Omitted Tehnical Proofs for "Exploring the Security Boundary of Data Reconstruction via Neuron Exclusivity Analysis"

Xudong Pan, Mi Zhang, Yifan Yan, Jiaming Zhu, Min Yang
Fudan University, China
{xdpan18, mi_zhang, yanyf20, 19210240146, m_yang}@fudan.edu.cn

I Proof for Proposition 2

For convenience, we denote the j-th element of D_i^m as $\alpha_{i,j}^m$, i.e., the activation state of the j-th neuron at the i-th layer when X_m is the input. Formally, the exclusive activation of a neuron is expressed as: $\alpha_{i,j}^m$ takes the value 1 for and only for a certain sample X_m. For intuition, readers may refer to Fig. 2 as an illustrative example.

- *Initial Step*: As a by-product of solving $\overline{g}_c^m$ and the assumed exclusivity, we already recovered at least two exclusive elements in D_H^m for each input X_m.
- *Recurrent Step*: Next, we consider the gradient equation w.r.t. W_{H-1}.

$$\overline{G}_{H-1} = \frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{K}\overline{g}_c(D_{H-1}^m ... W_0 X_m)([W_H]_c^T D_H^m) \tag{1}$$

Then, we expand it explicitly to individual scalar equations.

$$\begin{aligned} M[\overline{G}_{H-1}]_{ij} &= \sum_{m=1}^{M}\sum_{c=1}^{K}\overline{g}_c \alpha_{H-1,i}^m f_{H-2,i}^m [W_H]_{jc}\alpha_{H,j}^m \\ &:= \sum_{m=1}^{M} C_{ij}^m \alpha_{H-1,i}^m \alpha_{H,j}^m \end{aligned} \tag{2}$$

In the last line, we use the C_{ij}^m to replace the multiplier (which is non-zero almost surely in our threat model). The following is the key of the recurrent step. As $\{D_H^m\}_{m=1}^M$ have at least one exclusive nonzero position to each other, the terms in the summation above therefore have at most one non-vanishing term for this ExAN, indexed by e.g., j, which can be found based on the knowledge of $\{D_H^m\}_{m=1}^M$. In fact, the j-th column of $\overline{G}_{H-1}$, i.e., $[C_{ij}^m \alpha_{H-1,i}^m]$, immediately gives the diagonal terms of D_{H-1}^m, if we simply check the non-zero positions of $[\overline{G}_{H-1}]_{:,j}$. Similarly, with the solved $\{D_{H-1}^m\}_{m=1}^M$, the procedure can be done for the $(H-2)$-th layer, and so on, until the input layer.

II Proof for Theorem 1

This case corresponds to the situation when the gradient equation system is under-determined, i.e., the number of equations is smaller than the number of variables. We denote the total derivative operator $A := (\nabla_{W_0}\ell, \ldots, \nabla W_H \ell)$, where $\nabla_{W_i}\ell(X_1, \ldots, X_M) = \frac{1}{M}\sum_{m=1}^{M}\nabla_{W_i}\ell^m(X_m)$

(Here, $\ell^m(X_m)$ is defined similarly to the average loss while the accumulated activation patterns $(D_1, \ldots, D_H)$ are replaced by the m-th sample's own activation pattern $(D_1^m, \ldots, D_H^m)$). Therefore, the (under)-determined gradient equation writes $A(X_1, \ldots, X_M) = (G_0, \ldots, G_H) := b$, which has the ground-truth data inputs $X^* := (X_1^*, \ldots, X_M^*)$ as the least-square-error (LSE) solution. Then, we need to consider, when the attacker is only provided with an underdetermined equation system, i.e., $(A + \Delta A)X = b + \Delta b$, how the corresponding LSE solution $X := (X_1, \ldots, X_M)$ is perturbed. We introduce the following lemma. [Theorem 5.7.1[1]] Suppose $\text{rank}(A) = m \geq n$ and that $A \in \mathbb{R}^{m\times n}$, $\Delta A \in \mathbb{R}^{m\times n}$, $0 \neq b \in \mathbb{R}^m$, and $\Delta b \in \mathbb{R}^m$ satisfy $\epsilon = \max \epsilon_A, \epsilon_b < \lambda(A)$, where $\epsilon_A = \|\Delta A\|_2/\|A\|_2$ and $\epsilon_b = \|\Delta b\|_2/\|b\|_2$. If x and $\hat{x}$ are minimum norm solutions that satisfy $Ax = b$ and $(A + \Delta A)\hat{x} = b + \Delta b$, then

$$\frac{\|\hat{x} - x\|_2}{\|x\|_2} \leq \text{cond}(A)(\epsilon_A \min\{2, n - m + 1\} + \epsilon_b) + O(\epsilon^2), \tag{3}$$

