_k, $\overline{Z}_0$, $\overline{Z}_k$ be as defined in the lemma statement. Let $\overline{X}_k$ and $\overline{Y}_k$ denote first d rows and last row of $\overline{Z}_k$. Then by Lemma 10, $\overline{X}_k = X_k$ and $\overline{Y}_k = Y_k + \begin{bmatrix} 0 & 0 & \cdots & 0 & \langle w_\star, x^{(n+1)}\rangle \end{bmatrix}$. Therefore, (53) is equivalent to

$$\begin{aligned}
&\mathbf{E}_{(\overline{Z}_0,w_\star)}\left[\left([\overline{Z}_k]_{(d+1),(n+1)}\right)^2\right] \\
=&\mathbf{E}_{(\overline{Z}_0,w_\star)}\left[\left([\overline{Y}_k]_{(n+1)}\right)^2\right] \\
=&\mathbf{E}_{(\overline{Z}_0,w_\star)}\left[\left\|(I-M)\,\overline{Y}_k^\top\right\|^2\right] \\
=&\mathbf{E}_{(\overline{Z}_0,w_\star)}\left[\mathrm{Tr}\left((I-M)\,\overline{Y}_k^\top \overline{Y}_k\,(I-M)\right)\right].
\end{aligned}$$

This concludes the proof. □

C.2 Proof of Lemma 2 (Equivalence to Preconditioned Gradient Descent)

Proof of Lemma 2. Consider fixed samples $x^{(1)}...x^{(n)}$, and fixed $w_\star$. Let $P = \{P_i\}_{i=0}^k, Q = \{Q_i\}_{i=0}^k$ denote fixed weights. Let Z_i evolve as described in (4). Let X_i denote the first d rows of Z_k (under (9), $X_i = X_0$ for all I) and let Y_i denote the $(d+1)^{th}$ row of Z_i. Let $g(x,y,k) : \mathbb{R}^d \times \mathbb{R} \times \mathbb{Z} \to \mathbb{R}$ be a function defined as follows: let $x^{n+1} = x$ and let $y_0^{n+1} = y$, then $g(x,y,k) := y_k^{n+1}$. Note that $y_k^{n+1} = [Y_k]_{n+1}$.

We verify that, under (9), the formula for updating $y_k^{(n+1)}$ is given by

$$Y_{k+1} = Y_k - \frac{1}{n} Y_k M X_0^\top A_k X_0.$$

where M is a mask given by $\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$. We can verify the following facts

1. $g(x,y,k) = g(x,0,k) + y$. To see this, notice first that for all $i \in \{1...n\}$,

$$y_{k+1}^{(i)} = y_k^{(i)} - \frac{1}{n} \sum_{j=1}^n {x^{(i)}}^T A_k x^{(j)} y_k^{(j)}.$$

In other words, $y_k^{(i)}$ does not depend on $y_t^{(n+1)}$ for any t. Next, for $y_k^{(n+1)}$ itself,

$$y_{k+1}^{(n+1)} = y_k^{(n+1)} - \frac{1}{n} \sum_{j=1}^n {x^{(n+1)}}^T A_k x^{(j)} y_k^{(j)},$$

which depends on y_k^{n+1} only additively. We can verify under a simple induction that $g(x,y,k+1) - y = g(x,y,k) - y$.

2. $g(x,0,k)$ is linear in x. To see this, notice first that for $j \neq n+1$, $y_k^{(j)}$ is does not depend on $x_t^{(n+1)}$ for all t,j,k. Consequently, the update formula for $y_{k+1}^{(n+1)}$ depends only linearly on $x^{(n+1)}$ and $y_k^{(n+1)}$. Finally, $y_0^{(n+1)} = 0$ is linear in x, so the conclusion follows by induction.

