SETOL: A Semi-Empirical Theory of (Deep) Learning

Charles H. Martin* Christopher Hinrichs†

Abstract

We present a Semi-Empirical Theory of Learning (`SETOL`) that explains the remarkable performance of State-of-the-Art (SOTA) Neural Networks (NNs). We provide a formal explanation of the origin of the fundamental quantities in the phenomenological theory of Heavy-Tailed Self-Regularization (`HTSR`), the Heavy-Tailed Power Law Layer Quality metrics, `AlphaHat` (α) and `AlphaHat` ($\hat{\alpha}$). In prior work, these metrics have been shown to predict trends in the test accuracies of pretrained SOTA NN models, and, importantly, without needing access to the testing or even training data. Our `SETOL` uses techniques from Statistical Mechanics (`StatMech`) as well as advanced methods from Random Matrix Theory (`RMT`). Our derivation suggests new mathematical preconditions for *Ideal* learning, including the new `TRACE-LOG` metric (which is equivalent to applying the Wilson Exact Renormalization Group). We test the assumptions and predictions of our `SETOL` on a simple 3-layer Multi-Layer Perceptron (MLP), demonstrating excellent agreement with the key theoretical assumptions. For SOTA NN models, we show how to estimate the Model Quality of a trained NN by simply computing the Empirical Spectral Density (ESD) of the layer weight matrices and then plugging this ESD into our `SETOL` formulae. Notably, we examine the performance of the HTSR α and the `SETOL TRACE-LOG` Layer Quality metrics, and find that they align remarkably well, both on our MLP and SOTA NNs.

*Calculation Consulting, 8 Locksley Ave, 6B, San Francisco, CA 94122, `charles@CalculationConsulting.com`.
†Onyx Point Systems, chris@onyxpointsystems.com

Contents

1 Introduction

Deep Neural Networks (DNNs)—models in the field of Artificial Intelligence (AI)—have driven remarkable advances in multiple fields of science and engineering. AlphaFold has made significant progress in solving the protein folding problem.[1] Notably, the 2024 Nobel Prize in Physics was awarded to Hopfield and Hinton for developing early approaches to AI using techniques from *Statistical Mechanics* (`StatMech`), and Jumper and Hassabis, along with Baker, received the 2024 Nobel Prize in Chemistry for their contributions to AlphaFold and computational protein design.[2, 3] Self-driving cars now roam the streets of major metropolitan cities like San Francisco. Large Language Models (LLMs) like ChatGPT have gained worldwide attention and initiated serious conversations about the possibility of creating an Artificial General Intelligence (AGI). Clearly, not a single area of science or engineering has ignored these remarkable advances in the field of AI and Neural Networks (NNs).

Despite this remarkable progress in a research field spanning over 50 years, developing, training, and maintaining such complex models require staggering capital resources, limiting their development to only the largest and best-funded organizations. While many such entities have open-sourced some of their largest models (such as Llama and Falcon), using these models requires assuming they have been trained optimally, without significant defects that could limit, skew, or even invalidate their use downstream. Moreover, testing such models can be very expensive and complex to interpret.

Because training and evaluating NNs is so hard, significant issues can manifest in many obvious and non-obvious ways. A primary research goal is to improve the efficiency and reduce the cost of training large NNs. A less known but critical issue arises in many industrial settings, specifically "selecting the best models to test." This arises in industries such as ad-click prediction, search relevance, quantitative trading, and more. Frequently, one has several seemingly equally good models to choose from, but testing the model can be very expensive, time-consuming, and even risky to the business. Recently, researchers and practitioners have started to fine-tune such large open-source models using techniques such as LoRA and QLoRA. Such methods allow one to adapt a large, open-source NN to a small dataset, and very cheaply. However, in fine-tuning, one could unwittingly overfit the model to the small dataset, degrading its performance for its intended use. Despite these and many other problems, theory remains well behind practice, and there is an increasingly pressing need to develop *practical predictive theory* to both improve the training of these very large NN models and to design new methods to make their use more reliable.

Before discussing these methods, however, let us explain *What is a SemiEmpirical Theory*

1.1 Statistical Mechanics (`StatMech`) vs. Statistical Learning Theory (`SLT`)

Historically, there have been two competing theoretical frameworks for understanding NNs: *Statistical Mechanics* (`StatMech`) [4, 5, 6, 7, 8, 9, 10]; and *Statistical Learning Theory* (`SLT`) [11].

- **Statistical Mechanics (`StatMech`).** This framework has been foundational to the early development of NN models, such as the Hopfield Associative Memory (HAM) [12], Boltzmann Machines [13], [14], etc. `StatMech` has also been used to build early theories of learning, such as the Student-Teacher model for the Perceptron Generalization Error [4, 5], the Gardner model [6], and many others. Notably, the HAM was based on an idea by Little, who observed that, in a simple model, long-term memories are stored in the eigenvectors of transfer matrix [15]. (This general idea, but in a broader sense, is central to our approach below.) Moreover, `StatMech` predicts that NNs exhibit phase behavior. This has recently been rediscovered as the Double Descent phenomenon [16, 17], but it was known in `StatMech` long before it's recent rediscovery [18]. However, unlike other applied physics theories (e.g.,

Semi-Empirical methods in quantum chemistry), StatMech only offers qualitative analogies, failing to provide lacks quantitative predictions about large, modern NN models.[19]

- **Statistical Learning Theory (SLT).** SLT and related approaches (VC theory, PAC bounds theory, etc.) have been developed within the context of traditional computational learning problems [11], and they are based on analyzing the convergence of frequencies to probabilities (over problem classes, etc.). It was recognized early on, however, that they could not be directly applied to NNs [20]. Moreover, SLT cannot even reproduce quantitative properties of learning curves [21, 22] (whereas StatMech is very successful at this [8]). SLT also failed to predict the "Double Descent" phenomena [16]. More recently, it has been shown that in practical settings SLT can give vacuous [23] or even opposite results to those actually observed [24].

Technically, SLT focuses on obtaining bounds on a model's worst-case behavior, while StatMech seeks a probabilistic understanding of typical behaviors across different states or configurations. Unfortunately, neither of these general theoretical frameworks has proven particularly useful to NN practitioners. SETOL combines insights from both. Rather than being purely phenomenological like the HTSR approach, SETOL is derived from first-principles, and in form of a Semi-Empirical theory. As such, SETOL offers a practical, Semi-Empirical framework that bridges rigorous theoretical modeling and empirical observations for modern NNs.

1.2 Heavy-Tailed Self-Regularization (HTSR)

HTSR theory is an approach that combines ideas from StatMech with those of *Heavy-Tailed Random Matrix Theory* (RMT), providing eigenvalue-based quality metrics that correlate with model quality (i.e., out-of-sample performance). HTSR theory posits that well-trained models have extracted subtle correlations from the training data, and that these correlations manifest themselves in the *Shape* and *Scale* of the eigenvalues of the layer weight matrices $\mathbf{W}$. In particular, if one computes the empirical distribution of the eigenvalues, λ_i, of an individual $N \times M$ weight matrix, $\mathbf{W}$, then this density, $\rho^{emp}(\lambda)$, which is an ESD, is Heavy-Tailed (HT) and can be well-fit to a *Power Law* (PL), i.e., $\rho(\lambda) \sim \lambda^{-\alpha}$, with exponent α. HTSR theory provides a *phenomenology* for qualitatively-distinct phases of learning [25]. It can, however, also be used to define *Layer-level Quality metrics* and *Model-level Quality metrics*: e.g., the Alpha (α) and AlphaHat ($\hat{\alpha}$) PL metrics, described below.

Not needing any training data, HTSR theory has many practical uses. It can be directly applicable to large, open-source models where the training and test data may not be available. Model quality metrics can be used, e.g., to predict trends in the quality of SOTA models in computer vision (CV) [26] and natural language processing (NLP) [27, 28], both during and after training, and without needing access to the model test or training data. Layer quality metrics can be used to diagnose potential internal problems in a given model, or (say) to accelerate training by providing optimal layer-wise learning rates [29] or pruning rations [30]. Most notably, the HTSR theory provides *Universal* Layer Quality metrics encapsulated in what apears to be a critical exponent, $\alpha = 2$, that is empirically associated with optimal or Ideal Learning. Moreover, as argued below, the value $\alpha = 2$ appears to define a phase boundary between a generalization and overfitting, analogous to the phase boundaries seen in StatMech theories of NN learning.

These results both motivate the search for a first prinples understanding of the HTSR theory, and suggests a path for developing a practical predictive theory of Deep Learning. For this, however, we need to go beyond the phenomenology provided by HTSR theory, to relate it to some sort of (at least semi-rigorous/semi-empirical) derivations based on the StatMech theory of learning, and drawing

upon previous success (in Quantum Chemistry) in developing a first principles Semi-Empirical theory.

1.3 What is a Semi-Empirical Theory?

Historically, one of the most well known *Semi-Empirical* methods comes from Nuclear Physics. The Semi-Empirical Mass Formula, dating back to 1935, is based on the heuristic Liquid Drop Model of the nucleus, and it was used to predict experimentally-observed binding energies of nucleons. This model describes nuclear fission, and it was central to its development of the atomic bomb:

> Prior to WWII, Nuclear Physics was a phenomenological science, which relied upon experimental data and descriptive models [31].

In the Post-war era, the epistemological nature of nuclear theory changed, as it saw the development of Semi-Empirical shell models of the nucleus. These models were formulated with rigor (in the physics sense) but also relied on heuristic assumptions and experimental data for accurate predictions. They captured the structure of atomic nuclei and could accurately describe various nuclear properties[32, 33, 34]. The shell models, analogous to the electronic shell structure of atoms, represented a shift toward a more rigorous understanding of nuclear phenomena.

