GRADIENT ENTROPY DYNAMICS CHARACTERIZE MODEL ADAPTABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradients guide how a neural network learns, but can they also reveal how well a pretrained model will adapt to a downstream task? We uncover a systematic rise in the Rényi entropy of gradients across neuron units as task learning begins. This rise occurs when dominant gradient components diminish faster in relative terms, distributing gradient mass more evenly across units. Beyond this observation, we establish a theoretical connection between gradient Rényi entropy and adaptation dynamics, showing that higher gradient entropy yields a smaller upper bound on the remaining loss gap. This connection provides a theoretical basis for assessing adaptability through gradient entropy and motivates a Rényi-1/2 gradient-entropy score for ranking pretrained models, as well as a valley-restricted variant that limits the influence of high-gradient-energy regions. Experiments with CNN- and ViT-based models across ten downstream datasets show that the proposed score achieves stronger average ranking correlations than the compared transferability estimators. Even without labels, the method remains competitive with, and can outperform, most label-based alternatives. These findings highlight gradient entropy as an informative measure of pretrained-model adaptability to downstream tasks.

1 INTRODUCTION

Gradient signals reflect the interplay between input data and model properties, motivating their use in characterizing neural networks and learning tasks Jacot et al. (2018); Achille et al. (2019). Recent work further shows that the second moment of minibatch gradients induces an entropic loss term that can influence learning dynamics Ziyin et al. (2025). These findings motivate examining the distributional dynamics of gradients during learning.

In this work, we characterize the evolution of the Rényi entropy Rényi (1961) of the gradient distribution across neuron units during model training. We observe that, once the model begins to learn the task, as reflected by a sustained decrease in training loss, the gradient mass becomes increasingly dispersed across units and its entropy rises accordingly. To understand this phenomenon, we analyze the continuous-time dynamics of gradient entropy and show that entropy increases when units carrying larger gradient mass decay faster in relative terms. Intuitively, as meaningful learning takes place, optimization preferentially suppresses dominant gradient components, allowing the remaining gradient mass to spread more evenly across units. This systematic increase in gradient entropy suggests that it may reflect how effectively a model learns task-relevant patterns in the input data, raising a natural question: *does higher gradient entropy indicate that a model is better suited to a given dataset?*

To answer this question, we establish a theoretical connection between gradient entropy and adaptation dynamics through the effective dimension, defined as the exponential of Rényi entropy. Under a coordinate-wise smoothness condition motivated by the geometry of neural-network optimization Xie et al. (2025), we derive a one-step descent bound showing that the effective dimension controls the guaranteed decrease in loss. We then extend this result to the full adaptation trajectory under a Polyak–Łojasiewicz condition Dana et al. (2025), where a larger effective dimension yields a smaller upper bound on the remaining loss gap. These results imply that, once a pretrained model capture task-relevant patterns on the target data, its initial gradient entropy carries information about how effectively it can adapt. This makes the entropy-induced effective dimension a natural pre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

adaptation predictor of downstream transferability, suggesting whether a model is well suited to the downstream data.

Building on this connection, we propose an estimator guided by Rényi-$1/2$ gradient entropy as a transferability score for ranking pretrained models. To mitigate the influence of high-gradient-energy regions, which may not necessarily reflect task-relevant patterns, we further introduce a valley-restricted variant that focuses on relatively low-energy regions of the model. We evaluate our estimator on transferability benchmarks covering CNN- and ViT-based pretrained models across ten downstream image-classification datasets Shao et al. (2022); Khoba et al. (2025); Zheng et al. (2025). Without fine-tuning the candidate models, our gradient-entropy score achieve stronger average ranking correlations than the compared transferability estimators. Moreover, our framework naturally supports label-free estimation by computing gradient entropy with an unsupervised objective; despite using no target labels, the resulting score remains competitive with, and in our experiments can outperform, most label-based estimators.

Our contributions are summarized as follows:

- We identify gradient entropy growth during training, providing insight into model ranking for a given dataset.
- We theoretically connect the gradient effective dimension to loss convergence, motivating gradient entropy as a transferability signal.
- We propose gradient-entropy scores that improve pretrained-model ranking and support competitive label-free estimation.

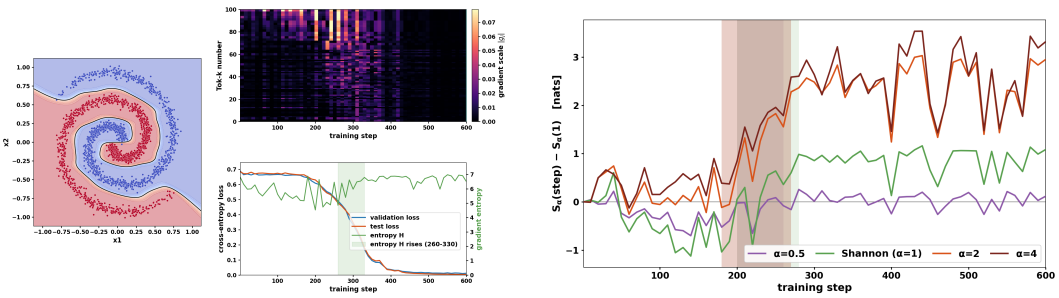
2 Realted Work

2.1 Gradient-Based Signals and Geometry

Gradient-based quantities have been widely used to characterize tasks, models, and learning dynamics. The neural tangent kernel (NTK) Jacot et al. (2018) relates parameter gradients to function-space optimization through a limiting kernel, while Task2Vec Achille et al. (2019) constructs task embeddings from Fisher information for task comparison and model selection. Beyond such gradient-based representations, recent work has studied how gradient statistics and geometry shape optimization itself. Ziyin et al. Ziyin et al. (2025) interpret learning from a thermodynamic perspective and show that minibatch-gradient statistics induce an entropic contribution to the training objective. Xie et al. Xie et al. (2025) analyze coordinate-wise optimization under ℓ_∞ smoothness, and Dana et al. Dana et al. (2025) establish gradient-flow convergence through a trajectory-dependent Polyak–Łojasiewicz analysis. In this paper, we instead focus on the distribution of gradient mass across neuron units, characterizing how its Rényi entropy evolves during learning and theoretically connecting the induced effective dimension to model adaptability and downstream transferability.

2.2 Transferability Estimation / Pretrained Model Ranking

Transferability estimation ranks pretrained models for downstream tasks without exhaustive fine-tuning. Early methods characterize source–target compatibility through conditional entropy (NCE Tran et al. (2019)) or expected empirical prediction (LEEP Nguyen et al. (2020)), while N-LEEP Li et al. (2021) replaces source predictions with Gaussian mixture assignments. Feature-based approaches assess transferability through statistical dependence (H-score Bao et al. (2019)), Bayesian label evidence (LogME You et al. (2021)), coding-rate-based mutual information (TransRate Huang et al. (2022)), or pairwise representation–annotation alignment (PARC Bolya et al. (2021)). Complementary approaches examine class geometry using Fisher discrimination (LDA Fisher (1936), SFDA Shao et al. (2022)), Gaussian class overlap (GBC Pándy et al. (2022)), and proximity to neural collapse (NCTI Wang et al. (2023)). ETran Gholami et al. (2023) combines energy-based distribution compatibility with task-specific scores. Moving beyond static feature assessment, PED Li et al. (2023) simulates representation dynamics through potential energy, while ITM Zheng et al. (2025) approximates embedding-space evolution using implicit modeling and variational approximation. Most existing methods rely on target labels to assess transferability from pretrained features or logits, whereas our approach can also be adapted to label-free settings.



(a) (b)

Figure 1: Figure 1a shows the toy experiment, which demonstrates an increase in the i entropy of gradients over the course of training; Figure 1b shows the entropy-increase phenomenon over different α in Rényi entropy. $S_\alpha(1)$ means the entropy of the gradients in the first step.