When the perturbation $\delta := \|\Delta A\|_2/\|A\|_2 < \lambda(A)$ (i.e., the smallest singular value of A), we have $\|X - X^*\|_2/\|X^*\|_2 < 2\delta\text{cond}(A) < 2\sum_{i=0}^H \delta_i \text{cond}(\nabla_{W_i}\ell)$, where $\delta_i := \|\Delta A_i\|_2/\|\nabla_{W_i}\ell\|_2$ and ΔA_i is the perturbation added to the i-th layer. First, we consider the perturbation condition to estimate δ_i. For the i-th layer, the condition requires $\delta_i < \lambda(\nabla_{W_i}\ell)$. Considering the underdetermined equation system built by the attacker, the perturbation ΔA_i should cancel out the rows of $\nabla_{W_i}\ell$ where the gradient is not captured, i.e., the $\|\Delta A_i\|_2/\|\nabla_{W_i}\ell\|_2 = (1-\beta(\overline{G}_i))$ almost surely (where $\overline{G}_i$ is the gradient at the i-th layer captured by the attacker). Next, we apply the following lemma from [2] to estimate the singular value of A. [Theorem 3 [2]] For every $i = 0, \ldots, H$, with probability $\geq 1 - \exp -\Omega(\sqrt{d_i d_{i+1}}/\text{poly}(M, H, \epsilon_i^{'-1}))$, it satisfies , and every W_i with $\|W_i - W_i^{(0)}\|_2 \leq \frac{1}{\text{poly}(M,H,\epsilon_i^{'-1})}$,

$$\Omega(\frac{\sqrt{d_i d_{i+1}}\epsilon_i^{'}}{M \dim \mathcal{X}}) \leq \|A\|_F^2 \leq O(\frac{\sqrt{d_i d_{i+1}}M}{\dim \mathcal{X}}). \tag{4}$$

In other words, the smallest and the largest singular values of A are controlled by the two ends of the inequality above. Therefore, the requirement above is reduced to $\delta = (1 - \beta(\overline{G}_i)) < \lambda(\nabla_{W_i}\ell) = \epsilon_i^{'}\frac{\sqrt{d_i d_{i+1}}}{M \dim \mathcal{X}}$ almost surely.

Using the two estimates in the lemma above, we can further upper bound the conditional number $\text{cond}(\nabla_{W_i}\ell) := \Lambda(\nabla_{W_i}\ell)/\lambda(\nabla_{W_i}\ell) < O(\frac{M^2}{\epsilon_i^{'}})$, where $\Lambda(\cdot)$ is the largest singular value. Finally, by inserting the estimations of $\text{cond}(\nabla_{W_i}\ell)$ and δ_i into the original bound and replacing $M/\epsilon_i^{'}$ with a new constant ϵ_i, we have $\|X - X^*\|_2/\|X^*\|_2 < O(M\sum_{i=0}^H \epsilon_i(1 - \beta(\overline{G}_i)))$, if for all $i \in \{0, \ldots, H\}$, $1 - \beta(\overline{G}_i) < \epsilon_i\frac{\sqrt{d_i d_{i+1}}}{M \dim \mathcal{X}}$. Expanding and moving $\|X^*\|_2$ to RHS gives the final form in Theorem 1.

Proof for Theorem 2 and Corollary 1. We prove the impossibility of unique reconstruction by directly constructing the linear space $\mathcal{Q}$ where every translation $\Delta \in \mathcal{Q}$ satisfies Eq. (7) & (8). To construct the perturbation $\Delta \in \mathbb{R}^{d_0\times M}$, we only need to consider solve the following equation system. $\begin{cases} A\Delta^T = 0 \\ W_0\Delta = 0, \end{cases}$ where $A = [\alpha_1^T, \ldots, \alpha_M^T] \in \mathbb{R}^{d_1\times M}$ and $\alpha_m = \sum_{c=1}^K \overline{g}_c^m([W_H]_c^T D_H^m \ldots W_1 D_1^m)$. It is easy to see, for any Δ satisfying the second equation above, we always have $W_0(X_m + \Delta_m) = W_0 X_m$, which guarantees the gradients w.r.t. each $(b_i)_{i=0}^H$ and each $(W_i)_{i=1}^H$ to be invariant. Meanwhile, to satisfy the first equation guarantees the gradients w.r.t. W_0 to be invariant. In the following, we show the solution set of the equation system above itself is a linear space of dimension $M \times (d_0 - d_1)$.