About this time, `RMT` itself was also introduced by *Wigner* [35] to model the statistical patterns of the nuclear energy spectra of strongly interacting heavy nuclei. These patterns were universal, independent of the specific nucleus, suggesting that a probabilistic approach would be fruitful. In the following decades, `RMT` saw many advances, including the development of the Marchenko-Patur model[36], and numerous other applications in physics[37]. By the 1990s, `RMT` was further expanded when *Zee* introduced the *Blue Function*, and reinterpreted the *R-transform* as a self-energy within the framework of many-body / quantum field theory (QFT) [38]. Also, so-called `HCIZ` integrals, integrals over random matrices, were being used both to model disordered electronic spectra[39], and, later, the behavior of spin glass models[40, 41].

Returning to the 1950s, and prior to the development of highly accurate, modern, computational *ab initio* theories of Quantum Chemistry,theoretical chemists introduced the Semi-Empirical PPP method for conjugated polyenes [42]The PPP model recast the electronic structure problem as an *Effective Hamiltonian* for the π-electrons. [1] For many years this and related Semi-Empirical methods worked remarkably well, even better than the existing *ab initio* theories[44, 45, 46, 47] Most importantly, these methods could be *fit* on a broad set of empirical molecular data, and then applied to molecules not in the original training set.

Around the same time, *Löwdin* first formalized the concept of the Effective Hamiltonian, which allowed the reduction of complex many-body problems to simplified *Effective Potentials* that still captured the essential physics. Then in the late 1960s Brandow developed an Effective Hamiltonian theory of nuclear structure, leveraging the *Linked Cluster Theorem* (LCT) (see [48]) and quantum mechanical many-body theory to describe the highly correlated effective interactions in a reduced model space. [2]

Like modern NNs, these Semi-Empirical methods of Quantum Chemistry worked well beyond their apparent range of validity, generalizing very well to out-of-distibution (OOD) data. This led to the search for a Semi-Empirical *Theory* to explain the remarkable performance of these phenomenological methods. Building on Brandow's many-body formalism, Freed and collaborators [49, 50] developed an *ab initio* Effective Hamiltonian Theory of Semi-Empirical methods to

[1] The PPP model resembles the later developed tight-binding model of condensed matter physics[43]

[2] Note also that the LCT shows that the log partition function (*i.e.*, $\ln Z$) can be expressed a sum of connected diagrams, which is very similar to our result below, which expresses the log partition function here as a sum of matrix cumulants from RMT.

explain the remarkable success of the Semi-Empirical methods. Specifically, the values of the PPP empirical parameters could be directly computed effective interactions, including both renormalized self-energies and higher-order terms. Somewhat later, in the 90s, Martin et. al.[51, 52, 53, 54] extended and applied this Effective Hamiltonian theory and demonstrated the Universality of the Semi-Empirical PPP parameters numerically. Indeed, it is this Universality that enabled the for-a-time inexplicable ODD performance of these methods. Crucially, this decades long line of work established a comprehensive analytic and numerical *Theory* of Semi-Empirical methods. That is, a framework that confirmed the empirically observed Universality, provided theoretical justification for this, and enabled systematic improvements of the methods using numerical techniques.

Finally, it is important to mention the Effective Hamiltonian approach provided by the Wilson *Renormalization Group* (RG).[55, 56] The RG approach provides a powerful framework for studying strongly correlated systems across different scales, enabling the construction of an Effective Hamiltonian by *integrating out* weakly-correlated degrees of freedom in a Scale-Invariant way. It is particularly suited for critical points and phase boundaries– such the phase boundary between generalization and memorization in spin glass models of neural networks– and, importantly, predicts the exisitence of Universal Power Law (PL) exponents .

Relevance to Deep Learning In this sense, Semi-Empirical theories of Nuclear Physics and Quantum Chemistry, (as well as the Renormalization Group approach), seem particularly appropriate for Deep Learning. DNN models are complex black boxes that defy statistical descriptions they are commonly pre-trained on a large set of data; and than applied to new data sets in new domains via transfer learning. Most recently, the inexplicable success of transfer-learning is seen in the GPT (Generative Pre-trained Transformer) models[57], and motivated early work by Jumper et. al. on protein folding[58]

In contrast, these Semi-Empirical approaches differ from more recently developed theoretical approaches to deep learning, which are typically based on SLT, rather than `StatMech` [59]. In particular, there have recently appeared several theories of deep learning, formulated using ideas from `RMT`. However, regarding realistic models, it has been explicitly stated that "These networks are however too complex in general for developing a rigorous theoretical analysis on the spectral behavior [60]. Even in recent work applying `RMT` to NNs, it has been noted "*that we make no claim about trained weights, only random weights*" [61]. The weight matrices of a trained NN, however, are clearly *not* simply random matrices—since they encode the specific correlations from the training data.

1.4 A Semi-Empirical Theory of Learning (`SETOL`)

We propose `SETOL`, a Semi-Empirical Theory for Deep Learning Neural Networks (NNs), as both a theoretical foundation for `HTSR` phenomenology and a novel framework for predicting the properties of complex NN models. This unified framework offers a deeper understanding of DNN generalization through a Semi-Empirical approach inspired by many-body physics, combined with a classic `StatMech` model for NN generalization. Specifically, `SETOL` combines theoretical and empirical insights to evaluate Model Quality, showing that the weightwatcher layer `HTSR` PL metrics (`Alpha` and `AlphaHat`) can be derived using a phenomenological Effective Hamiltonian approach. This approach expressses the `HTSR` Layer Quality in terms of the RMT matrix cumulants of the layer weight matrix $\mathbf{W}$, and is governed by a Scale-Invariant transformation equivalent to a single step of an exact Renormalization Group (RG) transformation. Here, we derive this from first principles, requiring no previous knowledge of statistical physics.

The `SETOL` approach unifies the `HTSR` principles with a broader theoretical framework for layer analysis. The `HTSR` theory identifies Universality (e.g., $\alpha = 2$) as a hallmark of the best-trained

DNN layers, and, here, ourSETOL introduces the closely related *Trace Log Condition*, a Scale-Invariant or Volume-Preserving transformation that reflects an underlying *Conservation Principle*. Together, these principles form the theoretical foundation for deriving HTSR Layer Quality metrics from first principles. By leveraging techniques from StatMech and modern RMT, SETOL offers a rigorous framework to connect empirical observations with theoretical predictions, advancing our understanding of generalization in neural networks.

- **Derivation of the HTSR Layer Quality metrics Alpha and AlphaHat** The SETOL approach takes as input the Empirical Spectral Density (ESD) of the layers of trained NN, and derives an expression for the approximate *Average Generalization Accuracy* of a multi-layer NN, We call this approximation the *Model Quality*, denoted $\bar{\mathcal{Q}}^{NN}$ This Model Quality is expressed as product of individual Layer Quality terms, $\bar{\mathcal{Q}}_L^{NN}$, which themselves can then be directly related to the HTSR Power Law (PL) empirical Alpha (α) and AlphaHat ($\hat{\alpha} = \alpha \log_{10} \lambda_{max}$) metrics.

  In particular, the Layer Quality-Squared, $\bar{\mathcal{Q}}^2 \approx [\bar{\mathcal{Q}}_L^{NN}]^2$, is expressed the logarithm of an HCIZ integral, the Thermal Average of an Annealed Error Potential for a matrix-generalized form the Linear Student-Teacher model of classical StatMech. This HCIZ integral evaluates into the sum of integrated R-transforms from RMT, or, equivalently, as a sum of integrated matrix cumulants. From this, the HTSR AlphaHat metric can be derived in the special case of Ideal Learning. [3]

- **Discovery of a Mathematical Condition for Ideal Learning.** By Ideal Learning, we mean that the specific NN layer has optimally converged, capturing as much of the information as possible in the training data without overfitting to any part of it. In defining this, and deriving our results, we have discovered (and are proposing) a new condition for Ideal Learning, which is associated with the Universality of the HTSR theory

  - HTSR **Condition for Ideal Learning.** This HTSR theory states that a NN layer is Ideal when the ESD can be well fit to a Power Law (PL) distribution, with PL exponent $\alpha = 2$. Importantly, this appears to be a Universal property of all well-trained NNs, independent of the training data, model architecture, and training procedure.
  - SETOL **TRACE-LOG Condition for Ideal Learning.** The SETOL condition for Ideal Learning states that the dominant eigencomponents associated with the ESD of layer form a reduced-rank *Effective Correlation Space* (ECS) that satisfies a new kind of Conservation Principle or *Volume Preserving Transformation* such that the largest eigenvalues $\tilde{\lambda}_i$ of the ECS satisfy the condition $\ln \prod \tilde{\lambda}_i = \sum \ln \tilde{\lambda}_i = 0$. This is called the *TRACE-LOG Condition*. The TRACE-LOG Condition is equivalent to the taking a single step of the Wilson Exact Renormalization (RG).

  The HTSR Condition has been proposed and analyzed previously [25, 26, 28]; but the TRACE-LOG Condition is new, based on our SETOL theory. When these two conditions align, we propose the NN layer is in the Ideal state.

- **Experimental Validation.** We present detailed experimental results on a simple model, along with observations on large-scale pretrained NNs, to demonstrate that the HTSR conditions for ideal learning ($\alpha = 2$) are experimentally aligned with the independent SETOL condition for ideal learning ($\det(\tilde{\mathbf{X}}) = 1$). See Section. 6.3. Our primary objective here is

[3] The SETOL approach to the HTSR theory resembles in spirit the derivation of the Semi-Empirical PPP models using the Effective Hamiltonian theory, where each phenomenological parameter is associated with a renormalized effective interaction, expressed as a sum of linked diagrams or clusters.[52, 54]

not to demonstrate performance improvements on SOTA NNs—this has been previously established [29]. Instead, our aim is to **validate the theoretical assumptions** of `SETOL`, test the **predictions of the `SETOL` framework**, and examine the **new, independent learning conditions** we discovered—on a model that is sufficiently simple that we can evaluate and stress test the theory.

