Why Pruning and Conditional Computation Work: A High-Dimensional Perspective

Erdem Koyuncu EKOYUNCU@UIC.EDU
University of Illinois Chicago

Abstract

We analyze the processes of pruning and conditional computation for the case of a single neuron in the asymptotic learning regime of large input dimension and training set size. For this purpose, we introduce conditional neurons, which implement an early exit strategy at the neuron level. Specifically, a conditional neuron considers the local field induced by a subset of its inputs. If this sub-local field is strong enough, then the rest of the inputs are ignored, saving computation. Conditional neurons provide an archetype of the well-known early exit or conditional computation architectures. As such, we formally analyze their generalization performance to understand why conditional computation is so effective in preserving performance despite significantly reduced average amount of computation. In the process, we introduce a concentration theorem for one-shot neuron-wise pruning, which is recently popularized in the context of large language models.

1. Introduction

In the context of neural networks, conditional computation refers to the idea of adapting the network computations based on the inputs or intermediate features produced at different layers [3, 6, 8, 11]. A recent example is mixture of experts in transformers [13, 17, 19]. Of particular interest to this work is one of the simplest cases of conditional computation, which is commonly referred to as early exit networks [10, 14, 16, 22]. The concept of early exit involves the utilization of intermediate classifiers, which are located in non-terminal layers of a neural network. When an intermediate classifier exhibits sufficient confidence in its decisions, it can perform an "early exit," bypassing the the subsequent layers and thus conserving computational resources. Previous works have shown that early exit architectures can significantly improve upon ordinary neural networks in terms of the trade-off between computational complexity and accuracy. However, a theoretical justification as to why early exit networks perform so well has remained elusive, which is the goal of the present work. To the best of our knowledge, our work is the first to formally study a conditional computation scheme. Thus, our results shed light into the performance of other conditional architectures as well.

The rest of this paper is organized as follows: In Section 2, we introduce the conditional perceptron. We prove a concentration theorem for pruning in Section 3. We provide our main generalization bounds for the conditional perceptron in Section 4. The proofs are provided in the appendices.

2. Conditional Perceptrons

In this section, we begin by introducing conditional perceptrons, which implement the idea of early exit at the neuron level. Conditional perceptrons thus serve as one of the simplest special cases of a deep neural network with an early exit.

Consider inputs $x_1, \ldots, x_n$ to a neuron with corresponding weights $w_1, \ldots, w_n$, respectively. Let $\mathbf{x} = [x_1 \cdots x_n]^T$ and $\mathbf{w} = [w_1 \cdots w_n]^T$ be the corresponding input and weight vectors. The Heaviside step function can be defined as $\sigma(v) = 1$ for $v \geq 0$ and $\sigma(v) = -1$ for $v < 0$. We recall that an ordinary, "unconditional" perceptron provides the output $y_{\mathrm{uc}} \triangleq \sigma(v_0)$, $v_0 \triangleq \mathbf{w}^T\mathbf{x}$, where the subscript "uc" stands for unconditional. To construct the conditional perceptron, let us order the weights $w_1, \ldots, w_n$ from the smallest to the largest in magnitude as $|w_{i_1}| \leq \cdots \leq |w_{i_n}|$, where $i_1, \ldots, i_n$ is a permutation of $1, \ldots, n$. The largest $n-k$ weights in magnitude, where $k \in \{1, \ldots, n\}$ is a design parameter, are thus given by $w_{i_{k+1}}, \ldots, w_{i_n}$. Ignoring the inputs from the remaining weights, we consider the local field induced by a normalized version of these weights. Namely, we define $v_1 \triangleq \mathbf{w}_{\mathrm{ee}}^T\mathbf{x}$, where $\mathbf{w}_{\mathrm{ee}} \triangleq f_{\mathrm{ee}}^k(\mathbf{w}) \triangleq \frac{\|\mathbf{w}\|}{\|\mathbf{w}_{\mathrm{p}}\|}\mathbf{w}_{\mathrm{p}}$, and $\mathbf{w}_{\mathrm{p}} \triangleq f_{\mathrm{p}}^k(\mathbf{w}) \triangleq [(w_{\mathrm{p}})_1 \cdots (w_{\mathrm{p}})_n]^T$ with $(w_{\mathrm{p}})_{i_j} = 0,\ 1 \leq j \leq k$, and $(w_{\mathrm{p}})_{i_j} = w_{i_j},\ k+1 \leq j \leq n$. The subscript "ee" indicates that $\mathbf{w}_{ee}$ corresponds to the early exit weight vector of the conditional perceptron, and "p" indicates a pruned (but unnormalized) version of the weight vector. The normalization factor $\|\mathbf{w}\|/\|\mathbf{w}_{\mathrm{p}}\|$ ensures that both v_0 and v_1 have the same variance of $\|\mathbf{w}\|^2$ when the input $\mathbf{x}$ is considered random with zero mean and identity covariance.

The input-output relationship of the conditional perceptron is then finally expressed as

$$y_{\mathrm{c}} \triangleq \begin{cases} \sigma(v_1), & |v_1| \geq \tau, \\ \sigma(v_0), & \text{otherwise}, \end{cases} \tag{1}$$

where $\tau > 0$ is a threshold hyperparameter, and the subscript "c" stands for conditional. Ideally, the conditional perceptron wishes to achieve the same class decision $\sigma(v_0)$ as the ordinary perceptron. However, by definition, if the local field $|v_1|$ is large enough, then an "early exit" is performed with the class decision $\sigma(v_1)$. We expect that the ignored weights, since they are relatively small in magnitude, will be unable to sway this decision to the disagreement $\sigma(v_1) \neq \sigma(v_0)$. Thus, $\sigma(v_1) = \sigma(v_0)$ holds for most inputs, at least when k and τ are both chosen to be large enough.

The motivation behind the early exit mechanism is to preserve computational resources. To demonstrate this fact, let us calculate the total floating point operations (FLOPs) required to calculate the conditional perceptron output v_1. Since only $n-k$ of the entries of $\mathbf{w}_{\mathrm{ee}}$ are non-zero, v_1 can be calculated using $n-k$ multiplications and $n-k-1$ additions, for a total of $2(n-k)-1$ floating point operations (FLOPs). Checking the condition $|v_1| \geq \tau$ in (1) is a mere extra FLOP. Hence, if $|v_1| \geq \tau$, the conditional perceptron consumes $2(n-k)+1$ FLOPs, where we have assumed that another FLOP is spent on the activation function. Otherwise, $1+2k$ more FLOPs need to be performed to obtain v_0 out of v_1 via the relationship $v_0 = (\|\mathbf{w}_{\mathrm{p}}\|/\|\mathbf{w}\|)v_1 + \sum_{j=1}^{n-k} w_{i_j}x_{i_j}$. Hence, letting μ_{uc} and μ_{c} denote the FLOPs required to implement an ordinary perceptron and a conditional perceptron, respectively, we have $\mu_{\mathrm{uc}} \triangleq 2n$, and

$$\mu_{\mathrm{c}} \triangleq \begin{cases} 2(n-k)+1, & |v_1| \geq \tau, \\ 2n+2, & \text{otherwise}. \end{cases} \tag{2}$$

On average, we thus expect the FLOPs with a conditional perceptron to be notably less than the FLOPs with an ordinary unconditional perceptron, without any significant penalty in terms of the classification performance.

3. A Concentration Theorem for Pruning

The operation of the conditional perceptron in (1) is intimately related to weight pruning in neural networks. Indeed, $\mathbf{w}_{\mathrm{ee}}$ is a pruned version of $\mathbf{w}$, followed by normalization. Pruning is achieved by retaining only the top k components of $\mathbf{w}$ with the highest magnitudes, while setting all other components to zero. Hence, a conditional perceptron first evaluates the local field using a pruned version of its weights. If this local field is confident enough, an early exit is performed. The main difference between the conditional perceptron and pruning is the former's switch to the unpruned full set of weights whenever the local field provided by the pruned weights is not confident enough.

Regardless of the particular emphasis, whether it is the conditional perceptron or sole pruning, the relationship between the original feature extracting vector $\mathbf{w}$ and its pruned counterpart $\mathbf{w}_{\mathrm{ee}}$ becomes a crucial point of interest. In particular, there has been a lot of recent interest on neuron-wise pruning for large language models [7, 21]. This line of work assumes that fine-tuning or retraining of models is not possible after pruning due to computational complexity, a condition we also adopt in our setting. To understand the statistical properties of the similarity $\mathbf{w}_{\mathrm{ee}}^T\mathbf{w}$, we assume $\mathbf{w}$ is uniform on the unit hypersphere, and derive the corresponding statistics of $\mathbf{w}_{\mathrm{ee}}^T\mathbf{w}$. We will consider the regime where the dimension of $\mathbf{w}$ grows to infinity. In this regime of high dimensions, an unexpected outcome emerges. Assuming that a positive fraction of the weights is retained during pruning, we demonstrate that the similarity denoted as $\mathbf{w}_{\mathrm{ee}}^T\mathbf{w}$ tends to converge in probability towards a non-zero constant. In other words, it essentially becomes deterministic. We will show that a similar result holds for conventional pruning in the sense that $\mathbf{w}_{\mathrm{p}}^T\mathbf{w}$ also concentrates to its mean.

