A PROOF OF LEMMA 3

Proof. To compute the pruning MSE for the noisy top-k method, we first calculate the pruning MSE for the original top-k vector I_0. As the top-k elements depend on the gradient distribution, we approximate the k-th percentile by $a\sigma_{\mathrm{g}}$ where a satisfies

$$2\Phi(a) - 1 = 1 - k \Longleftrightarrow a = \Phi^{-1}(1 - \frac{k}{2}). \tag{29}$$

And Φ represents the CDF for the normal distribution. Hence the pruning MSE for I_0 is:

$$\begin{aligned} MSE_{t_0} =& \mathbb{E}_{\mathrm{g}} \|\mathrm{g} - \mathrm{g} \odot I_0\|_2^2 = (d - k \cdot d)\sigma_{\mathrm{g}}^2 \int_{-a}^{a} \frac{x^2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\ =& \frac{d\sigma_{\mathrm{g}}^2(1-k)}{\sqrt{2\pi}} \left[-x \exp\left(-\frac{x^2}{2}\right)\Big|_{-a}^{a} + \int_{-a}^{a} \exp\left(-\frac{x^2}{2}\right) dx \right] \\ =& (1-k) d\sigma_{\mathrm{g}}^2 \left[1 - k - \sqrt{\frac{2}{\pi}} a \exp\left(-\frac{a^2}{2}\right) \right]. \end{aligned} \tag{30}$$

Hence Eq. (14) is proved.

To calculate the pruning MSE for the perturbed index vector I, we split I into two sets — $C_0(I)$ and $C_1(I)$, representing the set with index 0 and the set with index 1, respectively:

$$\begin{aligned} & C_0(I) = \{j | I_j = 0\},\ C_1(I) = \{j | I_j = 1\}, \\ & |C_0(I)| = |C_0(I_0)| = kd,\ |C_1(I)| = |C_1(I_0)| = (1-k)d, \end{aligned} \tag{31}$$

where $|\cdot|$ denotes the number of elements in the set. Therefore, the differed number of indices between $C_1(I)$ and $C_1(I_0)$ is equal to the differed number between $C_0(I)$ and $C_0(I_t)$, i.e.,

$$(1-k)d - |C_0(I) \cap C_0(I_0)| = kd - |C_1(I) \cap C_1(I_0)|. \tag{32}$$

We set the differed number of indices between $C_0(I)$ and $C_0(I_t)$ as i and its range is $0 \le i \le kd$. Thus, we denote MSE_t by i:

$$MSE_t = \sum_{i=0}^{k \cdot d} \|\mathrm{g} - \mathrm{g} \odot I\|_2^2 \frac{1}{\psi(\theta, \boldsymbol{d})} e^{-\theta 2i} \cdot |S_k(I_0, i)|. \tag{33}$$

The difference between I and I_0 can be regarded as randomly setting i elements from C_0 to be 1 and i elements from C_1 to be 0. Since each element in $C_0(I) \cap C_0(I_t)$ and $C_0(I) \cap C_1(I_t)$ has an equivalent chance as the rest elements to be selected, we have

$$\begin{aligned} & \mathbb{E}_{\mathrm{g}} \|\mathrm{g} - \mathrm{g} \odot I\|_2^2 = \mathbb{E}_{\mathrm{g}} \sum_{j \in C_0(I)} g_j^2 \\ =& \mathbb{E}_{\mathrm{g}} \sum_{j_1 \in C_0(I) \cap C_0(I_0)} g_{j_1}^2 + \sum_{j_2 \in C_0(I) \cap C_1(I_0)} g_{j_2}^2 \\ =& MSE_{t_0} \cdot \frac{(1-k)d - i}{(1-k)d} + (d\sigma_g^2 - MSE_{t_0}) \cdot \frac{i}{kd} \\ =& \sigma_g^2 [(1-k)d - i](1 - k\zeta) + \sigma_g^2 (1 + (1-k)\zeta) i \\ =& \sigma_g^2 [(1-k)(1-k\zeta)d + \zeta i] = MSE_{t_0} + \sigma_g^2 \zeta i, \end{aligned} \tag{34}$$

where $\zeta = \frac{2a}{k\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) + 1$. Therefore, Lemma 3 is proved. □