- **Observations on Overfit Layers ($\alpha < 2$).** Being a Semi-Empirical theory, `SETOL` can also identify violations of it's assumptions. For example, when empirical results show $\alpha < 2$ for a single layer, the layer's ESD falls into the `HTSR` Very Heavy-Tailed (VHT) Universality class. (See Section 6.6.) When this happens, the layer may be slightly overfit to the training data, resulting in **suboptimal performance** and potentially even exhibiting **hysteresis-like effects** (memory effects)—that we observe empirically. These effects indicate that overfit layers may retain memory-like behavior, affecting learning dynamics and generalization.

2 Heavy-Tailed Self-Regularization (HTSR)

In this section, we provide an overview of the HTSR phenomenology.[4] HTSR has been presented in detail previously [62, 63, 25].[5] Here, we provide a self-contained summary, with an emphasis on certain technical issues that will be important for our SETOL. We highlight its practical application for interpreting observed behaviors in trained weight matrices, and we distinguish the HTSR phenomenology from the analytical methods used in the SETOL approach. In Section 2.1, we summarize the basic HTSR setup and results; in Section 2.2, we summarize Gaussian (for RMT) and Heavy-Tailed (for HTRMT) Universality; and in Section 2.3, we describe *Shape* metrics and *Scale* metrics that arise from HTSR. (Here, we focus on basic methods for identifying HT correlations in the ESDs of pre-trained weight matrices; in Sections 6.1, 6.5 and 6.6, we show detailed experiments using theoretical constructs from the HTSR phenomenology.)

2.1 The HTSR Setup

We can write the Energy Landscape function (or NN output function) for a NN with L layers as

$$E_{NN} := h_L(\mathbf{W}_L \times h_{L-1}(\mathbf{W}_{L-1} \times (\cdots) + \mathbf{b}_{L-1}) + \mathbf{b}_L) \tag{1}$$

with activation functions $h_l(\cdot)$, and with weight matrices and biases $\mathbf{W}_l$ and $\mathbf{b}_l$.[6] For simplicity of exposition here (HTSR can be applied much more broadly), we ignore the structural details of the layers (dense or not, convolutions or not, residual/skip connections, etc.). We also ignore the biases b_l (because they can be subsumed into the weight matrices), and we treat each layer as though it contains a single weight matrix $\mathbf{W}_L$. We imagine training (or fine-tuning) this model on labeled data $\{\mathbf{x}_\mu, y_\mu\} \in \mathcal{D}$, where $\mathbf{x}_\mu$ is the μ-th input vector and y_μ is its corresponding label (e.g., for binary classification, $y_\mu \in \{-1, 1\}$). We expect to use backprop via some variant of stochastic gradient descent (SGD) to minimize some loss functional, $\mathcal{L}$ (such as ℓ_2, cross-entropy, etc.):

$$\underset{\mathbf{W}_l, \mathbf{b}_l}{\operatorname{argmin}} \sum_{\mu} \mathcal{L}[E_{NN}(\mathbf{x}_\mu), y_\mu] + \Omega, \tag{2}$$

where, Ω denotes some explicit regularizer (such as an ℓ_1 or ℓ_2 constraint on layer weight matrices) or some implicit regularization procedure (such as clipping the weight matrices or applying dropout).

Given a real-valued $N \times M$ layer weight matrix $\mathbf{W}$ (dropping the subscript), let $\mathbf{X}$ be the $M \times M$ layer *Correlation Matrix*:

$$\mathbf{X} := \frac{1}{N}\mathbf{W}^\top\mathbf{W}. \tag{3}$$

The Empirical Spectral Density (ESD) of $\mathbf{W}$, denoted $\rho^{emp}(\lambda)$, is formed from the M eigenvalues λ_j of $\mathbf{X}$:

$$\rho_{emp}(\lambda) := \sum_{j=1}^{M} \delta(\lambda - \lambda_j). \tag{4}$$

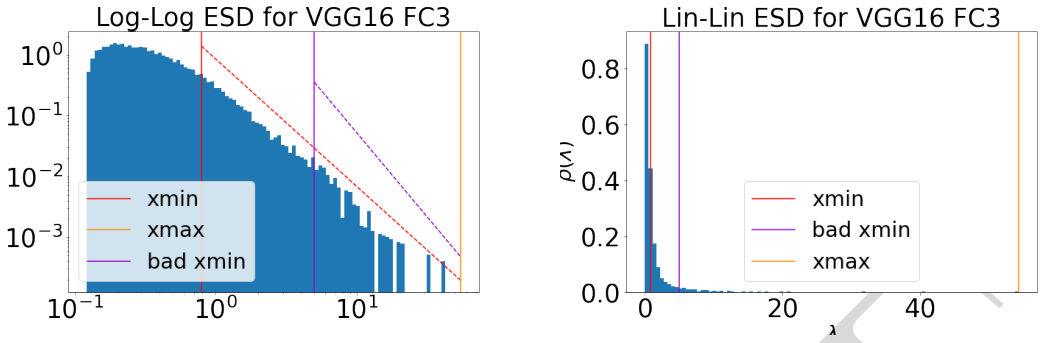
Given a model, we can compute the ESDs of all of its layers, as well as other metrics below, with the open-source WeightWatcher tool [66].[7]

[4] We may also refer to the HTSR phenomenology as the HTSR Theory; we use the term phenomenology to emphasize its empirical nature, and to distinguish it from the analytical methods used in the SETOL approach.

[5]

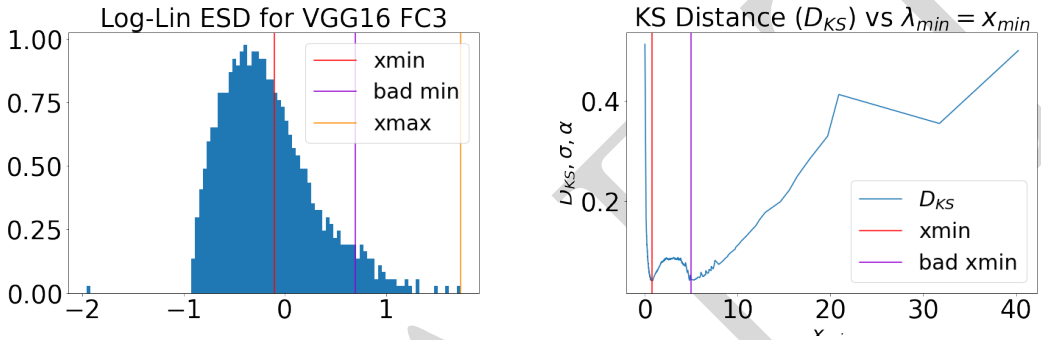
[6] The Energy Landscape function E_{NN} acts on a data instance and generates a list of energies, or un-normalized probabilities; [25]. This notation was chosen to make an analogy with Random Energy Models (REM) from spin-glass and protein folding theories [64, 65].

[7] For practical purposes, the WeightWatcher tool computes $\rho^{emp}(\lambda)$ by forming the Singular Value Decomposition (SVD) of the layer weight matrices $\mathbf{W}$, computing the eigenvalues $\lambda = \sigma^2$ from the singular values σ, and (when useful) smoothing them with a Kernel Density Estimator (KDE). For some calculations, such as the TRACE-LOG condition, we must also select the appropriate normalization of $\mathbf{W}$,



(a) Log-Log ESD

(b) Lin-Lin ESD



(c) Log-Lin ESD

(d) λ_{min} vs KS distance

Figure 1: **Fitting ESDs within `HTSR`.** Depiction of the ESD and results of PL fits for a typical well-trained layer of a modern NN (FC3 of VGG19), including both the actual and good PL fit (red) and a hypothetical bad PL fit (purple). The same ESD is plotted on a Log-Log (a), Lin-Lin (b) and Log-Lin (c) scales. (d) depicts how the start of the PL tail, λ_{min}, varies with the quality of the PL fit (the D_{KS} distance). All plots are generated using the open-source `WeightWatcher` tool. See the main text for details.

Based on empirical results based on thousands of pre-trained models and tens of thousand of layers [25, 26, 24, 28], it is generally observed that the best performing NNs have ESDs that are HT, and whose *tails* of these ESDs, $\rho_{tail}(\lambda)$, can be well fit to a PL, beyond some cutoff $\lambda \geq \lambda_{min}$.[8] For a PL fit,

$$\rho_{tail}(\lambda) := \rho^{emp}(\lambda \geq \lambda_{min}) \sim \lambda^{-\alpha}, \tag{5}$$

where λ_{min} is where the tail of the ESD starts (i.e., it is not the minimum eigenvalue, but the minimum eigenvalue in the tail of the ESD). See Figure 1. As such, the tail of the ESD "starts" at some value λ_{min}, called $xmin$ here, and it continues until the maximum eigenvalue λ_{max}, called $xmax$ here (labeled $xmax$ in the figure, shown by the orange line) We estimate $xmin$ and α jointly, using the method of [67], as implemented in the `powerlaw` python package [68], which is also integrated into the open-source `WeightWatcher` tool [26, 66].[9]

[8] Doing a large meta-analysis like this is tricky; but see [25, 26, 24, 28]. The `WeightWatcher` tool provides a systematic, reproducible way to compute a PL fit (using an MLE method of Clauset et al. [67]), as well other model metrics, including the `SpectralNorm`, `Rand-Distance`, and `AlphaHat` metrics [26]. Also, the ESD $\rho(\lambda)_{tail}$ is sometimes better fit by a Truncated Power Law (TPL), due to finite-size effects. This is important in practice, but we ignore this complexity in this initial discussion of `SETOL`.