Our results will follow from a concentration theorem related to order statistics of Gamma random variables. Hence, we consider a general Gamma random variable G with PDF

$$f_G(x) = \frac{1}{\Gamma(k)\theta^s} x^{s-1} e^{-x/\theta},\, x \geq 0, \tag{3}$$

where s and θ are known as the shape and the scale parameters, respectively, and $\Gamma(\cdot)$ is the Gamma function. Let $G_{\leq y}$ denote the random variable G conditioned on $G \leq y$, where $y \in [0, \infty)$. The notation $G_{\geq y}$ is defined similarly.

Theorem 1 *Let $G_1, G_2, \ldots, G_n$ be independent and identically distributed Gamma random variables with shape s and scale θ. Denote the corresponding order statistics by $G_{(1)} \leq \cdots \leq G_{(n)}$. For a given $0 < q < 1$, let*

$$\Xi_n \triangleq \frac{G_{(\lceil qn \rceil + 1)} + \cdots + G_{(n)}}{G_1 + \cdots + G_n}. \tag{4}$$

Let ω_q denote the unique real number such that $\mathrm{P}(G \leq \omega_q) = q$. Define the corresponding threshold

$$\tau_q(s, \theta) \triangleq \left(1 + \frac{q\mathrm{E}\big[G_{\leq \omega_q}\big]}{(1-q)\mathrm{E}\big[G_{\geq \omega_q}\big]}\right)^{-1} \in [0, 1]. \tag{5}$$

Then, we have, $\Xi_n \to \tau_q(s, \theta)$ as $d \to \infty$.

Corollary 2 *Suppose $\mathbf{W}$ is uniform on $\mathbb{S}^n$. Given $q \in (0, 1)$, let $\mathbf{W}_{\mathrm{p}} = f_{\mathrm{p}}^{\lceil nq \rceil}(\mathbf{W})$ and $\mathbf{W}_{\mathrm{ee}} = f_{\mathrm{ee}}^{\lceil nq \rceil}(\mathbf{W})$. As $n \to \infty$, we have $\mathbf{W}^T\mathbf{W}_{\mathrm{ee}} \to \sqrt{\tau_q}$, and $\mathbf{W}^T\mathbf{W}_{\mathrm{p}} \to \tau_q$, where $\tau_q \triangleq \tau_q(s = \frac{1}{2}, \theta = 2)$.*

We make two observations: First, both the pruned vector $\mathbf{W}_{\mathrm{p}}$ and the early exit vector $\mathbf{W}_{\mathrm{ee}}$ remain at a constant angle or similarity with respect to the original weight vector asymptotically for large feature dimension. Second, as Fig. 1 shows, the similarity has a non-linear relationship with respect to the pruning rate, and remains very high even for large pruning rates. For example, even when 60% of the components of a unit norm weight vector are pruned, which corresponds to $q = 0.4$, after normalizing the pruned vector to unit norm, the resulting vector will have a similarity of roughly 0.982, or an angle of only $10.8°$.

The phenomenon of concentration towards large similarities, even at high pruning rates, appears to be fundamental. This helps explain why pruned networks often perform nearly as well as their unpruned versions, a fact frequently noted in existing research. To elaborate further, we shall proceed in an informal fashion: Pruning a deep neural network at a certain rate would be roughly equivalent to pruning all its neurons at the same rate. In the pruned network, all neurons would approximately operate in the same manner as if they were in the unpruned network, provided that the pruning rate is not very high. This is because the weights of the pruned and unpruned networks then have a high similarity as a result of Corollary 2. Consequently, we anticipate the pruned network's performance to closely match that of the unpruned network. Next, we turn our attention back to the case of a single neuron, and in particular, the conditional perceptron, which is much easier to analyze in a formal fashion.

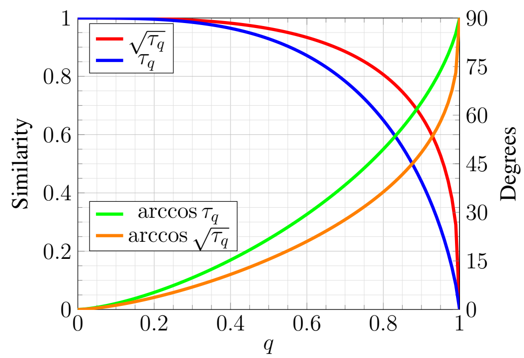

Figure 1: Concentrates for different sparsities.

4. Generalization Performance of Conditional Perceptrons

In this section, we analyze the generalization error of the conditional perceptron in the classical student-teacher framework of learning theory, building on the concentration results in Section 3. To the best of our knowledge, this represents the first instance of a generalization analysis for a neural conditional computation or early exit system in the literature.

4.1. Learning on the Unconditional Perceptron

We first provide an overview of learning on an ordinary, unconditional perceptron: Consider a dataset of N_t training vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_{N_t}\} \subset \mathbb{S}^n$ and a teacher $\mathbf{t} \in \mathbb{S}^n$. Given $i \in \{1, \ldots, N_t\}$, let $z_i \in \{-1, +1\}$ denote the desired output for training vector $\mathbf{y}_i$. The teacher determines the desired outputs in the sense that we set $z_i = 1$ whenever $\mathbf{t}^\dagger \mathbf{y}_i \geq 0$ and $z_i = -1$ if $\mathbf{t}^\dagger \mathbf{y}_i < 0$.

We consider now a student $\mathbf{w} \in \mathbb{R}^n$ acquired through some learning algorithm, as a function of only the input-output pairs $(\mathbf{y}_1, z_1), \ldots, (\mathbf{y}_{N_t}, z_{N_t})$. The student weights coincide precisely with the "main" weights of the conditional perceptron as defined in Section 1, and hence we used the same notation. The specific learning algorithm to obtain $\mathbf{w}$ out of the training vectors is not vital for our purposes. For example, the student can be chosen to be the vector that classifies the training data with the maximal margin, i.e. $\mathbf{w} = \arg\max_{\mathbf{w}_0 \in \mathbb{S}^n} \min_{i \in \{1,\ldots,N_t\}} z_i \mathbf{w}_0^T \mathbf{y}_i$. The generalization error provided by the student is the probability $P(\sigma(\mathbf{X}^T\mathbf{w}) \neq \sigma(\mathbf{X}^T\mathbf{t}))$ of mismatch between the student and teacher decisions when the input $\mathbf{X}$ to the perceptron is assumed to uniform on $\mathbb{S}^n$. For a fixed student and teacher vectors, the generalization error can simply be evaluated to be $\frac{1}{\pi}\arccos \mathbf{w}^t\mathbf{t}$.

We are often interested in the generalization error when averaged out over random datasets and teachers. For this purpose, suppose $\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_t}, \mathbf{T} \sim N(\mathbf{0}_d, \mathbf{I}_d)$ are mutually independent. The student is then the random vector $\mathbf{W} = \arg\max_{\mathbf{w}_0 \in \mathbb{S}^n} \min_{i\in\{1,\ldots,N_t\}} \sigma(\mathbf{T}^T\mathbf{Y}_i)\mathbf{w}_0^T\mathbf{Y}_i$, and the generalization error is given by $\epsilon_{\text{uc}} \triangleq \mathrm{P}(\sigma(\mathbf{X}^T\mathbf{W}) \neq \sigma(\mathbf{X}^T\mathbf{T})) = \frac{1}{\pi}\mathrm{E}[\arccos \mathbf{W}^T\mathbf{T}]$. It is difficult to calculate the generalization error exactly except for a few special cases. A special case is when both the number k of training vectors as well as the ambient dimension n grows to infinity. In other words, we have $n, N_t \to \infty$. If $\alpha \triangleq \lim_{n\to\infty} \frac{N_t}{n}$ exists, then it is known [18, 20] that there is a constant $C > 0$ such that $\epsilon_{\text{uc}} \sim \overline{\epsilon_{uc}} = \frac{C}{\alpha}$ as $\alpha \to \infty$. Here, we used the notation $\overline{\epsilon_{uc}}$ denote the asymptotic $n, N_t \to \infty$ generalization error for the unconditional perceptron.