3 Gradient Entropy Dynamics During Learning

Gradient descent is often characterized as an increasingly concentrated process: as training progresses, the local loss geometry becomes more confined and gradient magnitudes decrease. However, our measurements of the Rényi entropy of the gradient distribution reveal a distinct phenomenon during the intermediate phase of training. As illustrated in Figure 2, the Rényi entropy of the gradients rises markedly in mid-training, pointing to a previously underexplored broadening of the gradient distribution during the learning stage.

In the toy experiment, a two-layer linear network is trained to classify the data points sampled from two distinct distributions, as shown in Figure 1a. As training proceeds and the loss steadily decreases, the Shannon entropy of the gradient distribution increases markedly. This observation demonstrates that gradient entropy does not necessarily decrease monotonically with optimization progress; instead, it can increase during the learning phase as the model adapts to the underlying task patterns. Consistently, as shown in Figure 1b, we observe the same qualitative behavior for Rényi entropy across different values of α, indicating that this entropy-increasing phenomenon persists across a broader family of entropy measures.

To understand the reason behind the entropy-increasing phenomenon, we formulate the questions as follows. Let $\mathcal{L}(X, \theta)$ be the training loss on dataset X with parameters $\theta \in \mathbb{R}^{N_g}$, and let

$$g(X,\theta) \;:=\; \nabla_\theta \mathcal{L}(X,\theta), \qquad H(X,\theta) \;:=\; \nabla^2_\theta \mathcal{L}(X,\theta),$$

abbreviated $g = \nabla_\theta \mathcal{L}$ and $H = \nabla^2_\theta \mathcal{L}$ when the arguments are clear. We write g_i for the gradient mass attached to the i-th *unit*: throughout, indices $i \in [N_g]$ range over neurons rather than raw coordinates, and g averages the gradients over the parameters, which aligns with the calculation in deep learning framework Paszke et al. (2019). The induced distribution over units is

$$p_i \;=\; \frac{g_i^2}{\|g\|_2^2}, \qquad \textstyle\sum_i p_i = 1. \tag{1}$$

The probability p_i measures the share of the gradient carried by neuron i. In our experiments, we utilize this probability to calculate the Rényi entropy as follows.

Definition 3.1. *The Rényi entropy of order $\alpha > 0$, $\alpha \neq 1$, of the gradient distribution equation 1 is*

$$S_\alpha(p) \;=\; \frac{1}{1-\alpha} \log\Big(\sum_i p_i^\alpha\Big),$$

Specifically, the Shannon entropy can be presented as $S_1(p) = \lim_{a\to 1} S_\alpha(p)$.

Definition 3.1 characterizes the entropy of the gradients at a fixed time point. To understand how this entropy evolves during learning, however, we need to consider the continuous-time evolution of the model parameters. Accordingly, we model the parameter trajectory as

$$\dot\theta \;:=\; \frac{\partial \theta}{\partial t} \;=\; -\,u, \tag{2}$$

where u can be considered as the velocity of the model update during the learning state. With the formulation in continuous time, we have the proposition as follows.

Proposition 3.2. *With Equation 2,* $\frac{\partial \log \|g\|_2^2}{\partial t} = -\frac{2\, g^\top H u}{g^\top g}$,

Proposition 3.2 exhibits the relative change rate of the gradient norm. When $u = g$, the change rate can be expressed as a Rayleigh quotient. Apparently, the change of the gradient norm depends on the relation between g and Hu. To simplify the notation, we let $c_i := \frac{(Hu)_i}{g_i}$. The mass-weighted mean is $\bar{c} := \sum_j p_j c_j = g^\top H u / g^\top g$. Then, we can determine the evolution of the gradient entropy in the following theorem.

Theorem 3.3. *Under Definition 3.1, for every* $\alpha > 0$*,*

$$\frac{dS_\alpha}{dt} = \frac{\alpha}{(\alpha-1)\sum_k p_k^\alpha} \sum_{i,j} p_i p_j \left(p_i^{\alpha-1} - p_j^{\alpha-1}\right)(c_i - c_j).$$

Consequently, if $(p_i - p_j)(c_i - c_j) \geq 0$ *for all* i, j*, then* $\frac{\partial S_\alpha(p)}{\partial t} \geq 0$*.*

The case $\alpha = 1$ follows by continuity.

Corollary 3.4 (*Shannon case*)**.** *Letting* $\alpha \to 1$, $\frac{dS_1}{dt} = 2\sum_i p_i (c_i - \bar{c}) \log p_i$.

Since $\frac{d}{dt} \log |g_i| = -c_i$, Theorem 3.3 states that the gradient entropy increases whenever units carrying more gradient mass decay faster in relative terms. Equivalently, the alignment condition $(p_i - p_j)(c_i - c_j) \geq 0$ is exactly a positive rank correlation between gradient magnitude and relative decay rate.

Figure 1a tracks the 100 largest gradient magnitudes throughout training. As these initially large gradients decay, coinciding with the onset of loss reduction, the gradient entropy rises. This is consistent with Theorem 3.3, and it raises a further question: if learning drives the gradient entropy upward, does a higher entropy indicate a better model for a given dataset? We discuss this question in the next section.

4 Gradient Entropy and Adaptation Convergence

4.1 One-step Entropy Boundary

The velocity u defines the direction and magnitude of the model update in continuous time, but in practice the update happens in discrete steps. To model the discrete update, we let $\Delta_t = \theta_{t+1} - \theta_t = -\rho_t v_t$, where ρ_t represents the step size in the direction v_t and $v_t = u_t / \|u_t\|$ is the normalized u at time t.

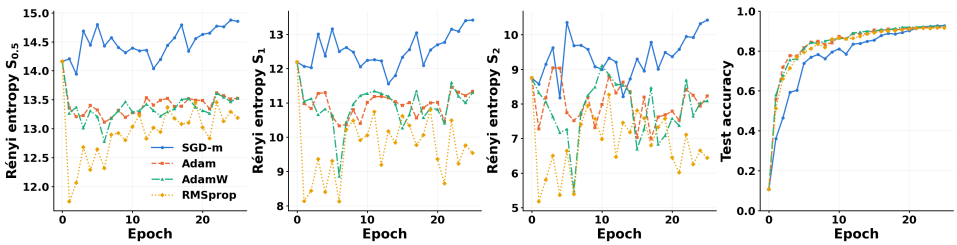
To preserve this unit-wise structure when moving from infinitesimal dynamics to finite updates, we explore the coordinate-wise ℓ_∞-geometry based on previous observation Xie et al. (2025). Specifically, we assume that the update of the model is coordinate-wise smoothness with the following condition:

Assumption 4.1. *There is* $\vec{L} = (L_1, \ldots, L_{N_g})$ $\vec{L} = (L_1, \ldots, L_{N_g})$ *satisfied that* $|\nabla_i \mathcal{L}(\theta + \Delta) - \nabla_i \mathcal{L}(\theta)| \leq L_i \|\Delta\|_\infty$.

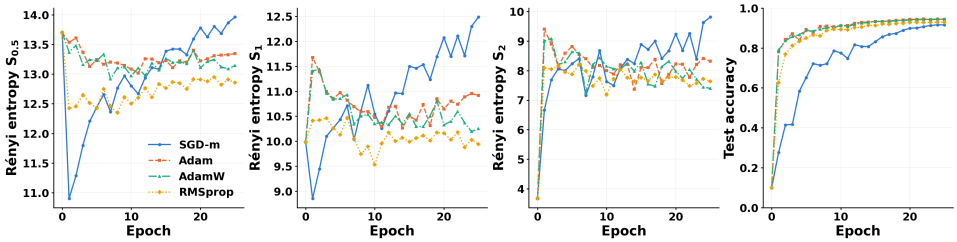
Assumption 4.1 keeps the local optimization geometry aligned with the neuron-wise representation used to define the gradient distribution. This is also consistent with Section 3, where entropy evolves because different gradient coordinates decay at different relative rates. With the duality between ℓ_∞ and ℓ_1 spaces, we have the following proposition.

