Information-Theoretic Foundations for Neural Scaling Laws

Hong Jun Jeon
Department of Computer Science
Stanford University
Stanford, CA 94305
hjjeon@stanford.edu

Benjamin Van Roy
Stanford University
Stanford, CA 94305
bvr@stanford.edu

Abstract

Neural scaling laws aim to characterize how out-of-sample error behaves as a function of model and training dataset size. Such scaling laws guide allocation of a computation resources between model and data processing to minimize error. However, existing theoretical support for neural scaling laws lacks rigor and clarity, entangling the roles of information and optimization. In this work, we develop rigorous information-theoretic foundations for neural scaling laws. This allows us to characterize scaling laws for data generated by a two-layer neural network of infinite width. We observe that the optimal relation between data and model size is linear, up to logarithmic factors, corroborating large-scale empirical investigations. Concise yet general results of the kind we establish may bring clarity to this topic and inform future investigations.

1 Introduction

In recent years, foundation models have grown immensely, with some embodying trillions of trainable parameters. While larger models have in general produced better results, they also require much more compute to train. It has become impractical to perform hyperparameter sweeps at the scale of these modern models. This has required bypassing the practice of tuning hyperparameters via extensive trial and error, as was previously common in deep learning.

Among other things, hyperparameters control $1)$ the size, measured in terms of the parameter count p, of the neural network model and $2)$ the number T of training tokens. If each parameter is adjusted in response to each token then the computational requirements of training scale will the product of these two quantities. For any compute budget C, one should carefully balance between p and T. Too few training tokens leads to model estimation error, while too few parameters gives rise to mispecification error. As evaluating performance across multiple choices of p and T becomes computationally prohibitive at scale, alternative kinds of analysis are required to guide allocation of computational resources.

Kaplan et al. [2020] and Hoffmann et al. [2022a] have proposed the following procedure for allocating a large compute budget: $1)$ Evaluate test errors of models produced using various small compute budgets C with many different allocations to parameters p versus training tokens T. $2)$ Extrapolate to estimate the relation between p and T for large C.

To give a sense of scales involved here, Hoffmann et al. [2022a] evaluate test errors across “small” models for which $p \times T$ ranges from around 10^{18} to 10^{22} and extrapolates out to “large” models at around 10^{24}. Kaplan et al. [2020] and Hoffmann et al. [2022a] each extrapolate based on a hypothesized *scaffolding function*. Kaplan et al. [2020] guess a scaffolding function based on results

38th Conference on Neural Information Processing Systems M3L Workshop (NeurIPS 2024).

observed in small scale experiments. Hoffmann et al. [2022a] carry out an informal and somewhat speculative mathematical analysis to guide their choice (see their Appendix D).

The analysis of Hoffmann et al. [2022a] is somewhat generic rather than specialized to the particular neural network architecture used in that paper. In this paper, building on the work of Jeon and Van Roy [2022a,b], we develop rigorous information-theoretic foundations and use them to derive similar scaling laws. To keep things simple and concrete, we carry out the analysis with a particular data generating process for which neural networks are well-suited. The sorts of arguments developed by Hoffmann et al. [2022a] are just as relevant to this context as they are to language models.

Hoffmann et al. [2022a] suggest that the compute optimal trade-off between parameter count and number of training tokens is linear, though the authors expressed some doubt and considered other possibilities that are near-linear as well. We establish an upper bound on the minimal information-theoretically achievable expected error as a function of p and T and derive the relation required to minimize this bound for each compute budget. For large compute budgets, this relation is linear, as suggested by Hoffmann et al. [2022a].

Our main contributions include a first rigorous mathematical characterization of the compute-optimal efficient frontier for a neural network model and development of information-theoretic tools which enable that. A limitation of our analysis is in its simplified treatment of computational complexity as the product of the model and data set sizes; we do not assume any constraints on computation beyond those imposed by choices of p and T. In particular, we analyze, algorithms which carry out perfect Bayesian inference with respect to a model that is misspecifiecified due to its restricted size. While this abstracts away the details of practical training algorithms, empirical evidence suggests that our idealized framework leads to useful approximations [Zhu et al., 2022]. In spite of these limitations, we hope our results set the stage for further mathematical work to guide hyperparameter selection when training large neural networks.

2 A Framework for Learning

2.1 Probabilistic Framework

We define all random variables with respect to a common probability space $(\Omega, \mathbb{F}, \mathbb{P})$. Recall that a random variable F is simply a measurable function $\Omega \mapsto \mathcal{F}$ from the sample space Ω to an outcome set $\mathcal{F}$.

The probability measure $\mathbb{P} : \mathbb{F} \mapsto [0,1]$ assigns likelihoods to the events in the $\sigma-\text{algebra}\ \mathbb{F}$. For any event $E \in \mathbb{F}$, $\mathbb{P}(E)$ to denotes the probability of the event. For events $E, G \in \mathbb{F}$ for which $\mathbb{P}(G) > 0$, $\mathbb{P}(E|G)$ to denotes the probability of event E conditioned on event G.