4.2. Analyzing the Conditional Perceptron

Let us now analyze the generalization performance of the conditional perceptron. Consider the same learning formulation as in Section 4.1. Define the early exit vector $\mathbf{W}_{\text{ee}} \triangleq f_{\text{ee}}^{\lceil nq \rceil}(\mathbf{W})$, where $0 < q < 1$. We extend the definition in (1) to random variables via $Y_{\text{c}} = \sigma(\mathbf{W}_{\text{ee}}^T\mathbf{X})$ if $|\mathbf{W}_{\text{ee}}^T\mathbf{X}| \geq \tau$, and $Y_{\text{c}} = \sigma(\mathbf{W}^T\mathbf{X})$ if $|\mathbf{W}_{\text{ee}}^T\mathbf{X}| < \tau$. The generalization error of the conditional perceptron is $\epsilon_{\text{c}} \triangleq \mathrm{P}(Y_{\text{c}} \neq \sigma(\mathbf{X}^T\mathbf{T}))$. We expect $\epsilon_{\text{c}} \geq \epsilon_{\text{uc}}$ as the conditional perceptron often operates with a pruned, normalized version $\mathbf{W}_{\text{ee}}$ of the student vector $\mathbf{W}$. We anticipate that any decrease in accuracy will be offset by a reduction in computational demand. To determine the corresponding tradeoff between the average computation and accuracy, we calculate the average FLOPs for the conditional perceptron, by averaging out (2) over the random inputs $\mathbf{X}$. This yields

$$\overline{\mu}_c \triangleq (2(n-k)+1)\mathrm{P}(|\mathbf{W}_{\text{ee}}^T\mathbf{X}| \geq \tau) + (2n+2)(1-\mathrm{P}(|\mathbf{W}_{\text{ee}}^T\mathbf{X}| \geq \tau)). \tag{6}$$

It is convenient to normalize the FLOPs with respect to the FLOPs $2n$ of the unconditional network in the $n \to \infty$ asymptotic regime. For this purpose, we define $\overline{\mu}'_c \triangleq \lim_{n\to\infty} \frac{\overline{\mu}_c}{2n}$. Then, the maximum FLOPs is 1, spent by the unconditional perceptron. The following theorem provides a set of achievable pairs of FLOPs and generalization errors provided by the conditional perceptron.

Theorem 3 *Let $\epsilon > 0$. For a given computation constraint $\overline{\mu}'_c \leq 1 - \epsilon$, a generalization error of $\epsilon_{\text{uc}} + (\epsilon/q)^{\frac{1}{2(1-\rho^2)}}$ is achievable, where $\rho = \cos(\arccos\sqrt{\tau_q} + \pi\overline{\epsilon_{\text{uc}}})$.*

Now, fix some "reasonable" sparsity rate q so that ρ is not too far from zero, and imagine ϵ as the only variable. Then, the theorem shows that near the full FLOPs of 1, the system performance approaches the unconditional performance exponentially fast with rate $O(\epsilon^A)$ for some constant $A > 1$. This helps explain why early exit networks do suffer significant performance loss despite one cuts of a decent chunk of the computation budget, i.e. ϵ is not close to zero. In fact, the exponent grows to infinity if $\rho \to 0$. Our analysis suggests this can happen if the sparsity rate nears zero and the training rate is high, so that $\overline{\epsilon_{\text{uc}}} \to 0$. A similar situation arises in the context of pruning and the behavior of similarity scores in Fig. 1: They remain relatively unchanged up to pruning rates of 0.4, and only then begin to decrease. General conditional computation networks, such as mixtures of experts, exhibit a similar behavior: the network performance remains robust unless the computation budget is overly restricted. The results presented in this paper also shed light on this phenomenon.

Acknowledgments

This work was supported in part by the Army Research Lab (ARL) under Grant W911NF-21-2-0272, in part by the Army Research Office (ARO) under Grant W911NF-24-1-0049, and in part by the National Science Foundation (NSF) under Grant CNS-2148182.

References

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

[2] Mohammad Ahsanullah, Valery B Nevzorov, Mohammad Shakil, Mohammad Ahsanullah, Valery Nevzorov, and Mohammad Shakil. Conditional distributions of order statistics. *An Introduction to Order Statistics*, pages 51–60, 2013.

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[4] Arnaud Buhot, Juan-Manuel Torres Moreno, and Mirta B Gordon. Finite size scaling of the bayesian perceptron. *Physical Review E*, 55(6):7434, 1997.

[5] Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

[6] Andrew Davis and Itamar Arel. Low-rank approximations for conditional feedforward computation in deep neural networks. *arXiv preprint arXiv:1312.4461*, 2013.

[7] Lucio Dery, Steven Kolawole, Jean-Francois Kagey, Virginia Smith, Graham Neubig, and Ameet Talwalkar. Everybody prune now: Structured pruning of llms with only forward passes. *arXiv preprint arXiv:2402.05406*, 2024.

[8] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

[9] Carl-Gustav Esseen. On the liapunoff limit of error in the theory of probability. *Arkiv för Matematik, Astronomi och Fysik*, A28:1–19, 1942. ISSN 0365-4133.

[10] Alperen Gormez, Venkat Dasari, and Erdem Koyuncu. Class means as an early exit decision mechanism. In *IJCNN*, 2022.

[11] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7436–7456, 2021.

[12] James J Heckman and Bo E Honore. The empirical content of the roy model. *Econometrica: Journal of the Econometric Society*, pages 1121–1149, 1990.

[13] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[14] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *International Conference on Machine Learning*, June 2019.

[15] Erdem Koyuncu and Hamid Jafarkhani. Distributed beamforming in wireless multiuser relay-interference networks with quantized feedback. *IEEE Transactions on Information Theory*, 58 (7):4538–4576, 2012. doi: 10.1109/TIT.2012.2191708.

[16] Yuhang Li, Tamar Geller, Youngeun Kim, and Priyadarshini Panda. Seenn: Towards temporal spiking early exit neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

[18] Manfred Opper. Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pages 922–925, 1995.

[19] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[20] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

[21] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

[22] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE, 2016.

[23] Shoufang Xu and Yu Miao. Limit behaviors of the deviation between the sample quantiles and the quantile. *Filomat*, 25(2):197–206, 2011.

Appendix A. Proof of Theorem 1

A.1. Some auxiliary results on Gamma random variables

We begin by presenting some auxiliary results concerning Gamma random variables that we will need to prove the theorem. As defined in Section 3, a general Gamma random variable G has PDF $f_G(x) = x^{s-1}e^{-x/\theta}/\Gamma(k)/\theta^s$, $x \geq 0$, where s and θ are known as the shape and the scale parameters, respectively, and $\Gamma(\cdot)$ is the Gamma function. The lower and upper incomplete gamma functions are defined as $\gamma(s,x) \triangleq \int_0^x t^{s-1}e^{-t}\mathrm{d}t$ and $\Gamma(s,x) \triangleq \int_x^\infty t^{s-1}e^{-t}\mathrm{d}t$, respectively.

We note the asymptotic expressions [1, Eqs. 6.5.29 and 6.5.32]

$$\gamma(s,x) \sim x^s/s,\ x \to 0, \tag{7}$$

$$\gamma(s,x) \to \Gamma(s),\ x \to \infty, \tag{8}$$

$$\Gamma(s,x) \to \Gamma(s),\ x \to 0, \tag{9}$$

$$\Gamma(s,x) \sim x^{s-1}e^{-x},\ x \to \infty. \tag{10}$$

According to (7), we have

$$\mathrm{P}(G \leq y) = \frac{1}{\Gamma(s)}\gamma(s, y/\theta) \sim \frac{(y/\theta)^s}{\Gamma(1+s)},\ y \to 0, \tag{11}$$

and by (10), we obtain,

$$\mathrm{P}(G \geq y) = \frac{1}{\Gamma(s)}\Gamma(s, y/\theta) \sim \frac{(y/\theta)^{s-1}e^{-y/\theta}}{\Gamma(s)},\ y \to \infty, \tag{12}$$