Proposition 4.2. *For the gradient* g_t, $\operatorname{sign}(g_t) \in \arg\max_{\|v\|_\infty \leq 1} g_t^\top v$.

For the finite update $\Delta_t = -\rho_t v_t$, the first-order loss change is $-\rho_t g_t^\top v_t$. Therefore, for a fixed update radius ρ_t, Proposition 4.2 shows that we can choose $v_t = \operatorname{sign}(g_t)$ which maximizes the first-order loss decrease under the ℓ_∞ constraint. We then obtain the following theorem.



(a) Gradient entropy dynamics during Resnet-18 training on CIFAR-10 from scratch



(b) Gradient entropy dynamics during Resnet-18 training on CIFAR-10 from ImageNet pretrained

Figure 2: Figure 2a shows the whole-model gradient Rényi entropy S_α ($\alpha \in \{0.5, 1, 2\}$) and test accuracy of ResNet-18 trained from scratch on CIFAR-10 with different optimizers, and Figure 2b shows the same quantities for an ImageNet-pretrained Resnet-18 on CIFAR-10. The two rows of figures demonstrate the increase in gradient entropy across different optimizers. Especially for the pretrained model, the increase happens apparently at the initial stage.

Theorem 4.3. *Suppose that Assumption 4.1 holds, and let* $L_\infty := \sum_{i=1}^{N_g} L_i$. *and* $v_t^\star = \operatorname{sign}(g_t)$ *be the normalized* ℓ_∞*-steepest direction given by Proposition 4.2, and consider the finite update*$\Delta_t = -\rho_t v_t^\star$. *Then, we have the optimal step size* $\rho_t^\star = \frac{\|g_t\|_1}{L_\infty}$ *and*

$$\mathcal{L}(\theta_t + \Delta_t^\star) \le \mathcal{L}(\theta_t) - \frac{\|g_t\|_2^2}{2L_\infty} \exp\big(S_{1/2}(p_t)\big)\,,$$

where $\Delta_t^\star = -\frac{\|g_t\|_1}{L_\infty} \operatorname{sign}(g_t)$. *Equivalently,*

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta_t + \Delta_t^\star) \ge \frac{\|g_t\|_2^2}{2L_\infty} \exp\big(S_{1/2}(p_t)\big)\,.$$

Theorem 4.3 indicates that, for fixed gradient magnitude $\|g_t\|_2$ and coordinate-wise curvature scale L_∞, a larger Rényi-$1/2$ gradient entropy implies a larger guaranteed local loss reduction under coordinate-wise adaptation. However, the bound still depends on the gradient magnitude $\|g_t\|_2^2$, and therefore does not by itself characterize how the loss evolves over an adaptation trajectory. To obtain a trajectory-level statement, we need to relate the gradient magnitude at the current iterate to the remaining optimization gap. We discuss this in the following.

4.2 Trajectory-level Entropy Boundary

As discussed in previous work Dana et al. (2025), a sufficient condition for global convergence is that the gradient norm remains bounded from below relative to the current loss gap, which is known as Polyak–Łojasiewicz (PL) condition. To characterize the optimization behavior of a model that continues to make progress along the adaptation trajectory, we therefore adopt the following assumption.

Assumption 4.4 (*Polyak–Łojasiewicz*)**.** *Along the adaptation trajectory, there exists a constant* $\mu > 0$ *such that*

$$\frac{1}{2}\|g_t\|_2^2 \ge \mu\,(\mathcal{L}(\theta_t) - \mathcal{L}^\star)\,, \qquad t = 0, \dots, T-1.$$

This assumption gives us a connection between different optimization steps, which leads to the following theorem.

Theorem 4.5. *Suppose that Assumptions 4.1 and 4.4 hold along the adaptation trajectory. At each iteration t, let the local update characterized in Theorem 4.3. Then the loss gap satisfies*

$$\mathcal{L}(\theta_T) - \mathcal{L}^\star \leq (\mathcal{L}(\theta_0) - \mathcal{L}^\star) \exp\left(-\frac{\mu}{L_\infty} \sum_{t=0}^{T-1} \exp\big(S_{1/2}(p_t)\big) \right).$$

Theorem 4.5 indicates that the gradient entropy along the trajectory controls the loss drop rate. If the model has already exhibited meaningful information for downstream learning, as we discussed in Section 3, the gradient may tends to increase. This situation is very common when we use pretrained models, which can be shown in Figure 2b. Thus, the entropy bound can be updated as follows.

Corollary 4.6. *Suppose that the conditions of Theorem 4.5 hold and that*

$$S_{1/2}(p_t) \geq S_{1/2}(p_0), \qquad t = 0, \ldots, T-1.$$

Then

$$\mathcal{L}(\theta_T) - \mathcal{L}^\star \leq \big(\mathcal{L}(\theta_0) - \mathcal{L}^\star\big) \exp\left(-\frac{\mu T}{L_\infty} \exp\big(S_{1/2}(p_0)\big) \right).$$

Corollary 4.6 connects the initial gradient entropy to the loss gap after adaptation, providing a pre-adaptation signal of the model's subsequent optimization behavior on the target task. This is closely aligned with the objective of model selection and transferability estimation, which seek to predict how well a pretrained model will adapt without carrying out full fine-tuning. Based on this theoretical connection, we introduce a gradient-entropy-based transferability score in the following section.

5 Experimental Analysis on Model Ranking

5.1 Method for Model Ranking

The analysis in Section 4 establishes a connection between the gradient entropy measured at a pretrained checkpoint and its subsequent adaptation behavior on a target task. In particular, a larger Rényi-$1/2$ gradient entropy leads to a larger guaranteed local loss reduction and, under the trajectory-level entropy boundary in Section 4, a faster contraction of the adaptation loss gap. Importantly, the gradient entropy can be evaluated at the initial checkpoint without performing full fine-tuning, making it a natural pre-adaptation signal for model selection. Thus, we use the exponentiated gradient entropy as a transferability criterion:

$$\text{score} = e^{S_{1/2}(p_0)} \tag{3}$$

Equation 3 corresponds to an effective dimension of the gradient distribution. A larger value indicates that the gradient mass is distributed over more units rather than concentrated on a small subset. In our experiments, we both evaluate this score with both supervised and unsupervised loss Oord et al. (2018).

Valley-restricted gradient entropy. The analysis in Section 3 further shows that gradient entropy increases when units carrying larger gradient mass decay faster during learning. This suggests that the entropy signal is particularly informative after the model has already captured meaningful task-related structure. In a pretrained deep network, however, gradient energy can vary substantially across different parts of the model, and high-energy regions may dominate the full-network gradient distribution. We therefore introduce a valley-restricted variant that focuses on regions with relatively low gradient energy.