With these two facts in mind, we verify that for each k, there exists a $\theta_k \in \mathbb{R}^d$, such that

$$g(x,y,k) = g(x,0,k) + y = \langle \theta_k, x \rangle + y$$

for all x, y. It follows from definition that $g(x,y,0) = y$, so that $\langle \theta_0, x \rangle = g(x,y,0) - y = 0$, so that $\theta_0 = 0$.

We now turn our attention to the third crucial fact: for all i,

$$g(x^{(i)}, y^{(i)}, k) = y_k^{(i)} = \left\langle \theta_k, x^{(i)} \right\rangle + y^{(i)}$$

To see this, suppose that we let $x^{(n+1)} := x^{(i)}$ for some $i \in 1...n$. Then

$$y_{k+1}^{(i)} = y_k^{(i)} - \frac{1}{n} \sum_{j=1}^n {x^{(i)}}^T A_k x^{(j)} y_k^{(j)}$$

$$y_{k+1}^{(n+1)} = y_k^{(n+1)} - \frac{1}{n} \sum_{j=1}^n {x^{(n+1)}}^T A_k x^{(j)} y_k^{(j)},$$

thus $y_{k+1}^{(i)} = y_{k+1}^{(n+1)}$ if $y_k^{(i)} = y_k^{(n+1)}$, and the induction proof is completed by noting that $y_0^{(i)} = y_0^{(n+1)}$ by definition. Let $\bar{X} \in R^{d \times n}$ be the matrix whose columns are $x^{(1)}...x^{(n)}$, leaving out $x^{(n+1)}$. Let $\bar{Y}_k \in \mathbb{R}^{1 \times n}$ denote the vector of $y_k^{(1)}...y_k^{(n)}$. Then it follows that

$$\bar{Y}_k = \bar{Y}_0 + \theta_k^T \bar{X}.$$

Using the above fact, the update formula for $y_k^{(n+1)}$ can be written as

$$
\begin{aligned}
y_{k+1}^{(n+1)} =& y_k^{(n+1)} - \frac{1}{n}\left\langle A_k X^\top Y_k, x^{(n+1)} \right\rangle \\
\Rightarrow \qquad \left\langle \theta_{k+1}, x^{(n+1)} \right\rangle =& \left\langle \theta_k, x^{(n+1)} \right\rangle - \frac{1}{n}\left\langle A_k \bar{X}\left(\bar{X}^T\theta_k + \bar{Y}_0\right), x^{(n+1)} \right\rangle \\
=& \left\langle \theta_k, x^{(n+1)} \right\rangle - \frac{1}{n}\left\langle A_k \bar{X}\left(\bar{X}^T\left(\theta_k + w_\star\right)\right), x^{(n+1)} \right\rangle
\end{aligned}
$$

Since the choice of $x^{(n+1)}$ is arbitrary, we get the more general update formula

$$
\theta_{k+1} = \theta_k - \frac{1}{n} A_k \bar{X}\bar{X}^T\left(\theta_k + w_\star\right).
$$

We can treat A_k as a preconditioner. Let $f(\theta) :== \frac{1}{2n}\left(\theta + w_\star\right)^T \bar{X}\bar{X}^T(\theta + w_\star)$, then

$$
\theta_{k+1} = \theta_k - \frac{1}{n} A_k \nabla f(\theta).
$$

Finally, let $w_k^{\mathsf{gd}} := -\theta_k$. We verify that $f(-w) = R_{w_\star}(w)$, so that

$$
w_{k+1}^{\mathsf{gd}} = w_k^{\mathsf{gd}} - \frac{1}{n} A_k \nabla R_{w_\star}(w_k^{\mathsf{gd}}).
$$

We also verify that for any $x^{(n+1)}$, the prediction of $y_k^{(n+1)}$ is

$$
g\left(x^{(n+1)}, y^{(n+1)}, k\right) = y^{(n+1)} - \left\langle \theta, x^{(n+1)} \right\rangle = y^{(n+1)} + \left\langle w_k^{\mathsf{gd}}, x^{(n+1)} \right\rangle.
$$

This concludes the proof. $\square$