Let $G_{\leq y}$ denote a truncated Gamma random variable, obtained by conditioning G on the event $G \leq y$, where $y \in [0, \infty)$. The notation $G_{\geq y}$ is defined similarly. A straightforward calculation reveals that for $a \geq 0$, we have

$$\mathrm{E}G_{\leq y}^a = \frac{\int_0^y x^a f_G(x)\mathrm{d}x}{\int_0^y f_G(x)\mathrm{d}x} = \frac{\int_0^y x^{a+s-1}e^{-x/\theta}\mathrm{d}x}{\int_0^y x^{s-1}e^{-x/\theta}\mathrm{d}x} = \frac{\int_{-\infty}^{y/\theta} u^{a+s-1}\theta^{a+s}e^{-u}\mathrm{d}u}{\int_{-\infty}^{y/\theta} u^{s-1}\theta^s e^{-u}\mathrm{d}u} = \frac{\theta^a\gamma(a+s, y/\theta)}{\gamma(s, y/\theta)}, \tag{13}$$

and similarly,

$$\mathrm{E}G_{\geq y}^a = \frac{\theta^a\Gamma(a+s, y/\theta)}{\Gamma(s, y/\theta)}. \tag{14}$$

In (13), the first equality is by definition. To obtain the second equality, we substituted the PDF of G. The third equality is by a change of variables $u = x/\theta$. The final equality is by the definition of the lower incomplete Gamma function. For $a = 1$, using (7), we can then obtain

$$\mathrm{E}G_{\leq y}^a \sim \frac{\theta^a(y/\theta)^{a+s}/(a+s)}{(y/\theta)^s/s} = \frac{sy^a}{a+s},\ y \to 0, \tag{15}$$

and using (8) yields

$$\mathrm{E}G_{\leq y}^a \to \frac{\theta^a\Gamma(a+s)}{\Gamma(s)},\ y \to \infty. \tag{16}$$

In a similar vein, using (9), we have

$$\mathrm{E}G^{a}_{\geq y} \to \frac{\theta^{a}\Gamma(a+s)}{\Gamma(s)},\ y \to 0. \tag{17}$$

and applying (10), we obtain

$$\mathrm{E}G^{a}_{\geq y} \sim \frac{\theta^{a}(y/\theta)^{a+s-1}e^{-y/\theta}}{(y/\theta)^{s-1}e^{-y/\theta}} = y^{a},\ y \to \infty. \tag{18}$$

We begin with a useful lemma on the expected values of truncated Gamma random variables.

Lemma 4 *The derivative of the function*

$$y \mapsto \mathrm{E}[G_{\leq y}]. \tag{19}$$

is bounded.

Proof We have

$$\mathrm{E}[G_{\leq y}] = \frac{\int_0^y xf(x)dx}{\int_0^y f(x)dx} \tag{20}$$

Hence, by the fundamental theorem of calculus,

$$\frac{\mathrm{dE}[G_{\leq y}]}{\mathrm{d}y} = \frac{yf(y)P(G \leq y) - \mathrm{E}[G_{\leq y}]P(G \leq y)f(y)}{P^2(G \leq y)} \tag{21}$$

$$= \frac{f(y)}{P(G \leq y)}(y - \mathrm{E}[G_{\leq y}]) \tag{22}$$

As $y \to 0$, according to (3), (11), and (15), we obtain

$$\lim_{y\to 0} \frac{\mathrm{dE}[G_{\leq y}]}{\mathrm{d}y} = \frac{s}{1+s} \tag{23}$$

On the other hand, by (3) and (16), we have

$$\lim_{y\to \infty} \frac{\mathrm{dE}[G_{\leq y}]}{\mathrm{d}y} = 0 \tag{24}$$

The statement of the lemma then follows from the continuity of $y \mapsto \mathrm{E}[G_{\leq y}]$. ■

The following useful lemma is a standard bound on powers of linear functions.

Lemma 5 ([15, Lemma 7]) *For any real numbers $x_1, \ldots, x_n \geq 0$, we have*

$$\left(\sum_{i=1}^n x_i\right)^{\beta} \leq n^{\beta-1} \sum_{i=1}^n x_i^{\beta}. \tag{25}$$

Our main technical result, Theorem 1, shows that a certain ratio related to order statistics of Gamma random variables converges in probability to a threshold given by (5). The following lemma shows that the derivatives of a more general form of the threshold function is bounded from above.

Lemma 6 *Let $q \in (0,1)$, $n \geq 1$ and $k = \lceil nq \rceil$. Define the function*

$$h(x) \triangleq \frac{(n-k)\mathrm{E}[G_{\geq x}]}{k\mathrm{E}[G_{\leq x}] + (n-k)\mathrm{E}[G_{\geq x}]}. \tag{26}$$

There is a constant $C_1 > 0$ that is independent of n such that $|\frac{\mathrm{d}h}{\mathrm{d}x}| < C_1$, $\forall x \in \mathbb{R}$ and for all large enough n.

Proof Let $\nu(x) = \mathrm{E}G_{\leq x}/\mathrm{E}G_{\geq x}$. We can rewrite (80) as

$$h(x) = \Big(1 + \frac{k}{n-k}\nu(x)\Big)^{-1}, \tag{27}$$

The derivative of (27) is calculated to be

$$\frac{\mathrm{d}h}{\mathrm{d}x} = -\frac{k}{n-k}\Big(1 + \frac{k}{n-k}\nu(x)\Big)^{-2}\frac{\mathrm{d}\nu}{\mathrm{d}x} \tag{28}$$

It follows that

$$\Big|\frac{\mathrm{d}h}{\mathrm{d}x}\Big| \leq \frac{k}{n-k}\Big|\frac{\mathrm{d}\nu}{\mathrm{d}x}\Big| \leq \frac{2q}{1-q}\Big|\frac{\mathrm{d}\nu}{\mathrm{d}x}\Big|. \tag{29}$$

The inequality follows since $\lim_{n\to\infty} \frac{k}{n-k} = \frac{q}{1-q}$. We used twice the limit as an upper bound, which will be valid for every large enough n.

What is now left to show is that the derivative $\frac{\mathrm{d}\nu}{\mathrm{d}x}$ is bounded. It is sufficient to prove that the limits $\lim_{x\to 0} \mathrm{d}\nu/\mathrm{d}x$ and $\lim_{x\to\infty} \mathrm{d}\nu/\mathrm{d}x$ exist and they are finite, and that $\mathrm{d}\nu/\mathrm{d}x$ is continuous on $(0,\infty)$. First, we calculate the derivative. Let

$$N(x) \triangleq \int_0^x yf(y)dy \int_x^\infty f(y)dy = [\mathrm{E}[G_{\leq x}]P(G \leq x)]P(G \geq x), \tag{30}$$

$$D(x) \triangleq \int_x^\infty yf(y)dy \int_0^x f(y)dy = [\mathrm{E}[G_{\geq x}]P(G \geq x)]P(G \leq x). \tag{31}$$

We note the alternate representation $\nu(x) = \frac{N(x)}{D(x)}$. By the fundamental theorem of calculus, we obtain

$$\frac{\mathrm{d}N}{\mathrm{d}x} = xf(x)\mathrm{P}(G \geq x) - \mathrm{E}[G_{\leq x}]P(G \leq x)f(x), \tag{32}$$

$$\frac{\mathrm{d}D}{\mathrm{d}y} = -xf(x)\mathrm{P}(G \leq x) + \mathrm{E}[G_{\geq x}]P(G \geq x)f(x). \tag{33}$$

Using the division rule for derivatives, after some cumbersome but straightforward calculations, we can obtain

$$\frac{\mathrm{d}\nu}{\mathrm{d}x} = f(x)\frac{x\mathrm{P}(G \geq x)\mathrm{E}[G_{\geq x}] + x\mathrm{P}(G \leq x)\mathrm{E}[G_{\leq x}] - \mathrm{E}[G_{\geq x}]\mathrm{E}[G_{\leq x}]\big[\mathrm{P}(G \leq x) + \mathrm{P}(G \geq x)\big]}{\mathrm{E}^2[G_{\geq x}]\mathrm{P}(G \leq x)\mathrm{P}(G \geq x)} \tag{34}$$

$$= \frac{f(x)\big[xE[G] - \mathrm{E}[G_{\leq x}]\mathrm{E}[G_{\geq x}]\big]}{\mathrm{E}^2[G_{\geq x}]\mathrm{P}(G \leq x)\mathrm{P}(G \geq x)}. \tag{35}$$

Since all functions involved in (35) are continuous, $\mathrm{d}\nu/\mathrm{d}x$ is continuous except possibly at 0 and ∞. Using (3), (11), (15), (17), and noting that $E[G] = k\theta$, we obtain

$$\lim_{x\to 0} \frac{\mathrm{d}\nu}{\mathrm{d}x} = \frac{1}{\theta(1+s)} \tag{36}$$

Also, substituting (3), (12), (16), and (18), to (35), we can show that the derivative at ∞ is zero. Together with (36), this shows that the derivative is bounded. This concludes the proof. ■

Let us now derive upper and lower bounds on the variances of truncated Gamma random variables.