For realization z of a random variable Z, $\mathbb{P}(Z = z)$ is a function of z. We denote its value evaluated at Z by $\mathbb{P}(Z)$. Therefore, $\mathbb{P}(Z)$ is a random variable (it takes realizations in $[0,1]$ depending on the value of Z). Likewise for realizations (y, z) of random variables Y, Z, $\mathbb{P}(Z = z|Y = y)$ is a function of (y, z) and $\mathbb{P}(Z|Y)$ is a random variable which denotes the value of this function evaluated at (Y, Z).

If random variable $Z : \Omega \mapsto \Re^K$ has density p_Z w.r.t the Lebesgue measure, the conditional probability $\mathbb{P}(E|Z = z)$ is well-defined despite the fact that for all z, $\mathbb{P}(Z = z) = 0$. If function $f(z) = \mathbb{P}(E|Z = z)$ and $Y : \Omega \mapsto \Re^K$ is a random variable whose range is a subset of Z's, then we use the $\leftarrow$ symbol with $\mathbb{P}(E|Z \leftarrow Y)$ to denote $f(Y)$. Note that this is different from $\mathbb{P}(E|Z = Y)$ since this conditions on the event $Z = Y$ while $\mathbb{P}(E|Z \leftarrow Y)$ indicates a change of measure.

2.2 Data

We consider a stochastic process which generates a sequence $(X_t, Y_{t+1} : t \in \mathbb{Z}_+)$ of data pairs. For all t, we let H_t denote the history $(X_0, Y_1, \ldots, X_{t-1}, Y_t, X_t)$ of experience. We assume that there exists an underlying latent variable F such that $(X_0, X_1, \ldots) \perp F$ and F prescribes a conditional probability measure $F(\cdot|H_t)$ to the next label Y_{t+1}. In the case of an *iid* data generating process, this conditional probability measure would only depend on H_t via X_t. Note that the current pre-training objective of foundation models falls under this iid setting in which for all t, X_t is a random segment of the training corpus and Y_{t+1} is the subsequent token. As our framework is Bayesian, we represent our uncertainty about F by modeling it as a random variable with prior distribution $\mathbb{P}(F \in \cdot)$.

2.3 A Learning Objective

We focus on a particular notion of error which facilitates analysis via Shannon-information theory and reflects the objective of modern foundation models. For all $t \in \mathbb{Z}_+$, our algorithm is tasked with providing a predictive distribution P_t of Y_{t+1} which may depend on the history of data which it has already observed H_t. We express such an algorithm as π for which $P_t = \pi(H_t)$. As aforementioned, an effective learning system ought to leverage data as it becomes available and perform well across all time. As a result, for any time horizon $T \in \mathbb{Z}_+$, we are interested in quantifying the cumulative expected log-loss:

$$\mathbb{L}_{T,\pi} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_\pi\left[-\ln P_t(Y_{t+1})\right].$$

Note that since we take all random variables to be defined with respect to a common probability space, the expectation $\mathbb{E}$ integrates over all random variables which we do not condition on. We use the subscript π in $\mathbb{E}_\pi$ to specify that all predictions P_t for all t are produced by π. As Y_{t+1} is the random variable which represents the next label that is generated by the underlying stochastic process, $P_t(Y_{t+1})$ denotes the probability that our algorithm's prediction P_t assigns to label Y_{t+1}.

It is important to note that even for an *omniscient* algorithm, the minimum achievable log-loss is not 0. Consider the *omniscient* algorithm which produces for all t the prediction $P^*_t = \mathbb{P}(Y_{t+1} \in \cdot | F, H_t)$. Even this agent incurs a loss of:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_\pi\left[-\ln \mathbb{P}(Y_{t+1}|F, H_t)\right] = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{H}(Y_{t+1}|F, H_t)$$

where our point follows from the fact that the conditional entropy $(\mathbb{H})$ of a discrete random variable Y_{t+1} is non-negative. As a result, we define the *reducible error* as:

$$\begin{aligned}\mathcal{L}_{T,\pi} &= \mathbb{L}_{T,\pi} - \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{H}(Y_{t+1}|F, H_t)\\ &= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(P^*_t(\cdot)\|P_t(\cdot)\right)\right].\end{aligned}$$

reducible error represents the error which is reducible via observing additional data and fitting a larger model. Therefore, we expect that this error will consist of two terms which reflect $1)$ the error due to estimation via finite data, $2)$ the error due to approximation with a finite parameter model.

3 Error of Constrained Predictors

We introduce a general upper bound on the reducible error of a *constrained* predictor. While the formulations remain abstract in this section, a useful running example is the following: Assume that F is an *infinite* width neural network which generates the data and $\tilde{F}$ is a *finite* width network.