Specifically, we partition the model into K non-overlapping parts according to their forward order. Let $\mathcal{B}_k$ denote the set of gradient units contained in the k-th part. We first compute the gradient energy of each part as

$$E_k = \|g_k\|_2^2, \qquad k = 1, \ldots, K, \tag{4}$$

where g_k denotes the gradient associated with $\mathcal{B}_k$. We then rank the K parts according to their gradient energies. Let π denote the resulting permutation such that $E_{\pi(1)} \leq E_{\pi(2)} \leq \cdots \leq E_{\pi(K)}$. The q parts with the smallest gradient energies are selected, and their union is denoted by $\mathcal{G} =$

$\bigcup_{r=1}^{q} \mathcal{B}_{\pi(r)}$. Finally, we restrict the gradient distribution to the selected region $\mathcal{G}$, renormalize the gradient mass within this region, and compute the Rényi entropy to obtain the transferability score:

$$\text{score}_{\text{valley}} = e^{S_{1/2}(p_0^{\mathcal{G}})}. \tag{5}$$

The complete procedure is summarized in Algorithm 1. In our experiments, we evaluate the ImageNet-pretrained models with Equation 5 as the measurement function on model ranking

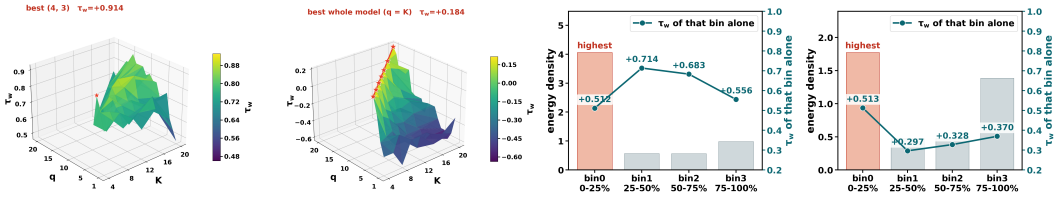
5.2 Experiment Setting

Data. We use two transferability-estimation benchmarks that share ten target datasets: CIFAR-10 Krizhevsky et al. (2009), CIFAR-100 Krizhevsky et al. (2009), DTD Cimpoi et al. (2014), Oxford Pets Parkhi et al. (2012), Flowers-102 Nilsback & Zisserman (2008), FGVC-Aircraft Maji et al. (2013), Caltech-101 Fei-Fei et al. (2004), Stanford Cars Krause et al. (2013), Food-101 Bossard et al. (2014) and SUN397 Xiao et al. (2010).

Models. The **CNN benchmark** Shao et al. (2022); Khoba et al. (2025) contains eleven supervised ImageNet CNNs (ResNet-34/50/101/152 He et al. (2016), DenseNet-121/169/201 Huang et al. (2017), MobileNetV2 Sandler et al. (2018), MnasNet-1.0 Tan et al. (2019), GoogLeNet Szegedy et al. (2015), Inception-v3 Szegedy et al. (2016)), with penultimate features as input. The **ViT benchmark** Zheng et al. (2025) contains six self-supervised ViTs (DINO ViT-S/8 and ViT-B/16 Caron et al. (2021), MoCo-v3 ViT-B/16 Chen et al. (2021), MAE ViT-B/16 and ViT-L/16 He et al. (2022), SimMIM ViT-B/16 Xie et al. (2022)), with backbone outputs as input. Ground truth is the fine-tuning accuracy released with each benchmark.

Setting and protocol All scores are computed on the training split. For each dataset, we correlate the per-model scores with the fine-tuning accuracies and report the weighted Kendall's τ_w and Spearman's ρ, averaged over the ten shared datasets. Every method is implemented once and applied to both benchmarks, and each method receives an independent copy of the features.

Compared methods We compare against ten label-based methods, taking the union of the methods reported by the two benchmarks and evaluating each on both: LogME You et al. (2021), SFDA Shao et al. (2022), NCTI Wang et al. (2023), GBC Pándy et al. (2022), LDA Fisher (1936), N-LEEP Li et al. (2021), PARC Bolya et al. (2021), PED Li et al. (2023), ETran Gholami et al. (2023), ITM **?**.



(a) CNN K-q search (b) ViT K-q search (c) CNN τ_w of each bin (d) ViT τ_w of each biny

Figure 3: Figures 3a and 3b show the ablation study of K and q on CNN models and ViT models respectively. CIFAR-10 is used for the ablation study; Figures 3c and 3d show the τ_w value of each bin on CNN and ViT models, respectively. The gradient energy density is also presented.

5.3 Analysis of K and q

Figure 3 presents the hyperparameter search over K and q on CIFAR-10. For CNN models, the best performance is achieved at $(K, q) = (4, 3)$, whereas for transformer-based models, using the gradients of the entire model works best. As shown in Figure 3a, performance generally improves as the ratio q/K increases, but drops sharply once the gradients of the whole model are used. Figure 3c further shows that, for CNN models, the most informative gradients reside mainly in the middle layers, while the segment with the highest energy density yields the lowest ranking performance. For ViT models, this trend is reversed. We hypothesize that this is because the gradient energy of ViT models is considerably lower in scale than that of CNN models, so that all parts of the model may carry meaningful information. This would explain why ViT models benefit from using the full-model gradients, while CNN models do not.

5.4 RESULTS ON TRANSFERABILITY ESTIMATION BENCHMARKS

Table 1: Transferability estimation performance on CNN-based models. We report the weighted Kendall's rank correlation coefficient (τ_w) and Spearman correlation (ρ) between predicted transferability scores and downstream performance across ten target datasets, with higher values indicating better transferability estimation. For the label-free setting, contrastive loss is employed as the unsupervised objective.

τ_w / ρ	CIFAR10	CIFAR100	DTD	Pets	Flowers	Aircraft	Cal101	Cars	Food	SUN397	Avg
LDA	0.842/0.864	0.869/0.888	0.319/0.387	0.713/0.873	0.218/0.345	-0.185/-0.127	0.548/0.436	0.152/0.127	0.754/0.727	0.547/0.564	0.478/0.508
LogME	0.852/0.891	0.725/0.702	0.700/0.733	0.411/0.564	0.147/0.309	0.439/0.264	0.463/0.664	0.605/0.482	0.385/0.382	0.511/0.373	0.524/0.536
N-LEEP	0.819/0.791	0.748/0.866	0.832/0.916	0.840/0.909	0.230/0.409	-0.289/-0.209	0.523/0.673	0.354/0.555	0.707/0.655	0.875/0.909	0.564/0.647
PARC	-0.196/-0.391	0.074/-0.009	0.628/0.683	0.044/0.055	0.055/0.291	-0.140/0.045	0.573/0.709	0.821/0.873	-0.511/-0.618	-0.560/-0.745	0.079/0.089
SFDA	0.949/0.964	0.866/0.879	0.575/0.642	0.522/0.655	0.492/0.520	0.614/0.636	0.363/0.618	0.518/0.709	0.815/0.827	0.558/0.700	0.627/0.715
GBC	0.735/0.682	0.664/0.733	0.703/0.670	0.514/0.618	0.214/0.455	0.423/0.536	0.269/0.591	0.567/0.600	0.548/0.655	0.324/0.591	0.496/0.613
NCTI	0.843/0.827	0.879/0.897	0.704/0.815	0.912/0.973	0.541/0.582	0.496/0.691	0.611/0.618	0.647/0.736	0.773/0.809	0.756/0.827	0.716/0.778
PED	0.604/0.782	0.546/0.788	0.442/0.547	0.594/0.745	0.329/0.347	0.429/0.636	0.359/0.600	-0.004/0.045	0.719/0.818	-0.065/-0.009	0.395/0.530
ETran	0.887/0.909	0.900/0.934	0.360/0.410	0.713/0.873	0.668/0.618	-0.091/-0.018	0.622/0.500	0.307/0.400	0.829/0.818	0.716/0.809	0.591/0.625
ITM	0.800/0.845	0.756/0.879	0.366/0.653	0.728/0.691	0.644/0.788	0.410/0.642	0.455/0.600	0.501/0.691	0.758/0.791	0.805/0.864	0.622/0.744
Gradient Entropy	0.914/0.936	0.598/0.743	0.854/0.920	0.794/0.864	0.412/0.564	0.806/0.818	0.646/0.882	0.757/0.773	0.821/0.891	0.663/0.809	0.727/0.820
Gradient Entropy (Label-free)	0.758/0.754	0.746/0.774	0.787/0.857	0.794/0.864	0.238/0.364	0.623/0.654	0.758/0.927	0.826/0.855	0.684/0.864	0.798/0.864	0.701/0.778

CNN benchmark. Table 1 shows the comparison with state-of-the-art (SOTA) methods on CNN models as we discussed above. Our score reaches $\tau_w = 0.727$ and spearman correlation $\rho = 0.82$ with the supervised probe, ahead of NCTI ($\tau_w = 0.716, \rho = 778$), SFDA ($\tau_w = 0.627, \rho = 715$), ITM ($\tau_w = 0.622, \rho = 0.744$), ETran ($\tau_2 = 0.591, \rho = 0.625$) and LogME ($\tau_2 = 0.524, \rho = 0.536$). The label-free variant reaches $\tau_w = 0.701$ and $\rho = 0.778$, outperforming most state-of-the-art methods while using no target labels at all.