[9] The authors of [69] failed to find evidence of a PL-like distribution in NN weight matrices, which is likely to be the case when α and $xmin$ are not estimated *jointly*, as can be seen in Figure 1(d).

HT/RMT Universality class	μ range	α range	Best Fit
RandomLike	NA	NA	MP
Bulk+Spikes	NA	NA	MP+Spikes
Weakly Heavy Tailed	$\mu > 4$	$\alpha > 6$	PL
Heavy (Fat) Tailed	$\mu \in (2, 4)$	$\alpha \in (2, 6)$	PL
Very Heavy Tailed	$\mu \in (0, 2)$	$\alpha \in (1, 2)$	(T)PL
Rank Collapse	NA	NA	NA

Table 1: HTSR Heavy-Tailed Universality classes of RMT. See Table 1 of [25] for more details.

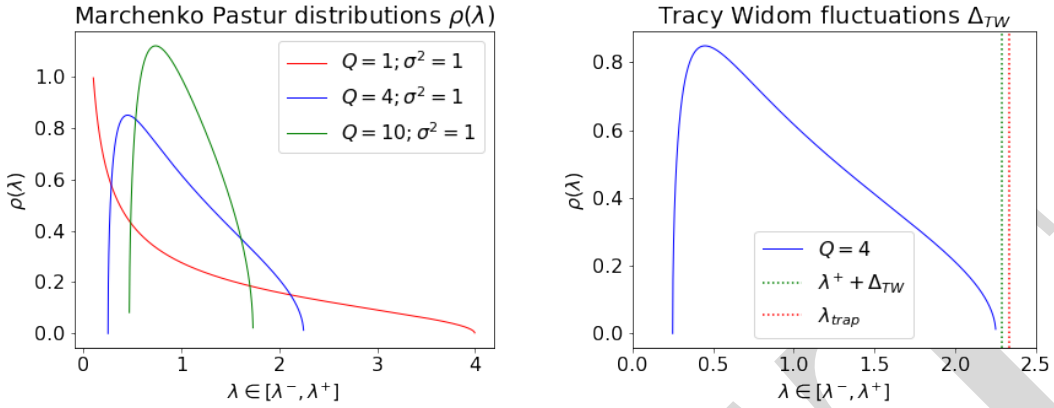
Fitting ESDs. Choosing the start of the tail, λ_{min}, is important for HTSR (and it will be very important for SETOL, as we will describe below). See Figure 1 for a depiction of how this was done within HTSR theory. Figures 1(a)-1(c) show the results of both a "good fit" and a "bad fit" on the same ESD, while Figure 1(d) indicates the quality of fit. For the good fit, the start of the tail is the optimal value $\lambda_{min} = xmin$ (in red); and for the bad fit, it is a suboptimal *bad xmin* (purple). Figure 1(d) depicts how the best fit is determined; it plots $xmin = \lambda_{min}$ versus the D_{KS} value, which is the Kolmogorov-Smirnov (KS) distance between the PL fit and the empirical data [67]. Notice that there are two nearly degenerate minima on Figure 1(d), corresponding to the good fit and the bad fit. It is not uncommon to face such practical challenges, as real-world ESDs are often slightly deformed from a perfect PL density, e.g., they may have two or more near degenerate solutions on the KS plots (d). (They may also have anomalously large eigenvalues; this is discussed in more detail in Section 3.3.)

When one finds a good PL fit for the ESD of a layer $\mathbf{W}$, it provides information about the *Shape* and *Scale* of the ESD of that layer. In particular: the SpectralNorm, λ_{max}, being a matrix norm, is a measure of the size *Scale* of the ESD [24]; the fitted PL exponent Alpha, α, being the slope of the tail of the ESD on a Log-Log plot, describes the *Shape* of the ESD; and the WeightWatcher AlphaHat metric combines *Shape* and *Scale* information. Also, as opposed to other applications of PL fits [67, 70], in our analysis, the start of the tail, $\lambda_{min} = \lambda_{min}^{PL}$, plays a particularly important role because it identifies the subspace of the strongest generalizing eigencomponents (i.e., $\tilde{\mathbf{X}}$, below) in each layer.

2.2 Gaussian and Heavy-Tailed Universality

The HTSR phenomenology uses RMT to classify of the ESD of a layer $\mathbf{W}$ into one of 5+1 Phases of Training, each roughly corresponding to a (Gaussian or HT) Universality class (of RMT or HTRMT). This is summarized in Table 1. A Universality class is a set of matrices having a common limiting spectral distribution, regardless of the other properties of their entries. Of those, the most familiar is the Gaussian class, characterized by the Marchenko Pastur (MP) results from traditional RMT [71, 72]. The Gaussian Universality class, however, is particularly poorly suited for analyzing realistic NNs—precisely becasue the ESDs of SOTA NNs are well-fit by HT distributions. This should not be surprising: weight matrices of realistic NNs do *not* have independent (i.i.d.) entries—their entries are strongly-correlated precisely because they provide a view into the correlated training data.

To model strongly-correlated NN layer matrices, the HTSR phenomenology characterizes NN layer weight matrices in terms of their ESDs (when a good PL fit can be found) by postulating that the (tail of the) eigenvalue spectrum $\rho(\lambda)$ detemines how each layer contributes to the overall generalization. To do this, the HTSR approach models the strong-correlated layer weight matrices *as if* they are actually i.i.d. HT random (i.e., entry-wise uncorrelated) matrices. By doing this,



(a) MP, varying $Q = \frac{N}{M}$

(b) Tracy Widom fluctuations

Figure 2: MP distributions for different aspect ratios Q and variance scales σ^2, and an example of the finite-sized TW fluctuation Δ_{TW}.

one can associate each $\rho(\lambda)$ with the corresponding HT Universality class, according to the PL exponent α fitted from the ESD. As we will see in Section 3.3, it can be critical to distinguish when the ESD is HT *Correlation-wise* vs HT *Element-wise*.

To understand Table 1 better, we first review basic results.

2.2.1 Random Matrix Theory (RMT): Marchenko-Pastur (MP) Theory and Tracy-Widom (TW) Fluctuations

The Marchenko-Pastur (MP) distribution predicts the (limiting) *Shape* of an ESD, $\rho_{MP}(\lambda)$, when the layer weight matrix has elements that are i.i.d. random from the Gaussian Universality class. In particular, the ESD will be MP when the matrix elements are drawn from a Normal distribution $W_{i,j} \in N(0, \sigma^2)$, e.g., as is typical at initialization, before NN training begins. Figure 2 (from Figure 4 of [25]) displays the MP distribution for different aspect ratios $Q = \frac{N}{M}$ and variances σ^2. Notice that the *Shape* is characterized by a well defined, compact envelope with sharp edges.

The MP distribution also predicts the *Scale* of an ESD, again when the layer weight matrix has elements that are i.i.d. random from the Gaussian Universality class. In particular, an MP distribution, $\rho_{MP}(\lambda)$, has very crisp, well-defined lower and upper bounds λ^-, λ^+ [25], and (importantly) the upper bound λ^+ exhibits finite-size Tracy-Widom (TW) fluctuations, $\Delta_{TW}(\lambda)$, which are on the order of $\mathcal{O}(M^{-2/3})$. Thus, any layer eigenvalue with *Scale* greater than this, i.e., $\lambda > [\lambda^+ + \Delta_{TW}(\lambda)]$, is an "outlier" or a "spike."

According to the HTSR phenomenology, these spikes carry significant generalizing information. (This is well-known for Bulk-Plus-Spike models [25], but the HTSR phenomenology generalizes this concept.) Relatedly, for layer matrices $\mathbf{W}$ with aspect ratio $Q > 1$ (i.e., rectangular matrices, where $N > M$), MP RMT predicts there should be no zero eigenvalues, i.e., $\lambda_i > 0$, for all i. Generally speaking, for well trained NNs, for layers with $Q > 1$, all eigenvalues are strictly larger than zero, i.e., well-trained layer weight matrices, with $Q > 1$, should have full rank and exhibit no "rank collapse." HTSR places random Gaussian and "Bulk-Plus-Spike" matrices into the first two rows of Table 1. The essential feature of Gaussian random matrices is that their entries have no correlations. When some correlations are injected, a few large spike eigenvalues form, without otherwise disturbing the shape of the ESD. To really understand how individual NN layers converge, we need to understand when and why their ESDs become HT.

2.2.2 Heavy-Tailed Random Matrix Theory (HTRMT) and Power Law (PL) fits

For very well-trained NN layers, ESDs are *not* MP at all. Frequently, if not always, their ESDs are HT—and they are HT *because* they are strongly-correlated matrices. Importantly, they are *not* HT element-wise. Instead, their entries have a scale, and they have ESDs that are HT due to correlations learned during training. Existing theoretical approaches, including SLT and even StatMech, cannot readily model such strongly-correlated systems.[10]

Such strongly-correlated systems, however, do frequently arise in other, related scientific domains, including in the StatMech of self-organizing systems [75, 76], in electronic structure theory [?, ?, ?], and in quantitative finance [77, 78, 72]. In these (and other) domains, correlated systems frequently exhibit characteristic PL signatures; and it is common practice to *model* correlated systems as random (uncorrelated) systems by using HT statistics (e.g., Levy distributions or PL random matrices), fully understanding that such systems are by no means actually i.i.d. random. The HTSR phenomenology builds upon this longstanding practice by delimiting families of HT NN weight matrices based on the corresponding Universality classes of Pareto matrices.