3.1 A Constrained Predictor

F may exhibit endless complexity, likely beyond what can be represented with finite memory hardware. To represent the predictions made by a constrained predictor, we first define a random variable $\tilde{F}$ whose range is a subset of F's. As aforementioned, this random variable can be a lossy compression of F i.e. if F is represented by an infinite-width neural network, $\tilde{F}$ could be a finite-width approximation. For all t, let the constrained predictor be:

$$\tilde{P}_t(\cdot) = \sum_{\tilde{f}} \mathbb{P}(\tilde{F} = \tilde{f}|H_t) \cdot \mathbb{P}(Y_{t+1} \in \cdot|F = \tilde{f}, X_t).$$

The predictor performs inference on $\tilde{F}$ but performs predictions as if $F = \tilde{F}$. We let

$$\mathcal{L}_T(\tilde{F}) = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(P_t^*(\cdot)\|\tilde{P}_t(\cdot)\right)\right].$$

3.2 Error of Constrained Predictor

We now upper bound the error of this constrained predictor as a sum of mutual information (denoted $\mathbb{I}$) which represents estimation error and an expected KL divergence which corresponds to error due to misspecification.

Theorem 3.1. *For all $T \in \mathbb{Z}_{++}$ and random variables $F : \Omega \mapsto \mathcal{F}, \tilde{F} : \Omega \mapsto \tilde{\mathcal{F}}$ for which $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, if $((X_t, Y_{t+1}) : t \in \mathbb{Z}_+)$ is iid conditioned on F, then*

$$\tilde{\mathcal{L}}_T(\tilde{F}) \leq \frac{\mathbb{I}(F;\tilde{F})}{T} + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(P_t^*(\cdot)\|\hat{P}_t(\cdot)\right)\right],$$

where $\hat{P}_t(\cdot) = \mathbb{P}(Y_{t+1} \in \cdot | F \leftarrow \tilde{F}, X_t)$.

The first term denotes the *estimation error* or the error which is reducible via access to more data. This is evident by the fact it decreases linearly in T and the numerator reflects the *complexity* of $\tilde{F}$. The more nats of information that $\tilde{F}$ contains about the data stream, the more data will be required to arrive at a good predictor.

The second term denotes the *misspecification error* or the error which is reducible via a larger learning model. The closer that $\tilde{F}$ approximates F, the smaller the KL divergence between P_t^* and $\hat{P}_t$ will be. In the following section, we will use Theorem 3.1 to derive a concrete neural scaling law for an infinite-width neural network example.

We note that while the above result provides a clean decomposition into estimation and misspecification error, the result is but an *upper bound*. Notably, the inequality comes from an application of the log-sum inequality for which equality only holds when the misspecified predictions $\tilde{P}_t(\cdot)$ match the correctly specified predictions $\mathbb{P}(Y_{t+1} \in \cdot | H_t)$ almost surely. A future analysis which tightens this results or provides suitable lower bounds would strengthen the following analysis which derives optimal scaling laws with respect to this upper bound.

3.3 Scaling Law

For a FLOP constraint $C = p \cdot T$, it is clear that there is a tension between p and T in minimizing the upper bound in Theorem 3.1. This can be seen by first fixing a FLOP count C and substituting $T = C/p$. The upper bound becomes:

$$\frac{p \cdot \mathbb{I}(F;\tilde{F})}{C} + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(P_t^*(\cdot)\|\hat{P}_t(\cdot)\right)\right].$$

Note that the first term is *increasing* in p whereas the second term is *decreasing* in p. Therefore, under a fixed FLOP budget, the designer ought to select a value of p which effectively balances the two sources of error.

4 An Illustrative Example

4.1 Data Generating Process

The generating process is described by a neural network with d inputs, a single asymptotically wide hidden layer of ReLU activation units, and a linear output layer. We denote by F the associated mapping from input to output. Inputs and binary labels are generated according to $X_t \overset{iid}{\sim} \mathcal{N}(0, I_d)$ and $\mathbb{P}(Y_{t+1} = 1 | F, X_t) = \sigma(F(X_t))$ where σ denotes the sigmoid function.

As alluded to by the asymptotic width, F is a nonparametric model which we will outline now. Let $\bar{\theta}$ be distributed according to a Dirichlet process with base distribution $\mathrm{uniform}(\mathbb{S}^{d-1})$ and scale parameter K. Realizations of this Dirichlet process are probability mass functions on a countably infinite subset of $\mathbb{S}^{d-1}$. Let $\mathcal{W} = \{w \in \mathbb{S}^{d-1} : \bar{\theta}_w > 0\}$ denote this set. For all $w \in \mathcal{W}$,

$$\theta_w = \begin{cases} \bar{\theta}_w & \text{with probability } 1/2, \\ -\bar{\theta}_w & \text{otherwise.} \end{cases}$$

Finally, we have that

$$F(X_t) \;=\; \sqrt{K+1} \cdot \sum_{w \in \mathcal{W}} \theta_w \text{ReLU}\left(w^\top X_t\right).$$

Since $\mathcal{W}$ has countably infinite cardinality, F is characterized by a neural network with infinite width. We let $\theta = (\theta_w : w \in \mathcal{W})$ and $W = (w : w \in \mathcal{W})$ denote the weights of such neural network and hence

$$F(X_t) \;=\; \sqrt{K+1} \cdot \theta^\top \text{ReLU}(W X_t).$$

Note that the mean and variance structure satisfy

$$\mathbb{E}[F(X)] = 0, \qquad \mathbb{E}[F(X)^2] = 1/2.$$

Therefore, this model remains nontrivial as d and K grow as all of the above quantities are invariant of d and K.