Lemma 7 *For every $x \geq 0$, the variances of truncated Gamma random variables follow the bounds*

$$\max\{s,1\}\theta^2 \geq \mathrm{var}(G_{\geq x}) \geq \min\{s,1\}\theta^2, \tag{37}$$

$$C_2 \geq \mathrm{var}(G_{\leq x}), \tag{38}$$

where C_2 is a constant that is independent of x.

Proof Assume that the shape parameter s of the Gamma random variable G satisfies $s \in (0,1]$. Then, G is log-convex, and according to [12, Proposition 2], the function $x \mapsto \mathrm{var}(G_{\geq x})$ is monotonically increasing. In particular, $\mathrm{var}(G_{\geq x}) \geq \mathrm{var}(G_{\geq 0}) = \mathrm{var}(G) = s\theta^2$. On the other hand, when $s \in [1,\infty)$, the density G is log-concave. In this case, [12, Proposition 1] shows that $x \mapsto \mathrm{var}(G_{\geq x})$ is monotonically decreasing. Hence, we have $\mathrm{var}(G_{\geq x}) \geq \lim_{x\to\infty} \mathrm{var}(G_{\geq x})$. In what follows, we calculate the limit. We have

$$\mathrm{var}(G_{\geq x}) = \mathrm{E}[G_{\geq x}^2] - \mathrm{E}^2[G_{\geq x}] \tag{39}$$

$$= \theta^2 \frac{\Gamma(2+s,\frac{x}{\theta})\Gamma(s,\frac{x}{\theta}) - \Gamma^2(1+s,\frac{x}{\theta})}{\Gamma^2(s,\frac{x}{\theta})} \tag{40}$$

We have a $\frac{0}{0}$ indeterminancy as $x \to \infty$. We can thus apply L'Hôpital's rule to obtain

$$\lim_{x\to\infty} \mathrm{var}(G_{\geq x}) = \lim_{x\to\infty} \frac{x^2\Gamma(s,\frac{x}{\theta}) - 2\theta x\Gamma(1+s,\frac{x}{\theta}) + \theta^2\Gamma(2+s,\frac{x}{\theta}))}{2\Gamma(s,\frac{x}{\theta})} \tag{41}$$

$$= \lim_{x\to\infty} \theta^s \frac{t\Gamma(1+s,\frac{x}{\theta}) - x\Gamma(s,\frac{x}{\theta})}{e^{-\frac{x}{\theta}}x^{s-1}} \tag{42}$$

$$= \lim_{x\to\infty} \frac{\theta^{1+s}\Gamma(s,\frac{x}{\theta})}{e^{-\frac{x}{\theta}}x^{-2+s}(\theta - s\theta + x)} \tag{43}$$

$$= \theta^2 \tag{44}$$

The second and the third equalities also follow from L'Hôpital's rule. In order to obtain the final equality, we have applied (10). Note that the derivatives of the upper incomplete Gamma function can be evaluated via the formulae $\frac{\mathrm{d}\Gamma(s,x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}\int_x^\infty t^{s-1}e^{-t}\mathrm{d}t = -x^{s-1}e^{-x}$, by the fundamental theorem of calculus. Therefore, for any s, we obtain $\mathrm{var}(G_{\geq x}) \geq \min\{s,1\}\theta^2$.

Since $\mathrm{var}(G_{\leq x}) = \mathrm{E}[G_{\leq x}^2] - [\mathrm{E}G_{\leq x}]^2 \leq \mathrm{E}[G_{\leq x}^2]$, according to (15) and (16), the lower conditional variance $\mathrm{var}(G_{\leq x})$ is bounded by a constant that is independent of x from above. For the upper conditional variance $\mathrm{var}(G_{\geq x})$, we consider the cases of $s \in (0,1)$ and $s \in [1,\infty)$ separately. In the former scenario, the monotonically increasing nature of $x \mapsto \mathrm{var}(G_{\geq x})$, as established in [12, Proposition 2], in conjunction with (44), demonstrates that $\mathrm{var}(G_{\geq x}) \leq \theta^2$ for every x. For $s \geq 1$, we obtain $\mathrm{var}(G_{\geq x}) \leq \mathrm{var}(G_{\geq 0}) = s\theta^2$, according to [12, Proposition 1]. Hence, for any s, we have $\mathrm{var}(G_{\geq x}) \leq \theta^2 \max\{1, s\}$. This concludes the proof of the lemma. ■

As a corollary, we obtain lower and upper bounds on a linear combination of variances truncated Gamma random variables.

Corollary 8 *Let $q \in (0,1)$, $n \geq 1$ and $k = \lceil nq \rceil$. Let*

$$\sigma^2 \triangleq y^2(k-1)\mathrm{var}(G_{\leq x}) + (1-y)^2(n-k)\mathrm{var}(G_{\geq x}). \tag{45}$$

Then, for every $\tau \in (0,1)$ and $y \in (0,1)$ with $|y-\tau| \leq \frac{1-\tau}{2}$, we have

$$C_3 n \leq \sigma^2 \leq C_4 n \tag{46}$$

for all sufficiently large n, where C_3 and C_4 are constants.

Proof According to Lemma 7, for any s, we obtain

$$\begin{aligned} \sigma^2 &\geq (1-y)^2(n-k)\mathrm{var}(G_{\geq x}) && (47)\\ &\geq (1-y)^2(n-k)\min\{s,1\}\theta^2 && (48) \end{aligned}$$

The bound $|y-\tau| \leq \frac{1-\tau}{2}$ implies

$$\sigma^2 \geq \left(\frac{1-\tau}{2}\right)^2 (n-k)\min\{s,1\}\theta^2 \tag{49}$$

Also, substituting $k = \lceil nq \rceil$, we obtain

$$\begin{aligned} \sigma^2 &\geq \left(\frac{1-\tau}{2}\right)^2 (n-\lceil nq \rceil)\min\{s,1\}\theta^2 && (50)\\ &\geq \frac{1-q}{2}\left(\frac{1-\tau}{2}\right)^2 \min\{s,1\}\theta^2 n, && (51) \end{aligned}$$

for sufficiently large n. This proves the lower estimate on the variance.

For the upper estimate, we can first obtain

$$\sigma^2 \leq \underbrace{y^2}_{\leq 1}\underbrace{(k-1)}_{\leq n}\mathrm{var}(G_{\leq x}) + \underbrace{(1-y)^2}_{\leq 1}\underbrace{(n-k)}_{\leq n}\mathrm{var}(G_{\geq x}) = n[\mathrm{var}(G_{\leq x}) + \mathrm{var}(G_{\geq x})], \tag{52}$$

and applying Lemma 7 proves the upper estimate on σ^2. This concludes the proof of the corollary. ■

The following lemma is utilized to bound the error terms resulting from the Berry-Esseen estimates.

Lemma 9 *Let $q, y \in (0,1)$, $n \geq 1$ and $k = \lceil nq \rceil$. Let σ^2 be as defined in (45) of Corollary 8, and*

$$\rho \triangleq y^3(k-1)\mathrm{E}\big|G_{\leq x} - \mathrm{E}G_{\leq x}\big|^3 + (1-y)^3(n-k)\mathrm{E}\big|G_{\geq x} - \mathrm{E}G_{\geq x}\big|^3. \tag{53}$$

For every large enough n, we have

$$\int_0^\infty \frac{\rho}{\sigma^3} f_{G_{(k)}}(x)\mathrm{d}x \leq C_6 n^{-\frac{1}{2}} \tag{54}$$

for some constant $C_6 > 0$ that is independent of y, k, n.

Proof Using the inequalities $y \leq 1$ and $k \leq n$, we obtain

$$\rho \leq n\mathrm{E}|G_{\leq x} - \mathrm{E}G_{\leq x}|^3 + n\mathrm{E}|G_{\geq x} - \mathrm{E}G_{\geq x}|^3. \tag{55}$$

Applying Lemma 5 yields

$$\begin{aligned}
\rho &\leq 4n(\mathrm{E}G_{\leq x}^3 + \mathrm{E}^3 G_{\leq x} + \mathrm{E}G_{\geq x}^3 + \mathrm{E}^3 G_{\geq x}) && (56)\\
&\leq 8n(\mathrm{E}G_{\geq x}^3 + \mathrm{E}^3 G_{\geq x}) && (57)\\
&\leq C_7 n(1+x^3), && (58)
\end{aligned}$$

where C_7 is a constant that is independent of n and x. The last inequality is a consequence of (17) and (18).