We explain briefly how to interpret Table 1 with respect to HTRMT. The 5+1 Phases of Training can be identified by fitting ESDs to MP or PL distributions, whichever gives the best fit, as shown in the last column. In case the PL distribution is a better fit, HTSR phenomenology treats the layer weight matrix as equivalent to an i.i.d. random matrix $\mathbf{W}(\mu)$, whose elements have been drawn from a Pareto distribution with exponent μ.

Heavy-Tailed Universality Classes of Random Pareto Matrices For such an element-wise HT matrix, the theoretical *limiting* ESD of a Pareto matrix is also PL, which allows us to related the fitted PL α with exponent $\alpha = a\mu + b$, to the Pareto exponent μ. Ideally, for an infite width matrix , $a = \frac{1}{2}$ and $b = 1$, but due to finite-size effects, however, we have found we must take $a \geq \frac{1}{2}$ and $b \geq 1$, giving

$$W_{i,j}(\mu) \sim \frac{C}{x^{\mu+1}}, \quad \rho(\lambda) \sim \lambda^{-(a\mu+b)}. \tag{6}$$

According to the above relation, we can use either the fitted PL exponent α, or the Pareto exponent μ, to index the HT Universality classes, Note, however, that the finite-size effects strongly depend on the and aspect ratio $Q = N/M$, at least when applied to i.i.d random Pareto matrices, and the (Clauset MLE) PL fit may overestimate the α of the ESD. Table 1 delimits the HT matrices into sub-categories (as shown in the bottom four rows) based on the behaviors of α as a function of μ.

Figure 3 illustrates how the fitted PL exponent α corresponds to the actual Pareto exponent μ for different aspect ratios $Q = M/M$.a Figure 3(a) displays the ESDs of three different i.i.d. 1000×1000 HT random matrices, with $\mu = 1, 3, 5$, on a Log-Log scale. Notice that smaller μ, and therefore smaller α, corresponds to heavier (i.e., larger) tails. Figure 3(b) shows how the empirically fit PL exponent α can vary with the theoretical μ for an associated $\mathbf{W}(\mu)$. For $\mu < 2$ and $Q = 1$, the fitted α follows the linear relation $\alpha = \frac{1}{2}\mu + 1$, albeit with some error. In contrast, for the more relevant $\mu \in (2, 4)$ regime, the relation now depends far more strongly on the aspect ratio Q, and $\alpha \in [2, 6]$. For $\mu > 4$, the fitted α saturates for each specific value of Q.

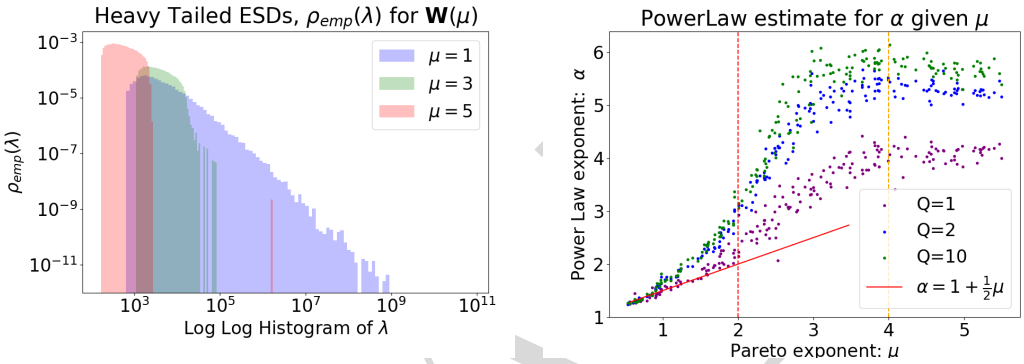
We emphasize that we only model the ESDs of the NN layer weight matrices using the same Universality class to that an associated with the ESD of an random, i.i.d, HT Pareto matrix. In fact, the elements $W_{i,j}$ do not at all appear as if they have been drawn from a HT Pareto distribution, and, in contrast, are almost always well fit to a Laplacian distribution. Also, despite

[10] For example, such theoretical approaches typically deal better with *Scale* information (such as λ_{max}) than with *Shape* information (such as α), e.g., by characterizing an "eigen-gap" separating large eigenvalues from "noise" [73] according to a noise plus low-rank perturbation model [74].

these strong finite-size effects, empirically one finds that the ESDs arising large, well trained, modern NNs can frequently be well fit to a PL (or TPL), and that the fitted $\alpha \in [2, 6]$ for $80-90\%$ of NN layers. Notably, we rarely find $\alpha < 2$ in the best performing, open source, pretrained DNNs.

As there is *no* ground truth whatsoever as to the limiting spectral density of a strongly correlated NN weight matrix (especially without HT elements) the `HTSR` phenomenology uses Pareto matrices as a guide. However, as we will see in Section 3.3, this analogy should be treated with caution because there are cases where it breaks down.

No matter why matrix ESD is HT, it can be difficult to reliably estimating the α parameter when the true α is large. For Pareto matrices of the size investigated here, an observed α above 6 is uninformative — the tail will decay very rapidly indeed, leaving very little of it to study. In this sense, the HT Universality classes are *larger* than the set of only strongly-correlated matrices or Pareto random matrices.



(a) Heavy Tailed ESDs

(b) PL α vs HT μ exponent

Figure 3: Comparison of ESDs and Power Law (PL) exponents α from Heavy-Tailed (Pareto) weight matrices $\mathbf{W}(\mu)$. Subfigure (a) depicts 3 typical ESDs with Pareto exponent $\mu = 1, 3, 5$, each decreasing in *Shape* and *Scale*. Subfigure (b) shows how the exponent α of the PL fit varies with μ, with significant finite-size effects emerging for $\mu > 2$ and $\alpha > 2$.

There is a particularly important boundary between Universality classes where $\alpha = 2$. Recall that one of the properties of power law distributions $\rho(\lambda) \sim \lambda^{-\alpha}$ is that if $\alpha < 2$, than the variance of $\rho(\lambda)$ is infinite. In such cases, the variance cannot be estimated empirically, making $\rho(\lambda)$ in some sense *atypical*. This implies that the NN will have substantially greater difficulty in applying any further load to such a weight matrix. Thus, the value of $\alpha = 2$ is a *critical value*. (See Figure 27 in Section 6.6 for an empirical study of this effect in a small MLP.)

Smaller PL exponent α values correspond to heavier tails, $\rho_{tail}(\lambda)$; and the `HTSR` phenomenology observes that smaller PL exponents α (at least for $\alpha \in (2, 6)$) tend to correspond to better models. This is the key idea of the `HTSR`: the generalizing components of a layer matrix $\mathbf{W}$ concentrate in larger singular vectors associated with the tail, and so that better models have more slowly-decaying (i.e., larger) ESD tails. This differs significantly than simply taking a general low-rank approximation to $\mathbf{W}$, where the rank is chosen without insight from the `HTSR` phenomenology. The `SETOL` theory formalizes this observation as a key assumption. We will revisit these model selection questions in Section 3.1 below.

2.3 Data-Free *Shape* and *Scale* Quality Metrics

The `HTSR` phenomenology provides quality metrics for both individual layers and (by averaging layers) for an entire NN model.

Layer-wise Quality Metrics. Using the HTSR phenomenology, we can define several other *Shape* and/or *Scale* based layer (quality) metrics. These are available in the WeightWatcher tool, and they work very well in practice.

- Alpha (α): $\rho_{tail}(\lambda) \sim \lambda^{-\alpha}$. A *Shape*-based quality metric.
- LogSpectralNorm: $\log_{10} \lambda_{max}$. A *Scale*-based quality metric.
- AlphaHat ($\hat{\alpha}$): $\alpha \log_{10} \lambda_{max}$. A *Scale*-adjusted *Shape*-based quality metric.
- Rand-Distance: $JSD[\rho^{emp}|(\rho^{emp}_{rand})]$. A *Shape*-based, non-parametric quality metric, suitable for highly-accurate, epoch-by-epoch analysis.[11]
- PL KS: D_{KS}. The KS-distance, or quality-of-fit, of the PL fits. For transformers, foundation models, and large, complex, modern NNs, this is frequently an even better model quality metric than the α of the PL fit itself.
- MP SoftRank: $\mathcal{R}_{MP}$. The MP-SoftRank, defined in [25], can be used to identify problems such as when there is significant label or data noise that causes spuriously small α, and also when it is difficult to fit a PL law.[12]

Each of these quality metrics provides a simple characterization of the *Shape* and/or *Scale* of the tail of the ESD of a given layer $\mathbf{W}$. These metrics are related to each other, and they have various trade-offs in practice [26, 24, 28]. Of particular interest here in our development of SETOL are the PL-based WeightWatcher Alpha and AlphaHat metrics.

From Layer-wise Quality Metrics to Layer-Averaged Model Quality Metrics. One can use the HTSR phenomenology to go beyond individual Layer Quality metrics, to construct model quality metrics by averaging Layer Quality metrics (over all layers that are not very small). Existing HTSR model quality metrics assume that all layers are statistically independent, so that the average model quality is just the average of the contributions from each weight matrix $\mathbf{W}$.[13] Given a Layer Quality metric, $\bar{\mathcal{Q}}^{NN}_L(\mathbf{W})$, one can define the *Model Quality* $\bar{\mathcal{Q}}^{NN}$ metric for an entire model as

$$\bar{\mathcal{Q}}^{NN} := \prod_L \bar{\mathcal{Q}}^{NN}_L(\mathbf{W}), \tag{7}$$

a product of each independent Layer Quality $\bar{\mathcal{Q}}^{NN}_L$, and then consider the layer average as the log *Layer Quality*,

$$\log \bar{\mathcal{Q}}^{NN} = \frac{1}{N_L} \sum_L \log \bar{\mathcal{Q}}^{NN}_L = \langle \log \bar{\mathcal{Q}}^{NN}_L \rangle_{\bar{L}} \tag{8}$$

where $\langle \cdots \rangle_{\bar{L}}$ denotes the layer average.