4.2 Constrained Predictor

We will study the scaling law associated with a particular constrained predictor characterized by a neural network of width n. Let $\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n$ be distributed iid $\text{Categorical}(\bar{\theta})$, where $\mathcal{W}$ are the classes. For any $\epsilon > 0$, let $\mathbb{S}^{d-1}_\epsilon$ be an ϵ-cover w.r.t $\|\cdot\|_2$ and for all $i \in [n]$, let

$$\tilde{w}_{i,\epsilon} = \underset{v \in \mathbb{S}^{d-1}_\epsilon}{\arg\min} \|\tilde{w}_i - v\|_2^2.$$

Finally, let

$$\tilde{F}_{n,\epsilon}(X_t) = \frac{\sqrt{K+1}}{n} \cdot \sum_{i=1}^{n} \text{sign}\left(\theta_{\tilde{w}_i}\right) \text{ReLU}\left(\tilde{w}_{i,\epsilon}^\top X_t\right).$$

Let $\tilde{\theta} \in \Re^n$ is $(\text{sign}(\theta_{\tilde{w}_i})/n : i \in [n])$ and $\tilde{W}_\epsilon \in \Re^{n \times d}$ is $(\tilde{w}_{i,\epsilon} : i \in [n])$. Therefore,

$$\tilde{F}_{n,\epsilon}(X_t) = \sqrt{K+1} \cdot \tilde{\theta}^\top \text{ReLU}\left(\tilde{W}_\epsilon X_t\right).$$

We consider the performance of a constrained agent which for all t, produces the prediction $\tilde{P}_t(\cdot) =$

$$\sum_{\tilde{f}} \mathbb{P}(\tilde{F}_{n,\epsilon} = \tilde{f} | H_t) \cdot \mathbb{P}(Y_{t+1} \in \cdot | F = \tilde{f}, X_t).$$

Note that this agent performs inference on the constrained model $\tilde{F}_{n,\epsilon}$ and produces predictions about Y_{t+1} as if $\tilde{F}_{n,\epsilon}$ were the function F which produced the data.

4.3 Error Bound

We will now study the error incurred by the constrained predictor described above. We define

$$\tilde{\mathcal{L}}_{T,n,\epsilon} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\text{KL}}\left(\mathbb{P}\left(Y_{t+1} \in \cdot | \theta, X_t\right) \| \tilde{P}_t(\cdot)\right)\right]$$

as the loss of interest.

Theorem 4.1. *For all $n, K, T \in \mathbb{Z}_{++}$ and $\epsilon \geq 0$, if for all $t \in \{0, 1, 2, \ldots, T-1\}$, (X_t, Y_{t+1}) is generated by F, then*

$$\tilde{\mathcal{L}}_{T,n,\epsilon} \;\leq\; \underbrace{\frac{K \ln\left(1 + \frac{n}{K}\right) \cdot \left(\ln(2n) + d \ln\left(\frac{3}{\epsilon}\right)\right)}{T}}_{\text{estimation error}} + \underbrace{\frac{3K(1 + d\epsilon^2)}{n}}_{\text{misspecification error}}.$$

The estimation error represents the error which is incurred in the process of learning $\tilde{F}$ from H_T. Notably, this error decays linearly in T, but only depends *logarithmically* in n. The misspecification error represents the error which persists due to the fact that we approximate F via $\tilde{F}_{n,\epsilon}$. As a result, this error decreases with greater n and smaller ϵ, but is *independent* of T. If we let $\tilde{\mathcal{L}}_{T,n} = \inf_{\epsilon>0}\ \tilde{\mathcal{L}}_{T,n,\epsilon}$, then

Corollary 4.2. *For all $n \geq 3, K \geq 2, T \in \mathbb{Z}_{++}$, if for all $t \in \{0, 1, 2, \ldots, T-1\}$, (X_t, Y_{t+1}) is generated by F, then*

$$\tilde{\mathcal{L}}_{T,n} \ \leq\ \frac{dK \ln\left(1+\frac{n}{K}\right)\left(\ln(e36TK) + \frac{2}{d}\ln(2n)\right)}{2T} + \frac{3K}{n}.$$

4.4 Resulting Scaling Law

Corollary 4.2 provides an upper bound on loss which we conjecture to be tight within logarithmic factors. This upper bound characterizes how the loss ought to grow/decay as functions of the network width n, the dataset size T, and complexity of the datat generating process d, K. We can therefore *analytically* derive a compute-optimal allocation by selecting n and T which *minimizes* the *upper bound* subject to the FLOP budget: $d \cdot n \cdot T \leq C$. The following Theorem states the resulting compute-optimal allocation.