It is a standard result in probability theory that the exact PDF for the kth order statistic $G_{(k)}$ is given by

$$\begin{aligned}
f_{G_{(k)}}(x) &= \frac{n!}{(k-1)!(n-k)!} f_G(x)[F_G(x)]^{k-1}[1-F_G(x)]^{n-k} && (59)\\
&\leq n2^{n-1} f_G(x)[1-F_G(x)]^{n-k} && (60)
\end{aligned}$$

Since $F_G(x) \to 1$, there exists $x_0 > 0$ and $c \in (0,1)$ such that for every $x \geq x_0$, we have $n2^{n-1}[1-F_G(x)]^{n-\lceil nq \rceil} \leq c^{-n}$. As a result, we obtain $f_{G_{(k)}}(x) \leq f_G(x)c^{-n}, \forall x \geq x_0$. Combining this with (58) and the lower bound on σ in Corollary 8, we obtain

$$\begin{aligned}
\int_0^\infty \frac{\rho}{\sigma^3} f_{G_{(k)}}(x)\mathrm{d}x &= \int_0^{x_0} \frac{\rho}{\sigma^3} f_{G_{(k)}}(x)\mathrm{d}x + \int_{x_0}^\infty \frac{\rho}{\sigma^3} f_{G_{(k)}}(x)\mathrm{d}x && (61)\\
&\leq \int_0^{x_0} \frac{C_7 n(1+x_0^3)}{(C_3 n)^{1.5}} f_{G_{(k)}}(x)\mathrm{d}x + \int_{x_0}^\infty \frac{C_7 n(1+x^3)}{(C_3 n)^{1.5}} f_G(x)c^{-n}\mathrm{d}x && (62)\\
&\leq \frac{C_7(1+x_0^3)}{C_3^{1.5}} n^{-\frac{1}{2}} + \frac{C_7(1+x_0^3)}{C_3^{1.5}} (\mathrm{E}G + \mathrm{E}G^3) o(n^{-\frac{1}{2}}). && (63)
\end{aligned}$$

The proof is now complete since the moments of a Gamma random variable are also finite. ■

A.2. Other auxiliary results

As the first result, we recall the central limit theorem for quantiles.

Proposition 10 ([23]) *Let $X_1, X_2, \ldots, X_n$ be IID copies of a random variable X. Given $p \in (0,1)$, let $\xi_p = \inf\{x : F_X(x) \geq 1-p\}$. Suppose F_X has a continuous first derivative f_X in the neighborhood of ξ_p and $f(\xi_p) > 0$. Then,*

$$\frac{\sqrt{n} f_X(\xi_p)}{\sqrt{p(1-p)}}\Big(X_{(\lceil n(1-p)\rceil)} - \xi_p\Big) \sim N(0,1) \text{ as } n \to \infty. \tag{64}$$

We then present the general form of Berry-Esseen theorem for non-identically distributed random variables.

Proposition 11 (Berry-Esseen Theorem [9]) *Let $X_1, X_2, \ldots,$ be independent random variables with $E[X_i] = 0$ and $\mathrm{E}[|X_i|^3] < \infty$ for every $i \in \mathbb{Z}_{>0}$. Let Φ denote the CDF of the standard normal distribution. For all n, we have*

$$\sup_{x\in\mathbb{R}}\left|\mathrm{P}\left(\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}} \leq x\right) - \Phi(x)\right| \leq 8\left(\sum_{i=1}^{n} \mathrm{var} X_i\right)^{-3/2} \sum_{i=1}^{n} \mathrm{E}|X_i|^3. \tag{65}$$

A.3. Proof of the theorem

We are now ready to prove Theorem 1. We organize the proof in four steps.

Step-1: In the first step, we approximate the cumulative distribution function of Ξ_n via a normal random variable using the Berry-Esseen theorem. Let $y \in [0,1]$ and $k = \lceil nq \rceil$. We have

$$\mathrm{P}\Big(\Xi_n \leq y\Big) = \mathrm{P}\Big(G_{(k+1)} + \cdots + G_{(n)} \leq y(G_{(1)} + \cdots + G_{(n)})\Big) \tag{66}$$

$$= \int_0^\infty \mathrm{P}\left((1-y)\sum_{j=k+1}^{n} G_{(j)} - y\sum_{j=1}^{k-1} G_{(j)} \leq yx \,\middle|\, G_{(k)} = x\right) f_{G_{(k)}}(x)\mathrm{d}x. \tag{67}$$

According to [2], the two sets of order statistics $(G_{(k+1)}, \ldots, G_{(n)})$ and $(G_{(1)}, \ldots, G_{(k-1)})$ that appear in (67) are conditionally independent given $G_{(k)} = x$. Moreover, we have

$$\left[\sum_{j=k+1}^{n} G_{(j)} \middle| G_{(k)} = x\right] \sim \sum_{j=1}^{n-k} G_{\geq x, j}, \tag{68}$$

and

$$\left[\sum_{j=1}^{k-1} G_{(j)} \middle| G_{(k)} = x\right] \sim \sum_{j=1}^{k-1} G_{\leq x, j}, \tag{69}$$

where $G_{\leq x,1}, G_{\geq x,1}, G_{\leq x,2}, G_{\geq x,2}, \ldots$ is a sequence of IID random variables with

$$G_{\leq x, j} \sim G_{\leq x}, \tag{70}$$

and

$$G_{\geq x,j} \sim G_{\geq x} \tag{71}$$

for every $j \in \mathbb{Z}_{>0}$.

Let us now normalize the mean of (70) and (71) as

$$G'_{\leq x,j} \triangleq G_{\leq x,j} - \mathrm{E}[G_{\leq x,j}] = G_{\leq x,j} - \mathrm{E}[G_{\leq x}], \tag{72}$$

$$G'_{\geq x,j} \triangleq G_{\geq x,j} - \mathrm{E}[G_{\geq x,j}] = G_{\geq x,j} - \mathrm{E}[G_{\geq x}], \tag{73}$$

where $j \in \mathbb{Z}_{>0}$. We have

$$\mathrm{P}\Big(\Xi_n \leq y\Big) = \int_0^\infty \mathrm{P}\left((1-y)\sum_{j=1}^{n-k} G'_{\geq x,j} - y\sum_{j=1}^{k-1} G'_{\leq x,j} \leq y'\right) f_{G_{(k)}}(x)\mathrm{d}x, \tag{74}$$

where

$$y' \triangleq yx + y(k-1)\mathrm{E}[G_{\leq x}] - (1-y)(n-k)\mathrm{E}[G_{\geq x}]. \tag{75}$$

By definition. the random variables $G'_{\leq x,j}$ and $G'_{\geq x,j}$ have zero mean. According to Lemma 7, they also have finite normalized moments. Hence, Proposition 11 is applicable, and we have

$$\mathrm{P}\Big(\Xi_n \leq y\Big) \leq \int_0^\infty \Phi(y'/\sigma) f_{G_{(k)}}(x)\mathrm{d}x + 8\int_0^\infty \frac{\rho}{\sigma^3} f_{G_{(k)}}(x)\mathrm{d}x, \tag{76}$$

where Φ is the CDF of the standard normal random variable with zero mean and unit variance,

$$\sigma^2 \triangleq y^2(k-1)\mathrm{var}(G_{\leq x}) + (1-y)^2(n-k)\mathrm{var}(G_{\geq x}), \tag{77}$$

and

$$\rho \triangleq y^3(k-1)\mathrm{E}\left[\Big|G_{\leq x} - \mathrm{E}[G_{\leq x}]\Big|^3\right] + (1-y)^3(n-k)\mathrm{E}\left[\Big|G_{\geq x} - \mathrm{E}[G_{\geq x}]\Big|^3\right]. \tag{78}$$

Let us now rewrite the mean-normalized threshold y' defined in (75) as

$$y' = yx - y\mathrm{E}[G_{\leq x}] - \Big[k\mathrm{E}[G_{\leq x}] + (n-k)\mathrm{E}[G_{\geq x}]\Big]\Big(h(x) - y\Big), \tag{79}$$

where

$$h(x) \triangleq \frac{(n-k)\mathrm{E}[G_{\geq x}]}{k\mathrm{E}[G_{\leq x}] + (n-k)\mathrm{E}[G_{\geq x}]}. \tag{80}$$

Step-2: Let $\omega_q = \inf\{x : F_G(x) \geq q\}$ be as defined in the theorem statement. Let us also recall the threshold

$$\tau \triangleq \tau_q(s,\theta) = \left(1 + \frac{q\mathrm{E}\big[G_{\leq\omega_q}\big]}{(1-q)\mathrm{E}\big[G_{\geq\omega_q}\big]}\right)^{-1} = \frac{(1-q)\mathrm{E}[G_{\geq\omega_q}]}{q\mathrm{E}[G_{\leq\omega_q}] + (1-q)\mathrm{E}[G_{\geq\omega_q}]}. \tag{81}$$

from the theorem statement. In this step, we find an upper bound on y' for $y = \tau - \epsilon$ and $|x - \omega_q| \in O(\epsilon)$.