ViT benchmark Table 2 presents a comparison with state-of-the-art methods on self-supervised ViT models. This setting is more representative of practical applications, as recent foundation models Siméoni et al. (2025); Kirillov et al. (2023) already exploit strong pretrained transformer-based models for a wide range of downstream tasks. Consequently, selecting an appropriate foundation model for a given downstream task has become increasingly important in real-world deployment. As the table shows, our gradient entropy attains 0.527 and the label-free variant 0.503, improving on the best baseline by 0.122 and 0.098, respectively. Even in the label-free setting, gradient entropy demonstrates stronger predictive capability for model selection.

The two tables demonstrate the robustness of gradient entropy across different model architectures. Moreover, unlike previous methods, gradient entropy does not rely on label information, making it more broadly applicable to real-world scenarios where labels may be unavailable or costly to obtain.

Table 2: Transferability estimation results on self-supervised ViT models, measured by the weighted Kendall's rank correlation coefficient (τ_w) and Spearman correlation (ρ) between predicted transferability scores and downstream performance across ten target datasets. Higher values indicate better transferability estimation. Contrastive loss is used as the unsupervised objective in the label-free setting.

τ_w / ρ	Cal101	Cars	CIFAR100	CIFAR10	DTD	Aircraft	Flowers	Food	Pets	SUN397	Avg
LDA	0.249/-0.029	0.008/-0.143	-0.105/-0.371	0.533/0.371	0.351/0.086	0.117/0.086	-0.010/-0.086	0.333/0.429	0.184/0.200	0.117/0.086	0.178/0.063
LogME	0.790/0.886	0.116/0.486	0.532/0.771	0.620/0.771	0.490/0.714	0.456/0.657	0.755/0.943	0.057/0.257	0.254/0.429	-0.020/0.429	0.405/0.634
N-LEEP	0.499/0.486	0.340/0.600	0.396/0.600	0.267/0.600	0.586/0.714	0.456/0.657	0.341/0.714	0.314/0.371	0.318/0.429	0.396/0.600	0.391/0.577
PARC	0.249/0.371	0.340/0.600	-0.020/0.429	0.518/0.771	0.646/0.771	0.456/0.657	0.401/0.771	0.314/0.371	0.380/0.543	0.396/0.600	0.368/0.589
SFDA	0.679/0.600	-0.140/-0.086	0.205/0.543	0.865/0.829	0.367/0.429	0.003/0.200	-0.408/-0.577	0.200/0.543	0.592/0.829	0.173/0.257	0.254/0.357
GBC	0.927/0.943	0.150/0.486	0.592/0.829	0.682/0.886	0.532/0.771	0.430/0.657	0.600/0.829	-0.257/-0.086	0.064/0.314	0.260/0.486	0.398/0.611
NCTI	0.561/0.600	0.139/0.086	0.090/0.429	0.865/0.829	0.456/0.657	0.390/0.600	0.518/0.771	0.135/0.086	0.380/0.543	0.456/0.657	0.399/0.526
PED	0.029/0.086	-0.535/-0.657	0.027/0.257	0.682/0.886	0.177/0.143	-0.244/-0.371	0.085/-0.131	0.476/0.486	-0.441/-0.543	-0.244/-0.371	0.001/-0.022
ETran	0.008/-0.143	0.139/0.086	-0.151/-0.143	0.261/0.257	0.445/0.257	0.248/0.314	-0.200/-0.143	0.456/0.657	0.314/0.371	0.248/0.314	0.177/0.183
ITM	0.404/0.543	0.116/0.486	0.396/0.600	0.620/0.771	0.265/0.600	0.592/0.829	0.411/0.638	0.248/0.314	0.518/0.771	0.456/0.657	0.403/0.621
Gradient Entropy	0.755/0.943	0.484/0.371	0.729/0.771	0.184/0.200	0.682/0.886	0.600/0.429	0.189/0.486	0.371/0.200	0.600/0.429	0.676/0.657	0.527/0.537
Gradient Entropy (Label-free)	0.682/0.886	0.254/0.429	0.267/0.600	0.690/0.543	0.682/0.886	0.676/0.657	0.414/0.600	0.188/0.200	0.414/0.600	0.762/0.829	0.503/0.623

5.5 ANALYSIS OF α.

Our model-ranking score adopts $\alpha = 0.5$, motivated by the assumptions in the ℓ_2 space. More generally, varying α can be interpreted as modifying the underlying norm-space geometry associated with this assumption. Figure 4 evaluates the effect of different values of α on transferability estimation benchmarks. As shown in the figure, performance generally deteriorates as α deviates from

0.5, with the peak of the curve occurring around $\alpha = 0.5$ in most cases. This empirical observation is consistent with our theoretical motivation for choosing $\alpha = 0.5$. In addition, across the evaluated values of α, the Rényi-entropy-based score consistently yields positive Kendall's τ_w and Spearman's ρ. This indicates that gradient entropy maintains a positive rank correlation with downstream model performance across different choices of α, suggesting that the proposed criterion is robust to the specific entropy parameterization.

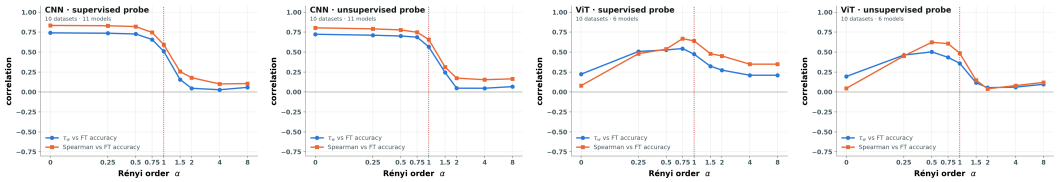


Figure 4: Transferability of the gradient Rényi effective-dimension score across α. At $\alpha = 0$, the score counts the number of non-zero coordinates; at $\alpha = 1$, it reduces to Shannon entropy (red dotted line). Curves show the weighted Kendall correlation τ_w and Spearman correlation ρ, averaged across datasets. **Left two panels:** 11 CNNs evaluated on 10 datasets. **Right two panels:** 6 self-supervised learning (SSL) ViTs evaluated on 10 datasets.

6 Discussion and Conclusion

Discussion on distribution shift Data distribution shift is one of the key factors underlying the performance gap across target domains. This phenomenon has been extensively observed in out-of-distribution (OOD) generalization studies Gulrajani & Lopez-Paz (2021). For example, on the PACS dataset Li et al. (2017), ImageNet-pretrained ResNet models generally achieve stronger performance on the Photo and Art domains, which are visually closer to the ImageNet pretraining distribution, while exhibiting a substantial performance drop on the more distinct Sketch domain. As shown in Figure 5, gradient entropy follows a consistent pattern across these domains: domains with better downstream performance tend to exhibit more favorable gradient-entropy responses. This observation suggests that gradient entropy is sensitive to the degree of alignment between pretrained representations and the target-domain distribution, indicating that gradient entropy may implicitly capture source–target compatibility. Therefore, gradient entropy may serve as a useful signal for characterizing the extent of distribution shift and, more broadly, for informing the analysis and development of methods aimed at improving robustness and generalization under distribution shifts.