First, we consider the equation $W_0\Delta = 0$. When $d_1 < d_0$, this equation has its solution written as $\Delta = (I - W_0^\dagger W_0)Q$, where $W_0^\dagger$ is the Moore-Penrose (MP) (pseudo-)inverse and Q is an arbitrary matrix in $\mathbb{R}^{d_0\times M}$. Denote the projection operator $P_0 := I - W_0^\dagger W_0$. Inserting the above equation into the first equation $A\Delta^T = 0$, we obtain the following constraint on $\tilde{Q}(:= Q^T)$: $A\tilde{Q}P_0^T = 0$. Next, we utilize the following results from [3]. [Theorem 2.13[3]] A necessary and sufficient condition for the matrix equation $AXB = C$ to have a solution is that $AA^\dagger CB^\dagger B = C$, in which case the general solution is $X = A^\dagger CB^\dagger + Q - A^\dagger AQBB^\dagger$. In our context, for the equation $A\tilde{Q}P_0^T = 0$, we set $C = 0$ in the above lemma, which states the equation always has infinitely many solutions written in $\tilde{Q} = Q - A^\dagger AQ(P_0^T P_0^{T\dagger})^T$, where Q is an arbitrary vector in $\mathbb{R}^{d_0\times m}$. Thus, we have $\Delta = P_0(Q - A^\dagger AQ(P_0^T P_0^{T\dagger}))^T$ for an arbirary $Q \in \mathbb{R}^{d_0\times m}$, which, as can be easily checked, forms a linear space $\mathcal{Q}$. Finally, as the projection operator P_0 projects the $\mathbb{R}^{m\times d_0}$ to a subspace of dimension $m \times (d_0 - d_1)$, we have $\dim \mathcal{Q} = m \times (d_0 - d_1)$.

Next, we show there exists a perturbation subspace $\mathcal{Q}$ such that for any $\Delta \in \mathcal{Q}$, the gradient equation becomes identical for X and $X + \Delta$, which in other words implies the impossibility of unique reconstruction from the gradient equation as the only information source. In this part, we further analyze the property of the perturbation subspace to answer how large such a perturbation can be. As a typical scenario, we estimate the upper bound of $\max \frac{1}{M}\sum_{i=1}^{M} \|\Delta_i\|_2^2$ where Δ satisfies the above equation system and respects the common box constraint on an image input, i.e., $X + \Delta \in [-1, 1]^{M\times d_0}$.

Denote the null space of W_0 as $W_0^\perp = \text{span}(e_1, \ldots, e_{d_0-d_1})$, where $(e_j)_{j=1}^{d_0-d_1}$ forms the orthogonal basis of $W_0^\perp$. Besides, we denote the remaining orthogonal basis as $\{e_{d_0-d_1+1}, \ldots, e_{d_0}\}$. We also denote the basis transformation matrix as $T = [e_1, \ldots, e_{d_0}]$. As $\Delta \in W_0^\perp$, we represent $\Delta_i = \sum_{j=1}^{d_0-d_1} \delta_{ij} e_j$. Also with the orthogonal basis of the null space, we reformulate the box constraint $X+\Delta \in [-1, 1]^{M\times d_0}$ as an inequality $-\mathbf{1}_{d_0} \preceq X_i+\Delta_i \preceq \mathbf{1}_{d_0}$ $(i = 1, \ldots, M)$. Applying the projection operator P_0 related with $W_0^\perp$ to both sides of the inequality, we have $-P_0\mathbf{1}_{d_0} \preceq P_0X_i + \Delta_i \preceq P_0\mathbf{1}_{d_0}$ (note $P_0\Delta_i = \Delta_i$), which gives $-|P_0|\mathbf{1}_{d_0} - P_0X_i \preceq \Delta_i \preceq |P_0|\mathbf{1}_{d_0} - P_0X_i$, where $|\cdot|$ denotes the elementwise absolute on the matrix. Similarly, applying the basis transformation matrix to the inequality, we have $-|T||P_0|\mathbf{1}_{d_0} - TP_0X_i \preceq T\Delta_i \preceq |T||P_0|\mathbf{1}_{d_0} - TP_0X_i$. The inequality is therefore transformed to another set of box constraints $\delta_{ij} \in [-a_{ij}, b_{ij}]$ $(i = 1, \ldots, M, j = 1, \ldots, d_0-d_1)$, where $a_{ij} := [|T||P_0|\mathbf{1}_{d_0}+TP_0X_i]_j$ and $b_{ij} := [|T||P_0|\mathbf{1}_{d_0}-TP_0X_i]_j$.