According to Lemma 6, we have

$$\sup_{x \in \mathbb{R}} \left| \frac{\mathrm{d}}{\mathrm{d}x} h(x) \right| \leq C_1, \tag{82}$$

for some constant $C_1 > 0$ that is independent of n. As a result, if $|x - \omega_q| \leq \delta \triangleq \frac{\epsilon}{2C_1}$, we have

$$\Big| h(x) - h(\omega_q) \Big| < C_1 \delta = \frac{\epsilon}{2}. \tag{83}$$

In comparison with

$$h(\omega_q) = \frac{(n - \lceil nq \rceil)\mathrm{E}[G_{\geq \omega_q}]}{\lceil nq \rceil \mathrm{E}[G_{\leq \omega_q}] + (n - \lceil nq \rceil)\mathrm{E}[G_{\geq \omega_q}]}, \tag{84}$$

it follows that

$$\lim_{n \to \infty} h(\xi_q) = \tau \tag{85}$$

Hence, for large enough n, we can argue that $|h(\omega_q) - \tau| \leq \frac{\epsilon}{4}$. Combining with (83), for $y = \tau - \epsilon$, we can thus obtain

$$h(x) - y \geq h(\omega_q) - \frac{\epsilon}{2} - y \tag{86}$$

$$= h(\omega_q) - \tau + \frac{\epsilon}{2} \tag{87}$$

$$\geq \frac{\epsilon}{4}. \tag{88}$$

On the other hand, we have the lower bounds

$$k\mathrm{E}[G_{\leq x}] + (n-k) \underbrace{\mathrm{E}[G_{\geq x}]}_{\geq \mathrm{E}[G_{\leq x}]} \geq n\mathrm{E}[G_{\leq x}] \tag{89}$$

$$\geq n\mathrm{E}[G_{\leq \omega_q - \delta}] \tag{90}$$

$$\geq n\Big(\mathrm{E}[G_{\leq \omega_q}] - C_8 \delta\Big) \tag{91}$$

$$\geq n\frac{\mathrm{E}[G_{\leq \omega_q}]}{2}, \tag{92}$$

where $C_8 > 0$ is a constant. In the above derivation, (90) follows since $x \mapsto \mathrm{E}[G_{\leq x}]$ is monotonically increasing. Inequality (91) is a consequence of Lemma 4, and the final inequality holds for all small enough ϵ. Substituting (88) and (92) to (79), we obtain

$$y' \leq yx - \frac{\mathrm{E}[G_{\leq \omega_q}]}{8} n\epsilon \tag{93}$$

$$\leq (\tau - \epsilon)(\omega_q + \delta) - \frac{\mathrm{E}[G_{\leq \omega_q}]}{8} n\epsilon \tag{94}$$

$$\leq 2\tau\omega_q - \frac{\mathrm{E}[G_{\leq \omega_q}]}{8} n\epsilon, \tag{95}$$

where the last inequality holds for all small enough ϵ.

Step-3: Let $y = \tau - \epsilon$ as in Step 2. We decompose the first integral in the upper bound in (76) as

$$\mathrm{P}\Big(\Xi_n \le y\Big) \le \int_{|x-\omega_q|\le\delta} \Phi(y'/\sigma) f_{G_{(k)}}(x)\mathrm{d}x + \int_{|x-\omega_q|\ge\delta} \underbrace{\Phi(y'/\sigma)}_{\le 1} f_{G_{(k)}}(x)\mathrm{d}x + 8\int_0^\infty \frac{\rho}{\sigma^3} f_{G_{(k)}}(x)\mathrm{d}x. \tag{96}$$

For the first term, we implement the upper limit on y' specified in (95), which holds a negative value for sufficiently large n. In this regime, we can apply the upper bound on σ as provided by Corollary 8 to obtain a valid upper limit for the first term in (96). Furthermore, an upper bound for the final term of (96) is obtained in Lemma 9. These estimates yield

$$\mathrm{P}\Big(\Xi_n \le y\Big) \le \int_{|x-\omega_q|\le\delta} \Phi\left[\frac{1}{\sqrt{C_4 n}}\Big(2\tau\omega_q - \frac{\mathrm{E}[G_{\le\omega_q}]}{8} n\epsilon\Big)\right] f_{G_{(k)}}(x)\mathrm{d}x + \int_{|x-\omega_q|\ge\delta} f_{G_{(k)}}(x)\mathrm{d}x + 8C_6 n^{-\frac{1}{2}}. \tag{97}$$

In (97), the $\Phi[\cdot]$-term decays to zero as $n \to \infty$. Moreover, the second integral also vanishes as $n \to \infty$ as a result of Proposition 10. Therefore, we obtain

$$\begin{aligned} \mathrm{P}\Big(\Xi_n \le y\Big) &\le o(1) \int_{|x-\omega_q|\le\delta} f_{G_{(k)}}(x)\mathrm{d}x + o(1) && (98)\\ &\le o(1). && (99) \end{aligned}$$

This shows that for any $\epsilon > 0$, we have $\mathrm{P}(\Xi_n \le \tau - \epsilon) \to 0$.

Step-4: Now, let $y = \tau + \epsilon$. We will show for any $\epsilon > 0$, we have $\mathrm{P}(\Xi_n \le \tau + \epsilon) \to 1$.

As in Step-2, for large enough n, we can argue that $|h(\xi_q) - \tau| \le \frac{\epsilon}{4}$. Combining with (83), for $y = \tau + \epsilon$, we obtain

$$\begin{aligned} h(x) - y &\le h(\xi_q) + \frac{\epsilon}{2} - y && (100)\\ &= h(\xi_q) - \tau - \frac{\epsilon}{2} && (101)\\ &\le -\frac{\epsilon}{4}. && (102) \end{aligned}$$

Combining with (92) and substituting to (79), we obtain

$$\begin{aligned} y' &\ge yx - y\underbrace{\mathrm{E}[G_{\le x}]}_{\le x} + \frac{\mathrm{E}[G_{\le\omega_q}]}{8} n\epsilon && (103)\\ &= \frac{\mathrm{E}[G_{\le\omega_q}]}{8} n\epsilon && (104) \end{aligned}$$

Analogous to the upper bound in (76), applying Proposition 11 yields the lower estimate

$$\begin{aligned}
\mathrm{P}\Big(\Xi_n \le y\Big) &\ge \int_0^\infty \Phi(y'/\sigma) f_{G_{(k)}}(x)\mathrm{d}x - 8\int_0^\infty \frac{\rho}{\sigma^3} f_{G_{(k)}}(x)\mathrm{d}x && (105)\\
&\ge \int_{|x-\omega_q|\le\delta} \Phi(y'/\sigma) f_{G_{(k)}}(x)\mathrm{d}x - o(1) && (106)\\
&\ge \int_{|x-\omega_q|\le\delta} \Phi\Bigg(\frac{1}{\sqrt{C_4 n}}\frac{\mathrm{E}[G_{\le\omega_q}]}{8}n\epsilon\Bigg) f_{G_{(k)}}(x)\mathrm{d}x - o(1) && (107)\\
&= (1-o(1))\int_{|x-\omega_q|\le\delta} f_{G_{(k)}}(x)\mathrm{d}x - o(1) && (108)\\
&= (1-o(1))(1-o(1)) - o(1) && (109)\\
&= 1-o(1). && (110)
\end{aligned}$$

The third inequality follows from (104) and Corollary 8. This shows that for any $\epsilon > 0$, we have $\mathrm{P}(\Xi_n \le \tau+\epsilon)$. Combining with the conclusion of Step-3, the proof of the theorem is now complete.

Appendix B. Proof of Corollary 2

Let $N_1, \ldots, N_n$ be independent and identically distributed $\mathrm{N}(0,1)$ random variables. We can set $\mathbf{W} = [N_1 \cdots N_n]^T / (\sum_{i=1}^n N_i^2)^{1/2}$. Let $|N_{i_1}| \le \cdots \le |N_{i_n}|$, where $i_1, \ldots, i_n$ is a permutation of $1, \ldots, n$. By definition, the random vector $\mathbf{W}_p$ is zero at all indices except at $\{i_j : j = \lceil nq \rceil + 1, \ldots, n\}$ where it equals $\mathbf{W}$. This implies $\mathbf{W}^T\mathbf{W}_\mathrm{p} = \sum_{j=\lceil nq\rceil+1}^n N_{i_j}^2 / \sum_{i=1}^n N_i^2$. The statement for $\mathbf{W}^T\mathbf{W}_\mathrm{p}$ now follows from Theorem 1 since $\mathrm{N}(0,1)^2$ is a Gamma random variable with shape $\frac{1}{2}$ and scale 2. The convergence of $\mathbf{W}^T\mathbf{W}_\mathrm{ee}$ is proved similarly.