In particular, prior work has used the following metrics:

[11] JSD is the Jensen-Shannon Divergence between the original ESD and the ESD of the layer weight matrix, randomized elementwise.

[12] The WeightWatcher tool also implements the WW-SoftRank, which is like the MP-SoftRank, but replaces λ^{+}_{bulk} with λ^{max}_{rand}; these are mostly equivalent for large matrices, but they can be different for very small matrices.

[13] This independence assumption, clearly a mathematical convenience, gets us closer to a workable theory. One could go beyond a "single layer theory" by adding in intra-layer correlations empirically. The WeightWatcher tool does support this, but doing so is outside the scope of this work.

- The layer-averaged model quality metric `Alpha`, $\log \bar{\mathcal{Q}}^{NN} = \langle \alpha \rangle_{\bar{L}}$, describes the *Shape* of the ESDs. One can use the averaged `Alpha` when studying a single model, and only varying the regularization hyperparameters, although `Alpha` also works very well as a model quality metric when comparing different transformer models [79].

- The layer-averaged model quality metric `LogSpectralNorm`, $\log \bar{\mathcal{Q}}^{NN} = \langle \log \lambda_{max} \rangle_{\bar{L}}$, describes the *Scale* of the ESDs. The averaged `LogSpectralNorm` does work as a model quality metric but not as well as `Alpha` (or `AlphaHat`). Notably, `SLT` predicts that smaller, not larger, `LogSpectralNorm` should be correlated with model quality; the opposite is observed in practice! This is because smaller layer α generally, but not always, corresponds to larger λ_{max}.[14]

- The layer-averaged model quality metric `AlphaHat`, $\log \bar{\mathcal{Q}}^{NN} = \langle \alpha \log \lambda_{max} \rangle_{\bar{L}} = \langle \hat{\alpha} \rangle_{\bar{L}}$, incorporates both *Shape* and *Scale* information. This can compensate for anomalies that can arise when (say) comparing models of different sizes or model qualities [24] or when other issues cause unusualy large λ_{max}. See Section 3.3.1).

The layer-averaged `AlphaHat` model quality metric has been been applied in a large meta-analysis of hundreds of SOTA pre-trained publicly-available NN models in CV and NLP [26, 27, 28, 80]. Generally speaking, `HTSR` shape-based metrics, when used appropriately, outperform all other metrics studied (including those from `SLT`, and with access to the training/testing data,) for predicting the quality of SOTA pre-trained publicly-available NN models. The `HTSR` theory predicts that the best-performing NN models have layers with `Alpha` $\in [2, 6]$, and with $\alpha = 2$ indicating optimal performance. Moreover, prior empirical results show that the `Alpha` and `AlphaHat` metrics can predict trends in the Quality (i.e., the Generalization Accuracy), of SOTA NN models—*even without access to any training or testing data* [26].

[14] The `LogSpectralNorm` can exhibit a Simpson's paradox when segmenting models by quality) [24]. Nevertheless, this metric may be useful when a PL fit can not be obtained, say, when $N \gg M$ and M is very small, as with LSTMs, U-Net architectures, etc.

3 A Semi-Empirical Theory of (Deep) Learning (SETOL)

Based on prior empirical results, and the success of the `Alpha` and `AlphaHat` metrics that are based on the `HTSR` phenomenology, this leads to the deeper question:

> *Why do the* ***`Alpha`*** *and* ***`AlphaHat`*** *metrics work so well as NN model quality metrics for SOTA NN models?*

That is, why do NN models with heavier-tailed layer ESDs tend to generalize better when compared to related models? Relatedly, can we derive these metrics from first principles? (If so, then under what conditions do they hold, and under what conditions do they fail?)
To answer these questions, we will derive a general expression for the Layer Quality, $\bar{\mathcal{Q}}$, of a NN. Although many modern NNs have many layers, we adopt a single-layer viewpoint (like a matrix-generalized Student–Teacher) because in `SMOG` theory the multi-layer generalization can be factorized or approximated. For this, we will obtain by simple averaging our model quality metrics, under effectively a single layer approximation, that correspond to `Alpha` and `AlphaHat`.

In deriving these quantities, we will introduce to NN theory a new Semi-Empirical approach that combines techniques from `StatMech` and `RMT` in a novel way. The Layer Quality $\bar{\mathcal{Q}}$ will estimate the contribution that an individual NN layer makes to the overall quality of a trained NN model. In deriving $\bar{\mathcal{Q}}$, we have discovered a new Layer Quality metric, called the `TRACE-LOG` condition, which indicates the generalizing components of the layer concentrate into a low-rank subspace (the *Effective Correlation Space*, or `ECS`). (Importantly, we have conducted detailed experiments to show that the key assumptions of our `SETOL` theory are valid (see Sections 6.2 and 6.3), and that the empirical estimates of the `SETOL` `TRACE-LOG` condition align remarkably well with predictions from the `HTSR` theory under Ideal conditions (see Sections 6.1). We also examine how the `HTSR` predictions (i.e., the HT PL exponent α) behave under non-Ideal conditions (see Sections 6.5 and 6.6).) In the following, we will outline key conceptual aspects of `SETOL`. In Section 3.1, ; In Section 3.2, ; and In Section 3.3.

3.1 `SETOL` Overview

Our `SETOL` formulates a parametric expression for the Layer Quality $\bar{\mathcal{Q}}$ using a matrix-generalization of the classic Student-Teacher (ST) model from the Statistical Mechanics of Generalization (`SMOG`) theory of the 1990s [8, 81], evaluated using recent advances in the evaluation of so-called HCIZ random matrix integrals [72, 82, 83], such that the final expression for $\bar{\mathcal{Q}}$ can be written in terms of empirically measured statistical properties of the layer ESD. We summarize our basic approach here; see Section 5 for a detailed derivation, and see Section 6 for a detailed empirical analysis.

Following the Student-Teacher (ST) model [8], we first formulate the Generalization Error ($\bar{\mathcal{E}}_{gen}^{ST}$) of the linear Perceptron (in the *Annealed Approximation*, and in the *High-Temperature* limit; see Section 4), and we then generalize this to the case of a NN ($\bar{\mathcal{E}}_{gen}^{ST} \rightarrow \bar{\mathcal{E}}_{gen}^{NN}$), so that we can analyze the Quality of each layer. For the Perceptron, the Generalization Error is an Energy, given as $\bar{\mathcal{E}}_{gen}^{ST} := \langle 1 - R \rangle_{\mathbf{s}}^{\beta}$, where R is the ST vector overlap, and $\langle \cdots \rangle_{\mathbf{s}}^{\beta}$ is a *Thermal Average* (defined in Section 4.2), a Boltzmann-weighted average. In this case, the model Quality, $\bar{\mathcal{Q}}^{ST}$ is exactly the AA, high-T Average Generalization Accuracy $\bar{\mathcal{Q}}^{ST} := 1 - \bar{\mathcal{E}}_{gen}^{ST} = \langle R \rangle_{\mathbf{s}}^{\beta}$. For an MLP or general NN, each layer Energy is associated with a Layer Quality $\bar{\mathcal{Q}}$, which we identify as the average contribution an individual layer makes to the overall generalized accuracy. (i.e $1 - \bar{\mathcal{E}}_{gen}^{NN}$) for a multilayer perceptron (MLP).

For technical reasons (below), we will seek the *Layer Quality (Squared)* $\bar{\mathcal{Q}}^2$, which is defined as

the Thermal Average of the matrix-generalized overlap ($\mathrm{Tr}\left[\mathbf{R}^2\right]$),

$$\bar{\mathcal{Q}}^2 \coloneqq \left\langle \mathrm{Tr}\left[\mathbf{R}^2\right]\right\rangle_{\mathbf{S}}^{\beta} \tag{9}$$

where $\mathbf{R}^2$ can be thought of as a Hamiltonian for the Quality-Squared ($\mathbf{H}_{\bar{\mathcal{Q}}^2} = \mathbf{R}^\top\mathbf{R}$).

In Eqn. 9, the so-called *Teacher* (T) is the NN model under consideration, and $\mathbf{R} \coloneqq \frac{1}{N}\mathbf{S}^\top\mathbf{T}$ denotes the ST overlap operator between the Teacher layer weight matrix $\mathbf{T}$ and a similar *Student* (S) layer weight matrix $\mathbf{S}$. The notation $\langle\cdots\rangle_{\mathbf{S}}^{\beta}$ denotes a Thermal Average over all Student weight matrices $\mathbf{S}$ that resemble the Teacher weight matrix $\mathbf{T}$. By "resemble", the `SETOL` approach assumes that the ESD of $\mathbf{S}$ has the same *limiting* form as $\mathbf{T}$, placing them in the same `HTSR` Universality class. This is made more precise below.

Let us now express the average matrix-matrix overlap $\mathbf{R}$ in squared form using:

$$\begin{aligned}\mathrm{Tr}\left[\mathbf{R}^2\right] &\coloneqq \mathrm{Tr}\left[\mathbf{R}^\top\mathbf{R}\right] \\ &= \frac{1}{N^2}\mathrm{Tr}\left[\mathbf{T}^\top\mathbf{S}\mathbf{S}^\top\mathbf{T}\right] = \frac{1}{N}\mathrm{Tr}\left[\mathbf{T}^\top\mathbf{A}_2\mathbf{T}\right]\end{aligned} \tag{10}$$

where $\mathbf{A}_2$ is the $N \times N$ form of the Student correlation matrix, $\mathbf{A}_2 \coloneqq \frac{1}{N}\mathbf{S}\mathbf{S}^\top$. We will also define the $M \times M$ matrix, $\mathbf{A}_1 = \frac{1}{N}\mathbf{S}^\top\mathbf{S}$, used later.