B PROOF OF THEOREM 3

Proof. By definition, the pruning MSE can be written as the summation over all S_k groups:

$$MSE_t = \sum_{I \in S_k} \mathbb{E}_{\mathrm{g}} \|\mathrm{g} - \mathrm{g} \odot I\|_2^2 \frac{1}{\psi(\theta, \boldsymbol{d})} e^{-\theta \boldsymbol{d}(I, I_0)}. \tag{35}$$

Each I in the same S_k has an equivalent probability to appear. And the number of elements in S_k is

$$|S_k(I_0, i)| = \binom{kd}{i} \binom{(1-k)d}{i}. \tag{36}$$

The difference between I and I_0 can be regarded as randomly turning i 0s in I_0 into 1s and i 1s in I_0 into 0s. According to the values of MSE at different is, we get

$$\begin{aligned} MSE_t =& \sum_{i=0}^{k \cdot d} \left[MSE_{t_0} + \sigma_g^2 \zeta i \right] \frac{1}{\psi(\theta, \boldsymbol{d})} e^{-\theta 2i} |S_k(I_0, i)| \\ =& MSE_{t_0} + \frac{\sigma_g^2}{\psi(\theta, \boldsymbol{d})} \left[1 + \frac{2a}{k\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) \right] \sum_{i=0}^{kd} \binom{kd}{i} \binom{(1-k)d}{i} i e^{-\theta 2i}. \end{aligned} \tag{37}$$

Thereby we have proved Thm. 3. □

C PROOF OF LEMMA 4

Proof. We verify the monotonicity of $F(\theta)$ by taking derivative over θ. We first analyze the derivative of $\psi(\theta, \boldsymbol{d})$:

$$\begin{aligned} \psi(\theta, \boldsymbol{d}) &= \sum_{i=0}^{kd} \binom{kd}{i} \binom{(1-k)d}{i} e^{-\theta 2i}, \\ \nabla_\theta \psi(\theta, \boldsymbol{d}) &= \sum_{i=0}^{kd} \binom{kd}{i} \binom{(1-k)d}{i} (-2i) e^{-\theta 2i}. \end{aligned} \tag{38}$$

For simplicity of presentation, we rewrite the equation by using $a_i = \binom{kd}{i}\binom{(1-k)d}{i} e^{-\theta 2i}$. Therefore, we calculate the derivative of $F(\theta)$ as:

$$\begin{aligned} \nabla_\theta F(\theta) &= \frac{1}{\psi^2(\theta, \boldsymbol{d})} \left[\sum_{i=0}^{kd} a_i(-2i^2) \sum_{i=0}^{kd} a_i - \sum_{i=0}^{kd} a_i(-2i) \sum_{i=0}^{kd} a_i i \right] \\ &= \frac{2}{\psi^2(\theta, \boldsymbol{d})} \left[(\sum_{i=0}^{kd} a_i i)^2 - \sum_{i=0}^{kd} i^2 a_i \sum_{i=0}^{kd} a_i \right]. \end{aligned} \tag{39}$$

By expanding the numerator part of the equation, we obtain

$$\begin{aligned} & (\sum_{i=0}^{kd} a_i i)^2 - \sum_{i=0}^{kd} i^2 a_i \sum_{i=0}^{kd} a_i \\ =& \sum_{i=0}^{kd} (a_i i)^2 + \sum_{0 \le i < j \le kd} 2 a_i a_j i j - \sum_{i=0}^{kd} (a_i i)^2 - \sum_{0 \le i < j \le kd} a_i a_j (i^2 + j^2) \\ =& \sum_{0 \le i < j \le kd} a_i a_j (2ij - i^2 - j^2). \end{aligned} \tag{40}$$

Since $2ij \le i^2 + j^2$, we have $\nabla_\theta F(\theta) \le 0$. Thus, $F(\theta)$ is monotonically decreasing with respect to θ when $\theta \ge 0$. □