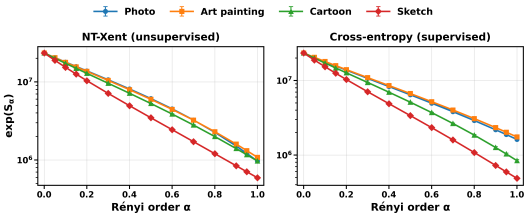


Figure 5: Effective dimension e^{S_α} of the gradient energy distribution from ImageNet-pretrained ResNet-50 on each PACS domain.

Conclusion. In this work, we study model transferability through the lens of gradient entropy. We first characterize the increase of gradient entropy during learning and show that it arises when units with larger gradient mass decay faster in relative terms. We then establish a theoretical connection between Rényi-$1/2$ gradient entropy and model adaptation through both one-step and trajectory-level descent bounds. Motivated by this connection, we develop a gradient-entropy-based transferability estimator that ranks pretrained models without full fine-tuning. Experiments on CNN and self-supervised ViT benchmarks demonstrate strong and consistent model-ranking performance across diverse downstream tasks. Importantly, the proposed estimator can also operate in a label-free setting using an unsupervised objective, while remaining competitive with label-based alternatives. Overall, our results suggest that gradient structure provides a useful pre-adaptation signal for understanding model adaptability and enabling efficient pretrained-model selection.

AI USE STATEMENT

In this work, we used generative AI tools solely to polish the writing and assist in identifying relevant literature. We did not use generative AI tools for any other aspect of the research. We have reviewed all AI-assisted work, including checking that language revisions preserved the intended meaning and verifying the accuracy and relevance of the identified references. We take responsibility for the final content of this work, including all text and claims produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The theoretical assumptions and results are presented in Section 4, with proofs provided in Appendix A. The proposed method and experimental setup are described in Sections 5.1 and 5.2, respectively. Hyperparameter selection is discussed in Section 5.3, and the algorithmic pseudocode is provided in Appendix B.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6430–6439, 2019.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, pp. 2309–2313. IEEE, 2019.

Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. *Advances in Neural Information Processing Systems*, 34:19301–19312, 2021.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9620–9629. IEEE, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Léo Dana, Loucas Pillaud-Vivien, and Francis Bach. Convergence of the gradient flow for shallow relu networks on weakly interacting data. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 5847–5878. Curran Associates, Inc., 2025. doi: 10.52202/085713-0206. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/08f805247677d42eac83664ccf9ef3ef-Paper-Conference.pdf.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

Mohsen Gholami, Mohammad Akbari, Xinglu Wang, Behnam Kamranian, and Yong Zhang. Etran: Energy-based transferability estimation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18567–18576. IEEE, 2023.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2261–2269. Ieee, 2017.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *International conference on machine learning*, pp. 9201–9225. PMLR, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Prafful Kumar Khoba, Zijian Wang, Chetan Arora, and Mahsa Baktashmotlagh. Feature space perturbation: A panacea to enhanced transferability estimation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1299–1308. IEEE, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551. Ieee, 2017.

Xiaotong Li, Zixuan Hu, Yixiao Ge, Ying Shan, and Ling-Yu Duan. Exploring model transferability through the lens of potential energy. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5406–5415. IEEE, 2023.

Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, and Boqing Gong. Ranking neural checkpoints. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2662–2672. IEEE, 2021.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International conference on machine learning*, pp. 7294–7305. PMLR, 2020.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9162–9172. IEEE, 2022.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 4510–4520. Ieee, 2018.

Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. Not all models are equal: Predicting model transferability in a self-challenging fisher space. In *European Conference on Computer Vision*, pp. 286–302. Springer, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2815–2823. IEEE, 2019.

Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1395–1405, 2019.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5526–5535. IEEE, 2023.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PUnD86UEK5.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9643–9653. IEEE, 2022.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International conference on machine learning*, pp. 12133–12143. Pmlr, 2021.

Yaoyan Zheng, Huiqun Wang, Nan Zhou, and Di Huang. Implicit modeling for transferability estimation of vision foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QnUNr9Xuoo.

Liu Ziyin, Yizhou Xu, and Isaac L. Chuang. Neural thermodynamics: Entropic forces in deep and universal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PPcM4cWgbp.

A PROOFS OF THE THEORETICAL RESULTS

In this appendix, we provide the proofs of the theoretical results in Sections 3 and 4. We first establish the evolution of the gradient-mass distribution under continuous-time dynamics, and then derive the entropy-dependent descent bounds used in our adaptation analysis.

Throughout this section, we suppress the dependence of the loss on the dataset and write

$$\mathcal{L}(\theta) := \mathcal{L}(X, \theta), \qquad g := \nabla_\theta \mathcal{L}(\theta), \qquad H := \nabla_\theta^2 \mathcal{L}(\theta).$$

Recall that the parameter dynamics are

$$\dot{\theta} = -u,$$

and that the gradient-mass distribution is defined as

$$p_i := \frac{g_i^2}{\|g\|_2^2}, \qquad \sum_i p_i = 1.$$

For each coordinate with $g_i \neq 0$, define

$$c_i := \frac{(Hu)_i}{g_i}, \qquad \bar{c} := \sum_i p_i c_i = \frac{g^\top H u}{g^\top g}.$$

The quantity c_i is the instantaneous relative decay rate of the i-th gradient component, since

$$\frac{\mathrm{d}}{\mathrm{d}t} \log |g_i| = -c_i.$$

For simplicity, the derivations below are stated on intervals on which the relevant gradient components are nonzero. The same formulas apply on a fixed active support, with boundary cases understood by continuity whenever the corresponding limits exist.

A.1 GRADIENT-ENTROPY DYNAMICS

Proof of Proposition 3.2. Because $g = \nabla_\theta \mathcal{L}(\theta)$, the chain rule gives

$$\dot{g} = H\dot{\theta} = -Hu.$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t} \|g\|_2^2 = 2g^\top \dot{g} = -2g^\top H u.$$

Dividing by $\|g\|_2^2 = g^\top g$ yields

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \|g\|_2^2 = -2\frac{g^\top H u}{g^\top g}.$$

This proves the proposition. □

We next record the evolution of the normalized gradient mass.

Lemma A.1. *For each i,*

$$\dot{p}_i = -2p_i(c_i - \bar{c}).$$

In particular,

$$\sum_i \dot{p}_i = 0.$$

Proof. Since

$$p_i = \frac{g_i^2}{\|g\|_2^2},$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \log p_i = 2\frac{\dot{g}_i}{g_i} - \frac{\mathrm{d}}{\mathrm{d}t} \log \|g\|_2^2.$$

Using

$$\dot{g}_i = -(Hu)_i = -c_i g_i$$

together with Proposition 3.2 gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \log p_i = -2c_i + 2\bar{c}.$$

Multiplying both sides by p_i yields

$$\dot{p}_i = -2p_i(c_i - \bar{c}).$$

Finally,

$$\sum_i \dot{p}_i = -2 \sum_i p_i(c_i - \bar{c}) = -2 \left(\sum_i p_i c_i - \bar{c} \sum_i p_i \right) = 0.$$

□

Proof of Theorem 3.3. For $\alpha \neq 1$, define

$$Z_\alpha := \sum_k p_k^\alpha.$$

The Rényi entropy can then be written as

$$S_\alpha(p) = \frac{1}{1-\alpha} \log Z_\alpha.$$

Differentiating Z_α and using Lemma A.1, we obtain

$$\begin{aligned} \dot{Z}_\alpha &= \alpha \sum_k p_k^{\alpha-1} \dot{p}_k \\ &= -2\alpha \sum_k p_k^\alpha (c_k - \bar{c}). \end{aligned} \tag{6}$$

Since

$$\bar{c} = \sum_j p_j c_j, \qquad \sum_j p_j = 1,$$

the term on the right-hand side can be rewritten as

$$\sum_k p_k^\alpha (c_k - \bar{c}) = \sum_{i,j} p_i^\alpha p_j (c_i - c_j). \tag{7}$$

Exchanging i and j in equation 7 and averaging the two expressions gives

$$\sum_{i,j} p_i^\alpha p_j (c_i - c_j) = \frac{1}{2} \sum_{i,j} p_i p_j \left(p_i^{\alpha-1} - p_j^{\alpha-1} \right) (c_i - c_j). \tag{8}$$

Combining equation 6–equation 8, we obtain

$$\begin{aligned} \frac{\mathrm{d}S_\alpha(p)}{\mathrm{d}t} &= \frac{\dot{Z}_\alpha}{(1-\alpha)Z_\alpha} \\ &= \frac{\alpha}{(\alpha-1)Z_\alpha} \sum_{i,j} p_i p_j \left(p_i^{\alpha-1} - p_j^{\alpha-1} \right) (c_i - c_j) \\ &= \frac{\alpha}{(\alpha-1)\sum_k p_k^\alpha} \sum_{i,j} p_i p_j \left(p_i^{\alpha-1} - p_j^{\alpha-1} \right) (c_i - c_j), \end{aligned} \tag{9}$$

which proves the first statement.