Theorem 4.3. (compute-optimal parameter count) *For all $d, K \in \mathbb{Z}_{++}$ and FLOP counts $C \in \mathbb{Z}_{++}$, if $K \geq 2, d \geq 3$, and n^* minimizes the upper bound of Corollary 4.2 subject to $d \cdot n \cdot T \leq C$, then*

$$d \cdot n^* = \tilde{\Theta}\left(\sqrt{C}\right).$$

We provide a proof in Appendix A.2. Note that $d \cdot n^*$ denotes the compute-optimal parameter count and hence, this result corroborates the insights of Hoffmann et al. [2022b] that, up to logarithmic factors, the optimal parameter count grows as square root of the FLOP count (equivalently *linearly* in the training dataset size).

5 Conclusion

Our results provide a first step in developing rigorous mathematics for the purposes of analyzing scaling laws for foundation models. We hope that this will inspire further theoretical research on the subject. Our analysis is based on an error upper bound and furthermore, our analysis restricts attention to single-hidden-layer feedforward neural networks. Generalizing the results to treat state-of-the-art architectures remains an open issue. Furthermore, we have only considered allocation of pretraining compute. State-of-the-art performance in modern application domains relies on subsequent fine-tuning (see, e.g., [Ziegler et al., 2019]) through reinforcement learning from human feedback. How best to allocate resources between pretraining and fine-tuning is another area that deserves attention. An information-theoretic framework that treats pretraining, fine-tuning, and decision making in a unified and coherent manner, perhaps in the vein of [Lu et al., 2021], might facilitate theoretical developments on this front.

References

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. v. d. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models, 2022a. URL https://arxiv.org/abs/2203.15556.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022b.

H. J. Jeon and B. Van Roy. An information-theoretic framework for deep learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022a.

H. J. Jeon and B. Van Roy. An information-theoretic framework for supervised learning, 2022b. URL https://arxiv.org/abs/2203.00246.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

X. Lu, B. Van Roy, V. Dwaracherla, M. Ibrahimi, I. Osband, and Z. Wen. Reinforcement learning, bit by bit, 2021. URL https://arxiv.org/abs/2103.04047.

Y. Zhu, H. J. Jeon, and B. Van Roy. Is stochastic gradient descent near optimal?, 2022. URL https://arxiv.org/abs/2209.08627.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences, 2019. URL https://arxiv.org/abs/1909.08593.

A Proofs of Theoretical Results

Theorem 3.1. *For all $T \in \mathbb{Z}_{++}$ and random variables $F : \Omega \mapsto \mathcal{F}, \tilde{F} : \Omega \mapsto \tilde{\mathcal{F}}$ for which $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, if $((X_t, Y_{t+1}) : t \in \mathbb{Z}_+)$ is iid conditioned on F, then*

$$\tilde{\mathcal{L}}_T(\tilde{F}) \leq \frac{\mathbb{I}(F; \tilde{F})}{T} + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(P^*_t(\cdot) \| \hat{P}_t(\cdot)\right)\right],$$

where $\hat{P}_t(\cdot) = \mathbb{P}(Y_{t+1} \in \cdot | F \leftarrow \tilde{F}, X_t)$.

Proof.

$$\begin{aligned}
&\tilde{\mathcal{L}}_T(\tilde{F}) \\
&= \frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(P^*_t(\cdot) \middle\| \sum_{\tilde{f} \in \tilde{\mathcal{F}}} \mathbb{P}(Y_{t+1} \in \cdot | F = \tilde{f}, H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | H_t)\right)\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; F | H_t) + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(Y_{t+1} \in \cdot | H_t) \middle\| \sum_{\tilde{f} \in \tilde{\mathcal{F}}} \mathbb{P}(Y_{t+1} \in \cdot | \mathcal{F} = \tilde{f}, H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | H_t)\right)\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; F | H_t) \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\sum_{\tilde{f} \in \tilde{\mathcal{F}}} \sum_{y \in \mathcal{Y}} \mathbb{P}(Y_{t+1} = y | H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | Y_{t+1} = y, H_t) \ln \frac{\mathbb{P}(Y_{t+1} = y | H_t)}{\sum_{\tilde{f} \in \tilde{\mathcal{F}}} \mathbb{P}(Y_{t+1} = y | F = \tilde{f}, H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | H_t)}\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; F | H_t) \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\sum_{\tilde{f} \in \tilde{\mathcal{F}}} \sum_{y \in \mathcal{Y}} \mathbb{P}(Y_{t+1} = y | H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | Y_{t+1} = y, H_t) \ln \frac{\sum_{\tilde{f} \in \tilde{F}} \mathbb{P}(Y_{t+1} = y | H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | Y_{t+1} = y, H_t)}{\sum_{\tilde{f} \in \tilde{\mathcal{F}}} \mathbb{P}(Y_{t+1} = y | F = \tilde{f}, H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | H_t)}\right] \\
&\overset{(a)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; F | H_t) \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\sum_{y \in \mathcal{Y}} \sum_{\tilde{f} \in \tilde{\mathcal{F}}} \mathbb{P}(Y_{t+1} = y | H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | Y_{t+1} = y, H_t) \ln \frac{\mathbb{P}(Y_{t+1} = y | H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | Y_{t+1} = y, H_t)}{\mathbb{P}(Y_{t+1} = y | F = \tilde{f}, H_t) \cdot \mathbb{P}(\tilde{F} = \tilde{f} | H_t)}\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; F | H_t) \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(\tilde{F} \in \cdot | Y_{t+1}, H_t) \| \mathbb{P}(\tilde{F} \in \cdot | H_t)\right)\right] + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(Y_{t+1} \in \cdot | H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | F \leftarrow \tilde{F}, H_t)\right)\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(Y_{t+1} \in \cdot | F, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | H_t)\right)\right] \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(Y_{t+1}; \tilde{F} | H_t) + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(Y_{t+1} \in \cdot | H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | F \leftarrow \tilde{F}, H_t)\right)\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(Y_{t+1}; \tilde{F} | H_t) + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(Y_{t+1} \in \cdot | F, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | F \leftarrow \tilde{F}, H_t)\right)\right] \\
&= \frac{\mathbb{I}(H_T; \tilde{F})}{T} + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(Y_{t+1} \in \cdot | F, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | F \leftarrow \tilde{F}, H_t)\right)\right],
\end{aligned}$$