Appendix C. Proof of Theorem 3

Let us first calculate and upper bound on the normalized FLOPs. It is easily seen that $|\mathbf{W}_\mathrm{ee}^T\mathbf{X}|^2$ is a Chi-squared random variable with 1 degree of freedom. We thus obtain $\mathrm{P}(|\mathbf{W}_\mathrm{ee}^T\mathbf{X}| \ge \tau) = \Gamma(\frac{1}{2}, \frac{\tau^2}{2})$, where $\Gamma(\cdot,\cdot)$ is the upper incomplete Gamma function. The normalized FLOPs is thus

$$\overline{\mu}_c' = (1-q)\Gamma\left(\frac{1}{2}, \frac{\tau^2}{2}\right) + \gamma\left(\frac{1}{2}, \frac{\tau^2}{2}\right), \tag{111}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete Gamma function. Further, we obtain

$$\overline{\mu}'_c = (1-q) + q\gamma\left(\frac{1}{2}, \frac{\tau^2}{2}\right) \tag{112}$$
$$= 1 - q + q\mathrm{erf}\left(\sqrt{\frac{\tau^2}{2}}\right) \tag{113}$$
$$\leq 1 - q + q\Bigg(1 - \underbrace{\sqrt{\frac{2e}{\pi}}}_{>1}\frac{\sqrt{\beta-1}}{\beta}e^{-\beta\tau^2/2}\Bigg) \tag{114}$$
$$\leq 1 - q\frac{\sqrt{\beta-1}}{\beta}e^{-\beta\tau^2/2} \tag{115}$$
$$\leq 1 - qe^{-\tau^2} \tag{116}$$

The upper bound on the error function in (114) follows from [5], and is valid for any $\beta > 1$. In (116), we substituted $\beta = 2$.

We now analyze the generalization error ϵ_c. We consider the following events:

- Let E_0 be the event where the decisions of the conditional perceptron and the teacher do not match so that $\epsilon_c = \mathrm{P}(E_0)$.
- We let E_1 denote the event that the student $\mathbf{W}$ and the teacher $\mathbf{T}$ are at least δ_1-close with respect to the angular distance. In other words, let E_1 denote the event that $\arccos \mathbf{W}^T\mathbf{T} \leq \delta_1$, where $\delta_1 \in [0, \frac{\pi}{2}]$.
- Let E_2 be the event that the student and the early exit vector (which are derived from the student weights) are δ_2-close. In other words, let E_2 represent the event $\arccos \mathbf{W}^{\dagger}\mathbf{W}_{\mathrm{ee}} \leq \delta_2$, where $\delta_2 \in [0, \frac{\pi}{2}]$.
- Finally, let E_3 be the event that $|\mathbf{X}^T\mathbf{W}_{\mathrm{ee}}| \geq \tau$, encoding the criterion of early exit in the definition of the conditional perceptron in Section 4.2.

We have

$$\epsilon_{\mathrm{c}} = \mathrm{P}(E_0E_1E_2) + \underbrace{\mathrm{P}(E_0|E_1^c \text{ or } E_2^c)}_{\leq 1}\underbrace{\mathrm{P}(E_1^c \text{ or } E_2^c)}_{\leq \mathrm{P}(E_1^c)+\mathrm{P}(E_2^c)} \tag{117}$$
$$= \mathrm{P}(E_0E_1E_2E_3) + \underbrace{\mathrm{P}(E_0E_1E_2E_3^c)}_{\leq \mathrm{P}(E_0E_3^c)\leq\epsilon_{\mathrm{uc}}} + \mathrm{P}(E_1^c) + \mathrm{P}(E_2^c) \tag{118}$$
$$= \epsilon_{\mathrm{uc}} + \mathrm{P}(E_1^c) + \mathrm{P}(E_2^c) + \mathrm{P}(E_0E_1E_2E_3) \tag{119}$$

Let $\overline{\epsilon_{uc}}$ denote the asymptotic $n, N_t \to \infty$ generalization error for the unconditional perceptron so that $\epsilon_{\mathrm{uc}} = \frac{1}{\pi}\mathrm{E}[\arccos \mathbf{W}^T\mathbf{T}] \to \overline{\epsilon_{uc}}$ as $n \to \infty$. Due to the self-averaging property learning [4], we have, in addition, the convergence in mean $\frac{1}{\pi}\arccos \mathbf{W}^T\mathbf{T} \to \overline{\epsilon_{uc}}$. Hence, if $\delta_1 > \pi\overline{\epsilon_{uc}}$, we have $P(E_1^c) \to 0$. With a similar argument, provided that $\delta_2 > \arccos\sqrt{\tau_q}$, we have $P(E_2^c) \to 0$ as a result of Corollary 2. What is left to analyze is the final term. By symmetry, we have

$$\mathrm{P}(E_0E_1E_2E_3) = 2\mathrm{P}(\mathbf{X}^T\mathbf{T} < 0, \arccos\mathbf{W}^{\dagger}\mathbf{T} \leq \delta_1, \arccos\mathbf{W}^T\mathbf{W}_{\mathrm{ee}} \leq \delta_2, \mathbf{X}^T\mathbf{W}_{\mathrm{ee}} \geq \tau). \tag{120}$$

Let us recall the triangle inequality for angular distances: For unit norm vectors a_1, a_2, a_3, we have $\arccos a_1^\dagger a_2 \leq \arccos a_1^\dagger a_3 + \arccos a_3^\dagger a_2$. Therefore,

$$\begin{aligned}
\mathrm{P}(E_0E_1E_2E_3) &\leq 2\mathrm{P}(\mathbf{X}^T\mathbf{T} < 0, \arccos\mathbf{T}^T\mathbf{W}_{\mathrm{ee}} \leq \delta_1 + \delta_2, \mathbf{X}^T\mathbf{W}_{\mathrm{ee}} \geq \tau) && (121)\\
&= 2\mathrm{P}(\mathbf{X}^T\mathbf{T} < 0, \mathbf{T}^T\mathbf{W}_{\mathrm{ee}} \geq \cos(\delta_1 + \delta_2), \mathbf{X}^T\mathbf{W}_{\mathrm{ee}} \geq \tau) && (122)\\
&\leq 2\mathrm{P}(\mathbf{X}^T\mathbf{a} < 0, \mathbf{X}^T\mathbf{b} \geq \tau) && (123)
\end{aligned}$$

where $\mathbf{a}$ and $\mathbf{b}$ are arbitrary unit-norm deterministic vectors with $\mathbf{a}^T\mathbf{b} = \cos(\delta_1 + \delta_2)$. The random variables $\mathbf{X}^T\mathbf{a}$ and $\mathbf{X}^T\mathbf{b}$ are jointly Gaussian with zero mean, unit variance, and covariance $\rho \triangleq \cos(\delta_1 + \delta_2)$. We can thus evaluate the joint probability as

$$\begin{aligned}
\mathrm{P}(\mathbf{X}^T\mathbf{a} < 0, \mathbf{X}^T\mathbf{b} \geq \tau) &= \int_\tau^\infty \int_{-\infty}^0 \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) \mathrm{d}x\mathrm{d}y && (124)\\
&= \frac{1}{2\sqrt{2\pi}} \int_\tau^\infty e^{-y^2/2} \mathrm{erfc}\left(\frac{\rho y}{\sqrt{2(1-\rho^2)}}\right) \mathrm{d}y && (125)
\end{aligned}$$

Using the upper bound $\mathrm{erfc} x \leq e^{-x^2}$, we obtain

$$\begin{aligned}
\mathrm{P}(\mathbf{X}^T\mathbf{a} < 0, \mathbf{X}^T\mathbf{b} \geq \tau) &\leq \frac{1}{2\sqrt{2\pi}} \int_\tau^\infty e^{-y^2/2} e^{-\frac{\rho^2 y^2}{2(1-\rho^2)}} \mathrm{d}y && (126)\\
&= \frac{1}{2\sqrt{2\pi}} \int_\tau^\infty e^{-\frac{y^2}{2(1-\rho^2)}} \mathrm{d}y && (127)\\
&= \frac{1}{4}\sqrt{1-\rho^2}\,\mathrm{erfc}\left(\frac{\tau}{\sqrt{2(1-\rho^2)}}\right) && (128)\\
&\leq \frac{1}{2} e^{-\frac{\tau^2}{2(1-\rho^2)}}. && (129)
\end{aligned}$$

A joint consideration with the previous bounds concludes the proof of the theorem.