We now establish the sign condition. Suppose that

$$(p_i - p_j)(c_i - c_j) \geq 0 \qquad \text{for all } i, j.$$

If $\alpha > 1$, the map $x \mapsto x^{\alpha-1}$ is increasing. Hence

$$\left(p_i^{\alpha-1} - p_j^{\alpha-1}\right)(c_i - c_j) \geq 0,$$

and the prefactor $\alpha/(\alpha-1)$ is positive.

If $0 < \alpha < 1$, the map $x \mapsto x^{\alpha-1}$ is decreasing. Hence

$$\left(p_i^{\alpha-1} - p_j^{\alpha-1}\right)(c_i - c_j) \leq 0,$$

while $\alpha/(\alpha-1) < 0$. Therefore, in both cases,

$$\frac{\mathrm{d}S_\alpha(p)}{\mathrm{d}t} \geq 0.$$

The Shannon case $\alpha = 1$ follows by continuity and is given explicitly below. □

Proof of Corollary 3.4. For $\alpha = 1$, the Shannon entropy is

$$S_1(p) = -\sum_i p_i \log p_i.$$

Differentiating and using $\sum_i \dot{p}_i = 0$, we obtain

$$\begin{aligned}\frac{\mathrm{d}S_1(p)}{\mathrm{d}t} &= -\sum_i \dot{p}_i(\log p_i + 1) \\ &= -\sum_i \dot{p}_i \log p_i. \end{aligned} \tag{10}$$

Applying Lemma A.1 gives

$$\frac{\mathrm{d}S_1(p)}{\mathrm{d}t} = 2\sum_i p_i(c_i - \bar{c}) \log p_i.$$

The same expression can also be recovered as the limit $\alpha \to 1$ in Theorem 3.3. Indeed,

$$p^{\alpha-1} = 1 + (\alpha-1)\log p + O\big((\alpha-1)^2\big),$$

and

$$\sum_k p_k^\alpha \to 1.$$

Consequently,

$$\begin{aligned}\lim_{\alpha\to 1} \frac{\mathrm{d}S_\alpha(p)}{\mathrm{d}t} &= \sum_{i,j} p_i p_j (\log p_i - \log p_j)(c_i - c_j) \\ &= 2\sum_i p_i(c_i - \bar{c}) \log p_i. \end{aligned} \tag{11}$$

Moreover, because the logarithm is increasing,

$$(p_i - p_j)(c_i - c_j) \geq 0$$

implies

$$(\log p_i - \log p_j)(c_i - c_j) \geq 0.$$

Thus the same pairwise alignment condition implies

$$\frac{\mathrm{d}S_1(p)}{\mathrm{d}t} \geq 0.$$

□

A.2 Gradient entropy and effective gradient dimension

For convenience, define the effective dimension associated with the Rényi entropy by

$$d_\alpha(p) := \exp\big(S_\alpha(p)\big).$$

Lemma A.2. *Let $p_i = g_i^2/\|g\|_2^2$, and let $s = 2\alpha$. For $\alpha > 0$, $\alpha \neq 1$,*

$$\frac{\|g\|_s^2}{\|g\|_2^2} = d_\alpha(p)^{(1-\alpha)/\alpha}.$$

In particular, when $\alpha = \frac{1}{2}$,

$$d_{1/2}(p) = \frac{\|g\|_1^2}{\|g\|_2^2}.$$

Proof. By the definition of p_i,

$$\begin{aligned}\sum_i p_i^\alpha &= \sum_i \left(\frac{g_i^2}{\|g\|_2^2}\right)^\alpha \\ &= \frac{\sum_i |g_i|^{2\alpha}}{\|g\|_2^{2\alpha}} = \frac{\|g\|_{2\alpha}^{2\alpha}}{\|g\|_2^{2\alpha}}. \end{aligned} \tag{12}$$

Therefore,

$$\begin{aligned} d_\alpha(p) = \exp(S_\alpha(p)) &= \left(\sum_i p_i^\alpha\right)^{1/(1-\alpha)} \\ &= \left(\frac{\|g\|_{2\alpha}}{\|g\|_2}\right)^{\frac{2\alpha}{1-\alpha}}. \end{aligned} \tag{13}$$

Rearranging gives

$$\frac{\|g\|_{2\alpha}^2}{\|g\|_2^2} = d_\alpha(p)^{(1-\alpha)/\alpha}.$$

Setting $\alpha = \frac{1}{2}$ yields

$$d_{1/2}(p) = \frac{\|g\|_1^2}{\|g\|_2^2}.$$

□

A.3 One-step entropy-dependent descent bound

We first prove the ℓ_∞-steepest-descent property used in Section 4.

Proof of Proposition 4.2. For any v satisfying $\|v\|_\infty \leq 1$,

$$\begin{aligned} g_t^\top v &\leq \sum_i |g_{t,i}|\,|v_i| \\ &\leq \sum_i |g_{t,i}| = \|g_t\|_1. \end{aligned} \tag{14}$$

Taking

$$v_t^\star = \operatorname{sign}(g_t)$$

achieves equality:

$$g_t^\top v_t^\star = \sum_i |g_{t,i}| = \|g_t\|_1.$$

Hence

$$\operatorname{sign}(g_t) \in \arg\max_{\|v\|_\infty \leq 1} g_t^\top v.$$

□

We next derive a descent lemma implied by Assumption 4.1. Let

$$L_\infty := \sum_{i=1}^{N_g} L_i.$$

For any perturbation Δ, the fundamental theorem of calculus gives

$$\begin{aligned}\mathcal{L}(\theta+\Delta) - \mathcal{L}(\theta) &= \int_0^1 \nabla\mathcal{L}(\theta+s\Delta)^\top \Delta \,\mathrm{d}s \\ &= g^\top\Delta + \int_0^1 \big(\nabla\mathcal{L}(\theta+s\Delta) - \nabla\mathcal{L}(\theta)\big)^\top \Delta \,\mathrm{d}s. \end{aligned} \tag{15}$$

Under Assumption 4.1,

$$|\nabla_i\mathcal{L}(\theta+s\Delta) - \nabla_i\mathcal{L}(\theta)| \le sL_i\|\Delta\|_\infty.$$

Therefore,

$$\begin{aligned}\mathcal{L}(\theta+\Delta) &\le \mathcal{L}(\theta) + g^\top\Delta + \int_0^1 s\sum_i L_i\|\Delta\|_\infty|\Delta_i|\,\mathrm{d}s \\ &\le \mathcal{L}(\theta) + g^\top\Delta + \frac{L_\infty}{2}\|\Delta\|_\infty^2. \end{aligned} \tag{16}$$

Proof of Theorem 4.3. Consider an update along the ℓ_∞-steepest direction,

$$\Delta = -\rho v_t^\star, \qquad v_t^\star = \mathrm{sign}(g_t).$$

By Proposition 4.2,

$$\|v_t^\star\|_\infty = 1, \qquad g_t^\top v_t^\star = \|g_t\|_1.$$

Applying equation 16 therefore gives

$$\mathcal{L}(\theta_t - \rho v_t^\star) \le \mathcal{L}(\theta_t) - \rho\|g_t\|_1 + \frac{L_\infty}{2}\rho^2.$$