As explained in Section 4.2, this Quality-Squared is more readily obtained as the derivative of the Layer Quality-Squared Generating Function, $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$, defined as

$$\bar{\mathcal{Q}}^2 \coloneqq \frac{1}{\beta}\frac{\partial}{\partial N}\lim_{N\gg 1}\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} \tag{11}$$

$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$ is essentially (β times) a *Free Energy* for the (approximate) Layer Quality-Squared. (see Section 4, and the Appendix, Section A.3). For more details, see Section 5, and the Appendix, Section A.2).

We can write $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$ as an HCIZ Integral,

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} = \ln\int d\mu(\mathbf{S})\exp\left(N\beta\,\mathrm{Tr}\left[\mathbf{T}^\top\mathbf{A}_2\mathbf{T}\right]\right), \tag{12}$$

The `SETOL` approach then seeks to express $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$ in Eqn. 12 as an HCIZ integral (and in the large-N limit) [72, 82, 83]. We evaluate this at large-N, and write

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1} \coloneqq \lim_{N\gg 1}\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} \tag{13}$$

With these definitions in place, moving forward, the following key assumptions, which can be tested empirically, must hold:

- **The Effective Correlation Space (`ECS`) Condition.** The generalizing components of the Student (and Teacher) layer weight matrices concentrate into a lower rank subspace—the `ECS`—spanned by the eigenvectors associated with the (heavy) tail of the layer ESD $\rho_{tail}(\lambda)$, such that the test error can be reproduced with only these components. We write $\tilde{\mathbf{A}}$ to denote the projection of the correlation matrix $\tilde{\mathbf{A}} \coloneqq \mathbf{P}_{ecs}\mathbf{A}$, onto this subspace, now with rank $\tilde{M} \ll M$. This restricts the measure $d\mu(\mathbf{A})$ to the `ECS`, $(d\mu(\mathbf{A}) \to d\mu(\tilde{\mathbf{A}}))$. This assumption will empirically examined using real-world Teacher weight matrices $\mathbf{T} = \mathbf{W}$ in Section 6.2.

- **The TRACE-LOG Condition.** The Student correlation matrix $\tilde{\mathbf{A}}_1$ (when properly normalized) satisfies the condition that $\mathrm{Tr}\left[\ln \tilde{\mathbf{A}}_1\right] = \ln \det\left(\tilde{\mathbf{A}}_1\right) = 0$, so that the change of measure $d\mu(\mathbf{S}) \rightarrow d\mu(\tilde{\mathbf{A}})$ is Volume Preserving. This condition is derived explicitly in terms of $\tilde{\mathbf{A}}_1$, and therefore will hold for $\tilde{\mathbf{A}}_2$ (and the Teacher Correlation matrix, $\tilde{\mathbf{X}} = \frac{1}{N}\mathbf{T}^\top\mathbf{T}$). Practically, this implies that the $\tilde{M}$ eigenvalues $\tilde{\lambda}$ of the tail of the ESD must satisfy $\sum_{i=1}^{\tilde{M}} \ln \tilde{\lambda}_i \approx 0$. Experiments will test this assumption explicitly in Section 6.3.

Remarkably, both conditions hold best empirically when the HTSR PL quality metric $\alpha \gtrsim 2$ is Ideal. Motivated from these empirical observations, we have:

- $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1}$ **is expressed as an HCIZ integral, at large-N.** We have

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1} = \lim_{N\gg 1} \ln \int d\mu(\tilde{\mathbf{A}}) \exp\left(\beta\, \mathrm{Tr}\left[\mathbf{T}^\top \tilde{\mathbf{A}}_2 \mathbf{T}\right]\right) \tag{14}$$

  where measure $d\mu(\tilde{\mathbf{A}})$ lets us average over all Student Correlation matrices $\tilde{\mathbf{A}}_2$ which lie in the ECS space and which "resemble" the Teacher, where by "resemble" we mean that they share the same form of the tail of their limiting ESDs, i.e., $\rho^{\infty}_{\tilde{\mathbf{A}}}(\lambda) \sim \rho^{\infty}_{\tilde{\mathbf{X}}}(\lambda)$.

- **The Layer Quality (Squared) $\bar{\mathcal{Q}}^2$ is a Norm Generating Function.** The final expression for $\bar{\mathcal{Q}}^2$ can be written as the derivative of $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1}$ as

$$\bar{\mathcal{Q}}^2 = \frac{1}{\beta}\frac{\partial}{\partial N}\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1} = \sum_{i=1}^{\tilde{M}} \mathcal{G}(\lambda_i) \tag{15}$$

  where $\mathcal{G}(\lambda_i)$ is a *Norm Generating Function*, and is defined as the integrated *R-transform* $R(z)$ of the Teacher layer ESD (where $z \in \mathbb{C}$), such that $\mathcal{G}(\lambda) := \int_{\lambda^{ECS}_{min}}^{\lambda} R(z)dz$ and λ^{ECS}_{min} encapsulates the ECS (and selects the desired branch-cut of $R(z)$ so that it is both single-valued and analytic).

To apply the theory, one must choose and R-transform $R(z)$ for the Teacher that models the tail of the ESD $\rho^{emp}_T(\lambda)$, and which can parameterized by some measurable property. This may include the number of Spikes λ^{spike}, the fitted PL exponent α, the maximum eigenvalue λ_{max}, or even the entire tail $\rho^{tail}_T(\lambda)$. This may be a formal expression, a computational procedure, or some combination.

To integrate $R(x)$, however, to have a physicaly meaningful result, one must ensure that $R(z)$ is both analytic and single-valued on domain of interest, namely, the ECS (and therefore the (PL) tail of the ESD), $z \geq \lambda^{ECS}_{min}$. Because the ESD is frequently Heavy-Tailed (HT), this R-transform $R(z)$ may have a branch-cut, and it is expected that this will occur at $z \leq \lambda^{ECS}_{min}$, corresponding roughly at or before the start of the ECS. In a sense, selecting the branch-cut $R(z)$ forces one to define the ECS.

To complete the theopry, we will also show that the HTSR PL Layer Quality metrics Alpha (α) and AlphaHat ($\hat{\alpha}$) can be formally derived directly from the SETOL Layer Quality $\bar{\mathcal{Q}}$ by selecting the appropriate R-transform $R(z)$. In Section 5.4 we provide several possible model $R(z)$ and the resulting Layer Quality $\bar{\mathcal{Q}}$.

Renormalization Group Effective Hamiltonian The formulation of SETOL closely parallels the construction of an Effective Hamiltonian $\mathbf{H}^{ECS}_{\bar{\mathcal{Q}}^2}$ via the Wilson Exact Renormalization Group (RG) approach. Consider a *bare* Hamiltonian $\mathbf{H}_{\bar{\mathcal{Q}}^2}$ for the Layer Quality-Squared, defined as $\mathbf{H}_{\bar{\mathcal{Q}}^2} := \mathbf{R}^\top\mathbf{R}$. We can express Eqn.12 in terms of this bare Hamiltonian $\mathbf{H}_{\bar{\mathcal{Q}}^2}$, and rewrite Eqn.14

in terms of an *renormalized* Effective Hamiltonian $\mathbf{H}_{\bar{\mathcal{Q}}^2}^{ECS}$ that spans the Effective Correlation Space (ECS). Formally, we have:

$$\ln \int d\mu(\mathbf{S}) e^{N\beta \operatorname{Tr}[\mathbf{H}_{\bar{\mathcal{Q}}^2}]} \xrightarrow{RG} \lim_{N\gg 1} \ln \int d\mu(\tilde{\mathbf{A}}) e^{N\beta \operatorname{Tr}[\mathbf{H}_{\bar{\mathcal{Q}}^2}^{ECS}]} \tag{16}$$

where the RG transformation is defined by the Scale-Invariant TRACE-LOG condition, applied at large-N, and where $\mathbf{H}_{\bar{\mathcal{Q}}^2}^{ECS}$ is defined implicitly through result for $\bar{\mathcal{Q}}^2$ (Eqn. 15). The result is, formally, a sum of the integrated R-transforms $\mathcal{G}(\lambda_i)$. In a sense, this result resembles (a non-perturbative form of) the Linked Cluster Theorem in that the log Partition Function is expressed as a sum of integrated matrix-generalized cumulants and/or self-energy. And in analogy with Semi-Empirical theories of Quantum Chemistry, the HTSR Alpha (α) and AlphaHat $\hat{\alpha}$) enter as renormalized empirical parameters. MoSs importantly, the Scale-Invariant TRACE-LOG condition can be verified empirically (See Section 6. Importantly, in analogy with the Wilson RG theory, the HTSR $\alpha = 2$ resembles in spirit an RG *Universal Critical Exponent* at a phase boundary being between the Heavy-Tailed (HT) and the Very Heavy-Tailed (VHT) phase of learning of the HTSR theory. This observation strengthens our argument that the HTSR HT and VHT phases are analogous to the generalizing and overfit phases, resp., of the classical SMOG theories of NN learning.

3.2 Comparing SETOL with HTSR: Conditions for Ideal Learning

The SETOL approach establishes a starting point for developing a first-principles theory for modern NNs. Among other things, by connecting with the HTSR phenomenology, it lets us identify conditions for an Ideal state of learning for an individual NN layer, under the Single Layer Approximation. By Ideal, we mean that the layer is being used most effectively i.e., in some sense it is at its optimal data load, and thus it is conjectured to result in the best model quality.