where (a) follows from the log-sum inequality. □

A.1 Proof of Dirichlet Process Results

Lemma A.1. **(squared error upper bounds KL)** *For all real-valued random variables G and $\tilde{G}$, if Y is a binary random variable for which $\mathbb{P}(Y = 1|G) = \frac{1}{1+e^{-G}}$, then*

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot|G)\|\mathbb{P}(Y \in \cdot|G \leftarrow \tilde{G}))\right] \leq \mathbb{E}\left[\left(G - \tilde{G}\right)^2\right].$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot|G)\|\mathbb{P}(Y \in \cdot|G \leftarrow \tilde{G}))\right] &= \mathbb{E}\left[\frac{1}{1+e^{G}} \ln\left(\frac{1+e^{\tilde{G}}}{1+e^{G}}\right)\right] \\ &\quad + \mathbb{E}\left[\frac{1}{1+e^{-G}} \ln\left(\frac{1+e^{-\tilde{G}}}{1+e^{-G}}\right)\right] \\ &\overset{(a)}{\leq} \mathbb{E}\left[\left(G - \tilde{G}\right)^2\right]\end{aligned}$$

where (a) follows from the fact that for all $x, y \in \Re$, $\frac{1}{1+e^{x}} \ln\left(\frac{1+e^{y}}{1+e^{x}}\right) + \frac{1}{1+e^{-x}} \ln\left(\frac{1+e^{-y}}{1+e^{-x}}\right) \leq (x-y)^2$. □

Lemma A.2. *For all $d, n, N \in \mathbb{Z}_{++}$, if $X \sim \mathcal{N}(0, I_d)$, then*

$$\mathbb{E}\left[\left(F(X) - \tilde{F}_{n,0}(X)\right)^2\right] \leq \frac{K+1}{n}.$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\left(\theta^\top \mathrm{ReLU}(WX) - \tilde{\theta}^\top \mathrm{ReLU}(\tilde{W}X)\right)^2\right] &\overset{(a)}{\leq} \frac{K+1}{n^2} \cdot \mathbb{E}\left[\mathrm{ReLU}(\tilde{W}X)^\top \left(\tilde{\theta}\tilde{\theta}^\top\right) \mathrm{ReLU}(\tilde{W}X)\right] \\ &= \frac{K+1}{n^2} \cdot \mathbb{E}\left[\mathrm{ReLU}(\tilde{W}X)^\top I_n \mathrm{ReLU}(\tilde{W}X)\right] \\ &\leq \mathbb{E}\left[\frac{K+1}{n^2} \cdot \sum_{i=1}^{n} (\tilde{w}_{i,0}^\top X)^2\right] \\ &= \frac{K+1}{n}.\end{aligned}$$

where (a) follows from the fact that the two functions have equal conditional expectation conditioned on θ. □

Lemma A.3. *For all $d, n, K \in \mathbb{Z}_{++}$,*

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(Y \in \cdot|F, X)\|\mathbb{P}(Y \in \cdot|F \leftarrow \tilde{F}_{n,0}, X)\right)\right] \;\leq\; \frac{K+1}{n}.$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot|F, X)\|\mathbb{P}(Y \in \cdot|F \leftarrow \tilde{F}_{n,0}, X))\right] &\overset{(a)}{=} \mathbb{E}\left[\left(F(X) - \tilde{F}_{n,0}(X)\right)^2\right] \\ &\overset{(b)}{\leq} \frac{K+1}{n},\end{aligned}$$

where (a) follows from Lemma A.1, (c) follows from the fact that the distribution of θ is the limiting distribution $\lim_{N\to\infty}$ of a Dirichlet $[K/N, \ldots, K/N]$ random variable, (d) follows from the dominated convergence theorem, and (e) follows from Lemma A.2. □