The quadratic upper bound on the right-hand side is minimized at

$$\rho_t^\star = \frac{\|g_t\|_1}{L_\infty}.$$

Thus, with

$$\Delta_t^\star = -\frac{\|g_t\|_1}{L_\infty}\,\mathrm{sign}(g_t),$$

we obtain

$$\mathcal{L}(\theta_t + \Delta_t^\star) \le \mathcal{L}(\theta_t) - \frac{\|g_t\|_1^2}{2L_\infty}.$$

For $\alpha = \frac{1}{2}$,

$$\begin{aligned}S_{1/2}(p_t) &= 2\log\left(\sum_i \sqrt{p_{t,i}}\right) \\ &= 2\log\left(\frac{\|g_t\|_1}{\|g_t\|_2}\right). \end{aligned} \tag{17}$$

Equivalently,

$$\exp\big(S_{1/2}(p_t)\big) = \frac{\|g_t\|_1^2}{\|g_t\|_2^2},$$

or

$$\|g_t\|_1^2 = \|g_t\|_2^2\exp\big(S_{1/2}(p_t)\big).$$

Substituting this identity into equation A.3 yields

$$\mathcal{L}(\theta_t + \Delta_t^\star) \le \mathcal{L}(\theta_t) - \frac{\|g_t\|_2^2}{2L_\infty}\exp\big(S_{1/2}(p_t)\big).$$

Equivalently,

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta_t + \Delta_t^\star) \ge \frac{\|g_t\|_2^2}{2L_\infty}\exp\big(S_{1/2}(p_t)\big).$$

This proves the theorem. □

A.4 TRAJECTORY-LEVEL ENTROPY BOUND

Proof of Theorem 4.5. Let

$$\Delta_t := \mathcal{L}(\theta_t) - \mathcal{L}^\star.$$

By Theorem 4.3, the update at iteration t satisfies

$$\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t) - \frac{\|g_t\|_2^2}{2L_\infty} \exp\big(S_{1/2}(p_t)\big).$$

Subtracting $\mathcal{L}^\star$ from both sides gives

$$\Delta_{t+1} \le \Delta_t - \frac{\|g_t\|_2^2}{2L_\infty} \exp\big(S_{1/2}(p_t)\big).$$

By the PL condition in Assumption 4.4,

$$\frac{1}{2}\|g_t\|_2^2 \ge \mu\Delta_t.$$

Substituting this inequality into equation A.4 yields

$$\Delta_{t+1} \le \left(1 - \frac{\mu}{L_\infty} \exp\big(S_{1/2}(p_t)\big)\right) \Delta_t.$$

For completeness, the contraction factor is nonnegative whenever $\Delta_t > 0$. Indeed, because $\mathcal{L}^\star$ is a lower bound on the loss, Theorem 4.3 implies

$$\frac{\|g_t\|_2^2}{2L_\infty} \exp\big(S_{1/2}(p_t)\big) \le \Delta_t.$$

Combining this inequality with the PL condition gives

$$0 \le \frac{\mu}{L_\infty} \exp\big(S_{1/2}(p_t)\big) \le 1.$$

Applying equation A.4 recursively over $t = 0, \dots, T-1$ therefore yields

$$\Delta_T \le \Delta_0 \prod_{t=0}^{T-1} \left(1 - \frac{\mu}{L_\infty} \exp\big(S_{1/2}(p_t)\big)\right).$$

Using $1 - x \le e^{-x}$ for $x \ge 0$, we obtain

$$\Delta_T \le \Delta_0 \exp\left(-\frac{\mu}{L_\infty} \sum_{t=0}^{T-1} \exp\big(S_{1/2}(p_t)\big)\right).$$

Restoring the definition of Δ_t gives

$$\mathcal{L}(\theta_T) - \mathcal{L}^\star \le \big(\mathcal{L}(\theta_0) - \mathcal{L}^\star\big) \exp\left(-\frac{\mu}{L_\infty} \sum_{t=0}^{T-1} \exp\big(S_{1/2}(p_t)\big)\right).$$

□

Proof of Corollary 4.6. Suppose, in addition, that the gradient entropy does not decrease from its value at the pretrained initialization, i.e.,

$$S_{1/2}(p_t) \ge S_{1/2}(p_0), \qquad t = 0, \dots, T-1.$$

Then

$$\sum_{t=0}^{T-1} \exp\big(S_{1/2}(p_t)\big) \ge T \exp\big(S_{1/2}(p_0)\big).$$

Applying this inequality to Theorem 4.5 gives

$$\mathcal{L}(\theta_T) - \mathcal{L}^\star \le \big(\mathcal{L}(\theta_0) - \mathcal{L}^\star\big) \exp\left(-\frac{\mu T}{L_\infty} \exp\big(S_{1/2}(p_0)\big)\right).$$

Thus, under the nondecreasing-entropy condition, the initial Rényi-$\frac{1}{2}$ gradient entropy alone provides a lower bound on the cumulative entropy term governing contraction along the subsequent adaptation trajectory. □

B Details of Method for the selection of K and q

Algorithm 1 shows the detailed algorithm calculating the valley-restricted score.

Algorithm 1 Valley layer selection and Rényi gradient-entropy score.

Require: Per-tensor statistics $\{(S_i, P_i)\}_{i=1}^{L}$ in forward order, with $S_i = \sum_j g_{ij}^2$ and $P_i = \sum_j \left(g_{ij}^2\right)^{\alpha}$; number of bins K; number of kept bins $q \geq 1$; Rényi order $\alpha \neq 1$

Ensure: Transferability score s

function ValleyScore($\{(S_i, P_i)\}_{i=1}^{L}$, K, q, α)
  if $q \geq K$ **then**
    $\mathcal{G} \leftarrow \{1, \ldots, L\}$ ▷ whole network
  else
    $\mathcal{B}_k \leftarrow \emptyset$ **for** $k = 1, \ldots, K$
    for $i = 1, \ldots, L$ **do** ▷ equal-depth bins by tensor position
      $k \leftarrow \min\left(\lfloor (i-1)K/L \rfloor,\ K-1\right) + 1$
      $\mathcal{B}_k \leftarrow \mathcal{B}_k \cup \{i\}$
    end for
    $E_k \leftarrow \sum_{i \in \mathcal{B}_k} S_i$ **for** $k = 1, \ldots, K$ ▷ gradient energy of each bin
    if $\sum_{k=1}^{K} E_k \leq 0$ **then**
      $\mathcal{G} \leftarrow \{1, \ldots, L\}$
    else
      $\pi \leftarrow$ bin indices ordered so that $E_{\pi(1)} \leq \cdots \leq E_{\pi(K)}$
      $\mathcal{G} \leftarrow \bigcup_{r=1}^{q} \mathcal{B}_{\pi(r)}$ ▷ the q lowest-energy bins
    end if
  end if
  $S_{\mathcal{G}} \leftarrow \sum_{i \in \mathcal{G}} S_i, \quad P_{\mathcal{G}} \leftarrow \sum_{i \in \mathcal{G}} P_i$
  if $S_{\mathcal{G}} \leq 0$ **then**
    return 0
  end if
  $H_\alpha \leftarrow \frac{1}{1-\alpha}\left(\ln P_{\mathcal{G}} - \alpha \ln S_{\mathcal{G}}\right)$ ▷ $= \frac{1}{1-\alpha} \ln \sum_{i \in \mathcal{G}} \sum_j p_{ij}^{\alpha},\ p_{ij} = g_{ij}^2 / S_{\mathcal{G}}$
  return $\exp\left(\max(H_\alpha,\ 0)\right)$
end function