Then, our problem reduces to estimate the upper bound of $\sum_{i=1}^{M}(\sum_{j=1}^{d_0-d_1} \delta_{ij}^2)^{1/2}$, where (δ_{ij}) satisfy the interval constraints $\delta_{ij} \in [-a_{ij}, b_{ij}]$ and the first matrix equation $A\Delta^T = 0$. Inserting the orthogonal basis representation of Δ into the equation, we have $\sum_{i=1}^{M}\sum_{j=1}^{d_0-d_1} \delta_{ij}(\alpha_i \otimes e_j) = 0$, which can be reformulated as the following linear equation w.r.t. $\delta := (\delta_{ij})_{i=1,j=1}^{M,d_0-d_1}$:

$$(A \otimes E)\text{vec}(\delta) = 0 \tag{5}$$

where $A = [\alpha_1^T, \ldots, \alpha_M^T] \in \mathbb{R}^{d_1\times M}$ and $E = [e_1^T, \ldots, e_{d_0-d_1}^T] \in \mathbb{R}^{d_0\times(d_0-d_1)}$. As $\text{rank}(A \otimes E) = \text{rank}(A)\text{rank}(E) = d_1(d_0 - d_1) < M(d_0 - d_1)$, the linear vector equation above always have infinitely many non-trivial solutions. Denote the projection operator w.r.t. $A\otimes E$ as $P_1 = I - (A\otimes E)^\dagger A\otimes E = I-(A^\dagger A\otimes E^\dagger E) = I-(A^\dagger A\otimes I_{d_0-d_1})$ (as the matrix E formed by the orthogonal basis is of full column rank). With the above definition, the general solution of $A\Delta^T = 0$ is written as P_1q, where $q \in \mathbb{R}^{M(d_0-d_1)}$ satisfies the interval constraints $[-a_{ij}, b_{ij}]$. As the norm of the perturbation $\sum_{i=1}^{M}\sum_{j=1}^{d_0-d_1} \delta_{ij}^2$ is equal to $\|P_1q\|_2^2 = q^T P_1^T P_1q$, a quadratic function with the critical point at $q = 0$ with a positive curvature (as $P_1^T P_1 \succ 0$), we therefore assert that the maximum norm solution is taken at the boundary points of the interval constraints. Formally, it gives $\max_q \|P_1q\|_2^2 = \|P_1q^*\|_2^2$, where $(q^*)_{ij} = \max\{a_{ij}, b_{ij}\} = \max\{([|T||P_0|\mathbf{1}_{d_0} + TP_0X_i]_j, [|T||P_0|\mathbf{1}_{d_0} - TP_0X_i]_j\} \geq |TP_0X_i|_j$. Denote $\eta_i = |TP_0X_i|$ and therefore $q^* = \eta_1 \oplus \cdots \oplus \eta_M$. Denote $Y = [\eta_1^T, \ldots, \eta_1^T] \in \mathbb{R}^{(d_0-d_1)\times M}$. Finally, we have $\|P_1q^*\|_2^2 = q^{*T} P_1^T P_1q^* = q^{*T} P_1q^* = \|q^*\|_2^2 - q^{*T}(A^\dagger A \otimes I_{d_0-d_1})q^* = \|q^*\|_2^2 - \text{Tr}(A^\dagger AY^TY) = \sum_{i=1}^{M} \|\eta_i\|_2^2 - \text{Tr}(A^\dagger AY^TY)$, where the second equality comes from the fact that the projection operator P_1 is symmetric and idempotent. Corollary 1 is immediate as the removal of the first ReLU layer is equivalent to $D_1^m \equiv I_{d_1}$, for which Theorem 2 is then applicable.

References

[1] G. Golub and C. Loan. *Matrix computations (2nd ed.)*. The Johns Hopkins University Press, 1989.

[2] Zeyuan Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. *ICML*, 2019.

[3] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics (Revised Edition)*. John Wiley & Sons, 1999.