Lemma A.4. *For all $d, n, K \in \mathbb{Z}_{++}$,*

$$\mathbb{E}\left[\left(F_{n,0}(X) - F_{n,\epsilon}(X)\right)^2\right] \;\leq\; \frac{d(K+1)\epsilon^2}{n}.$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\left(F_{n,0}(X) - F_{n,\epsilon}(X)\right)^2\right] &= \mathbb{E}\left[(K+1)\cdot\left\|\sum_{i=1}^{n}\tilde{\theta}_i\left(\text{ReLU}(\tilde{w}_{i,0}^\top X) - \text{ReLU}(\tilde{w}_{i,\epsilon}^\top X)\right)\right\|^2\right]\\ &= \mathbb{E}\left[(K+1)\cdot\sum_{i=1}^{n}\tilde{\theta}_i^2\cdot\left\|\left(\text{ReLU}(\tilde{w}_{i,0}^\top X) - \text{ReLU}(\tilde{w}_{i,\epsilon}^\top X)\right)\right\|^2\right]\\ &\leq \frac{d(K+1)\epsilon^2}{n}\end{aligned}$$

□

Lemma A.5. *For all $d, n, K \in \mathbb{Z}_{++}$ and $\epsilon \geq 0$,*

$$\mathbb{E}\left[\mathbf{d}_{\text{KL}}\left(\mathbb{P}(Y \in \cdot|F, X)\|\mathbb{P}(Y \in \cdot|F \leftarrow \tilde{F}_{n,\epsilon}, X)\right)\right] \;\leq\; \frac{3K(1+d\epsilon^2)}{n}.$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\mathbf{d}_{\text{KL}}\left(\mathbb{P}(Y \in \cdot|F, X)\|\mathbb{P}(Y \in \cdot|F \leftarrow \tilde{F}_{n,\epsilon}, X)\right)\right] &\overset{(a)}{\leq} \mathbb{E}\left[\left(F(X) - \tilde{F}_{n,\epsilon}(X)\right)^2\right]\\ &= \mathbb{E}\left[2\left(F(X) - F_{n,0}(X)\right)^2 + 2\left(F_{n,0}(X) - F_{n,\epsilon}(X)\right)^2\right]\\ &\overset{(b)}{\leq} \frac{2(K+1)}{n} + \frac{2d(K+1)\epsilon^2}{n}\\ &\leq \frac{3K(1+d\epsilon^2)}{n},\end{aligned}$$

where (a) follows from Lemma A.1 and (b) follows from Lemmas A.4 and A.5. □

Lemma A.6. (entropy upper bound) *For all $d, n, K \in \mathbb{Z}_{++}$ and $\epsilon > 0$,*

$$\mathbb{H}(\tilde{F}_{n,\epsilon}) \leq K\ln\left(1 + \frac{n}{K}\right)\cdot\left(\ln(2n) + d\ln\left(\frac{3}{\epsilon}\right)\right).$$

Proof.

$$\begin{aligned}\mathbb{H}(\tilde{F}_{n,\epsilon}) &\overset{(a)}{\leq} \mathbb{E}\left[\sum_{w\in\mathcal{W}}\mathbb{1}_{|\tilde{\theta}_w|>0}\cdot\left(\ln(2n) + d\ln\left(\frac{3}{\delta}\right)\right)\right]\\ &\overset{(b)}{\leq} K\ln\left(1 + \frac{n}{K}\right)\left(\ln(2n) + d\ln\left(\frac{3}{\delta}\right)\right)\end{aligned}$$

where (a) follows from the fact that the output weight can take on at most $2n$ different values $(-\frac{n\sqrt{K+1}}{n}, -\frac{(n-1)\sqrt{K+1}}{n}, \ldots, -\frac{\sqrt{K+1}}{n}, \frac{\sqrt{K+1}}{n}, \ldots \frac{n\sqrt{K+1}}{n})$ and the fact that $|\mathbb{S}_\epsilon^{d-1}| \leq (3/\delta)^d$ and (b) follows as a commonly known fact about the number of unique classes of a dirichlet-multinomial distribution. □

Theorem 4.1. *For all $n, K, T \in \mathbb{Z}_{++}$ and $\epsilon \geq 0$, if for all $t \in \{0, 1, 2, \ldots, T-1\}$, (X_t, Y_{t+1}) is generated by F, then*

$$\tilde{\mathcal{L}}_{T,n,\epsilon} \;\leq\; \underbrace{\frac{K\ln\left(1+\frac{n}{K}\right)\cdot\left(\ln(2n) + d\ln\left(\frac{3}{\epsilon}\right)\right)}{T}}_{\text{estimation error}} + \underbrace{\frac{3K(1+d\epsilon^2)}{n}}_{\text{misspecification error}}.$$

Proof. The result follows from Theorem 3.1 and Lemmas A.6 and A.5 □

Corollary 4.2. *For all $n \geq 3, K \geq 2, T \in \mathbb{Z}_{++}$, if for all $t \in \{0, 1, 2, \ldots, T-1\}$, (X_t, Y_{t+1}) is generated by F, then*

$$\tilde{\mathcal{L}}_{T,n} \leq \frac{dK \ln\left(1+\frac{n}{K}\right)\left(\ln(e36TK)+\frac{2}{d}\ln(2n)\right)}{2T} + \frac{3K}{n}.$$

Proof. The result holds from Theorem 4.1 by setting $\epsilon^2 = \frac{nK\ln(1+\frac{n}{K})}{4T(K+1)}$ and the fact that for $n \geq 3, K \geq 2$, $\frac{36T(K+1)}{nK\ln(1+n/K)} \leq 36KT$. □

A.2 Optimal Width

Theorem 4.3. **(compute-optimal parameter count)** *For all $d, K \in \mathbb{Z}_{++}$ and FLOP counts $C \in \mathbb{Z}_{++}$, if $K \geq 2, d \geq 3$, and n^* minimizes the upper bound of Corollary 4.2 subject to $d \cdot n \cdot T \leq C$, then*

$$d \cdot n^* = \tilde{\Theta}\left(\sqrt{C}\right).$$

Proof.

$$\begin{aligned}
n^* &= \underset{n \in \left[\frac{C}{d}\right]}{\arg\min} \frac{3K}{n} + \frac{K\ln\left(1+\frac{n}{K}\right)\cdot\ln(2n)}{t} + \frac{dK\ln\left(1+\frac{n}{K}\right)\left(1+\frac{1}{2}\ln(36KT)\right)}{t}; \ \text{ s.t. } n \cdot d \cdot t \leq C \\
&= \underset{n \in \left[\frac{C}{d}\right]}{\arg\min} \frac{3}{n} + \frac{\ln\left(1+\frac{n}{K}\right)\cdot\ln(2n)}{t} + \frac{d\ln\left(1+\frac{n}{K}\right)\left(1+\frac{1}{2}\ln(36KT)\right)}{t}; \ \text{ s.t. } n \cdot d \cdot t \leq C \\
&= \underset{n \in \left[\frac{C}{d}\right]}{\arg\min} \frac{3}{n} + \frac{nd\ln\left(1+\frac{n}{K}\right)\cdot\ln(2n)}{C} + \frac{nd^2\ln\left(1+\frac{n}{K}\right)\left(1+\frac{1}{2}\ln\left(\frac{36KC}{nd}\right)\right)}{C} \\
&\overset{(a)}{=} n \text{ s.t. } \frac{3}{n^2} = \frac{d\left(\ln\left(1+\frac{n}{K}\right)+\ln\left(1+\frac{n}{K}\right)ln(2n)+\frac{n}{K+n}\ln(2n)\right)}{C} \\
&\quad + \frac{d^2\ln\left(1+\frac{n}{K}\right)\ln\left(\frac{36KC}{nd}\right)}{2C} + \frac{d^2n}{C(n+K)} + \frac{nd^2\ln\left(\frac{36KC}{nd}\right)}{2C(n+K)} \\
&= n \text{ s.t. } C = \frac{dn^2\left(\ln\left(1+\frac{n}{K}\right)+\ln\left(1+\frac{n}{K}\right)ln(2n)+\frac{n}{K+n}\ln(2n)\right)}{3} \\
&\quad + \frac{d^2n^2\ln\left(1+\frac{n}{K}\right)\ln\left(\frac{36KC}{nd}\right)}{6} + \frac{d^2n^3}{3(n+K)} + \frac{n^3d^2\ln\left(\frac{36KC}{nd}\right)}{6(n+K)}.
\end{aligned}$$

where (a) follows from 1st order optimality conditions

Due to monotonicity, we can drive upper and lower bounds for the value of n via lower and upper bounds of the above RHS respectively.

We begin with the upper bound for n^*:

$$\begin{aligned}
n^* &\overset{(a)}{\leq} n \text{ s.t. } C = \frac{d^2n^2}{3} \\
&= \frac{\sqrt{3C}}{d}.
\end{aligned}$$

where (a) follows from the fact that $d^2n^2/3$ is a lower bound of the above RHS.

We now derive the lower bound for n^*:

$$\begin{aligned}
n^* &\overset{(a)}{\geq} n \text{ s.t. } C = dn^2\ln(2n)\ln(36KC) \\
&= \tilde{\Omega}\left(\frac{\sqrt{C}}{d}\right).
\end{aligned}$$

where (a) follows from the fact that $dn^2 \ln(2n) \ln(36KC)$ is an upper bound of the above RHS. The result follows. □

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We develop a rigorous information-theoretic framework to analyze scaling laws which provides clarity to the topic and corroborates empirical phenomena in the space.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We outline how the theoretical foundations of Hoffmann et al. and Kaplan et al. are ad-hoc and the purpose of our piece is to study the problem rigorously.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, assumptions are in the theorem statements and proofs are in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: No experiments in this paper.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: No code or data in this paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No experiments in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: No experiments in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No experiments in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper is purely mathematical so it does not have any of the ethical concerns listed in the NeurIPS code.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The work is mathematical aimed at understanding the phenomenon of neural scaling laws. Any societal impact would be many layers removed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not applicable.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.