The Ideal State of Learning is conjectured to be characterized by the following three conditions:

1. the tail of ESD, $\rho_{tail}^{emp}(\lambda)$, can be well fit to a PL of $\alpha \approx 2$: $\rho_{tail}^{emp}(\lambda) \sim \lambda^{-2}$;
2. the eigenvalues in the tail, λ_i, satisfy the TRACE-LOG condition: $\sum_i \ln \lambda_i = 0$; and
3. the generalizing components of the layer concentrate in the singular vectors associated with the tail of the ESD, (Effective Correlation Space).

In Section 6, we will test and justify this conjecture.

These claims are fundamentally about NN learning itself. They are motivated by our formulation of the SETOL approach in our search for a practical predictive theory behind the HTSR phenomenology. When (1) and (2) conditions hold for any layer, we conjecture that (3) holds as well. Moreover, when (1−3) hold for all layers, we conjecture the NN has the lowest Generalization Error (and highest Model Quality) possible for given model architecture and dataset.

Previous results have shown that the HTSR quality metrics (Alpha, AlphaHat, etc.) correlate very well with reported test accuracies, as well as model quality on epoch-by-epoch basis. These results hold because, as indicated by the HTSR theory, the PL exponent α characterizes both the quality of the layer and provides an after-the-fact measure of the amount of regularization.[15] However, the HTSR approach says nothing about the SETOL TRACE-LOG condition; and neither does

[15] By "after-the-fact", we mean that it provides a measure of the regularization in a layer, along the lines of the self-regularization interpretation of HTSR Theory [25]. However, we do *not* recommend that it be used as an explicit regularization parameter. Informally, this is since the "easiest way to obtain HT ESDs is to make weight matrices HT element-wise; but this is *not* what is observed in practice, and thus this is precisely *not* what HTSR Theory and our new SETOL approach are designed to model.

the SETOL approach require a minimum of $\alpha = 2$ to obtain the best model quality, as observed by the HTSR phenomenology. Remarkably, we can show that (1) and (2) do hold *simultaneously*, both in carefully designed experiments on a small model, as well as for many pre-trained, high quality open-source models (such as VGG, ResNet, Llama, Falcon, etc).

HTSR theory, however, has been developed as a phenomenology describing the best-trained, most accurate open-source models available. As such, it may be biased towards such models, and it may not describe less optimal learning scenarios. The keys goals of this work are to derive independent conditions, both theoretical and experimental, that can identify the conditions for Ideal Learning, and to stress-test these conditions in carefully designed, reproducible experiments.

3.3 Detecting Non-Ideal Learning Conditions

The HTSR phenomenology posits that SGD training reduces the *Training Error* by accumulating correlations into the large eigenvalues in NN layer weight matrices $\mathbf{W}$ such that they *self-organize* into a HT with a PL signature, and that this successful self-organization leads to good model quality. Conversely, it also posits that when training has gone awry in some way, the resulting ESD, $\rho^{emp}(\lambda)$, will be deformed in some way. In many practical situations, there can be other, competing factors that give rise to large eigenvalues ($\lambda > \lambda^{+}$) that do not contribute to the generalization capacity of the model, and, consequently, can affect the Scale (i.e., the largest eigenvalue(s) λ_{max}) and Shape (i.e., the PL exponent α, or goodness of fit D_{KS}) of the layer ESDs. These large λ could be Dragon Kings, *Correlation Traps*, or some other anomaly.

In order to apply the HTSR phenomenology most effectively, one must be able to identify various spurious factors and distinguish real correlations from any other large eigenvalues, including the effects of both extreme eigenvalues λ, individual matrix elements $W_{i,j}$, and rank-1 perturbations in $\mathbf{W}$ In one case the ESD is HT primarily due to correlations that help the model generalize, whereas in another when the ESD may be more HT than expected due to suboptimal training, mis-labeled data, etc. In extreme cases, spurious eigenvalues can push the weight matrix into the Very Heavy-Tailed Universality class (i.e., $\alpha < 2$. See Table 1, Section 2), or disrupt the formation of a HT, resulting in a poor PL fit, undermining the core proposition of the HTSR approach.

When training a model with SGD, one may only achieve a sub-optimal result when using overly large learning rates / small batch sizes, (see Section 6), from poor hyper-parameter settings, or simply because direct regularization fails. In such cases, the HTSR approach allows one to detect potential problems by looking for large eigenvalues *not* resulting from correlations.[84]

Importantly, in the context of the SETOL theory, we can now identify such empirical anomalies due to atypical layer weight matrices $\mathbf{W}$, a key factor when models break down. The SETOL approach formalizes the empirical HTSR phenomenology, but, in doing so, assumes that the underlying layer effective correlation matrix $\tilde{\mathbf{X}}$, is typical, meaning that it can describe out-of-sample / test data. Conversely, if the underlying weight matrix $\mathbf{W}$ is atypical, then it is in some sense overfit to the training data and can not fully represent out-of-sample / test data. Consequently, when $\mathbf{W}$ is atypical, we argue that we can observe this, either in the ESD $\rho^{emp}(\lambda)$ directly (i.e., when $\alpha < 2$), and/or having unusually large matrix elements W_{ij}.

We conjecture such sub-optimal results, and particularly those occurring from overfitting, actually arise when the underlying layer weight matrix $\mathbf{W}$ is atypical in some way, in analogy to the results from the classic SMOG theory (see Section 4.1), and, importantly, that we can use the SETOL approach to detect when $\mathbf{W}$ is atypical and therefore a layer is overfit in some way.

Here, we identify two specific cases of atypical weight matrices—Correlation Traps and *Over-Regularization*.[16]

[16] Later, in Section 6, we will show that we can systematically induce both phenomena and observe their effects on the HTSR HT PL metric α and the SETOL TRACE-LOG condition.

3.3.1 **Correlation Traps.** $\mathbf{W}$ is atypical in that $\mathbf{W}$ exhibits an anomalously large mean ($\bar{\mathbf{W}}$). We can observe these by randomizing the layer weight matrix, $\mathbf{W} \rightarrow \mathbf{W}^{rand}$, and then looking for eigenvalues that extend significantly beyond the MP edge of the random bulk (i.e., Spikes).. We call such random spikes *Correlation Traps*, denoted as λ_{trap}, because they appear, in some exteme cases, to be associated with distorted ESDs and, importantly, lower test accuracies. Examples of Correlation Traps are shown in Section 6.5.

3.3.2 **Over-Regularization.** $\mathbf{W}$ is atypical in that $\mathbf{W}$ exhibits an anomalously large variance ($\sigma^2(\mathbf{W})$). We can observe this when the layer $\alpha < 2$. Also, since `Alpha` a measure of implicit regularization, we say the layer with $\alpha < 2$ is *Over-Regularized*. In particular, when one layer is undertrained, having $\alpha > 6$, it appears that other layers may become overtrained to compensate, and this can be seen with having $\alpha < 2$. These effects are studied in Section 6. Additionally, we also observe that when evaluating the `SETOL TRACE-LOG` condition, when $\alpha < 2$, then $\Delta\lambda_{min} < 0$ (see Section 6.3).

3.3.1 Correlation Traps

The first way we identify $\mathbf{W}$ as atypical is when it has an anomalously large mean ($\bar{\mathbf{W}}$); detecting this in general, however, requires more than just examining which elements $W_{i,j}$ are anomalously large.

The `HTSR` phenomenology states that NNs generalize better when their layers ESDs are more HT—precisely because the tail eigenvalues arise from correlations in the weight matrices. So one way is to identify *atypicality* is to look for large eigenvalues that do not arise from correlations in $\mathbf{X}$, but, rather, from one or a few spuriously large matrix elements $W_{i,j}$ and/or rank-1 perturbations in $\mathbf{W}$. We call these eigenvalues Correlation Traps, denoted by λ_{trap} (i.e., see Section 6.5).

Indeed, if we randomize $\mathbf{W}$ elementwise, i.e $\mathbf{W} \rightarrow \mathbf{W}^{rand}$, we expect the $W^{rand}_{i,j}$ matrix elements to be i.i.d and with a small mean (unless something odd happens during SGD training). Likewise, we expect the singular values of $\mathbf{W}^{rand}$ to follow the MP distribution, to within finite-size / TW fluctuations. If we observe an eigenvalue λ_{trap} extending beyond the MP bulk region, $\lambda_{trap} > \lambda^{+}_{bulk}$, then the mean $W_{i,j}$ matrix element will also be anomalously large, and we can identify $\mathbf{W}$ as atypical. We must be careful, however, as we do not fully understand the origin of these atypicalities and do not claim that every one is associated with suboptimal generalization.

By a *Correlation Trap*, we mean that some anomaly in the training of $\mathbf{W}$ resulted in one or more spuriously large eigenvalues λ_{trap} in W^{rand}, and that whatever caused them also may, in some pronounced cases, tend to "trap the correlations in $\mathbf{X}$ itself, preventing them from coalescing into a well defined PL Heavy Tail, or otherwise distorting the ESD. Whether they are a signature of training gone wrong, or whether they distort the dynamics of the tail correlations simply by being there, Correlation Traps can be expected to alter the shape of the ESD, reducing the quality of the PL fits, and sometimes producing spurious α values.

Why would such anomalies occur in a NN? It is conceivable that SGD will, when it fails to find usable correlations, instead produce spurious correlations in the form of large elements and/or rank-1 perturbations. Also, the matrix itself may simply undergo an innocuous mean-shift because the mean is not explicitly controlled during training. Here, mean-recentering may be beneficial. [17]

We will see, below in Section 6, that we can induce a Correlation Trap both by shrinking the batch size, or, equivalently, increasing the learning rate, and that this is associated with degraded model performance and small α. We seek to identify specific ways of identifying such traps because

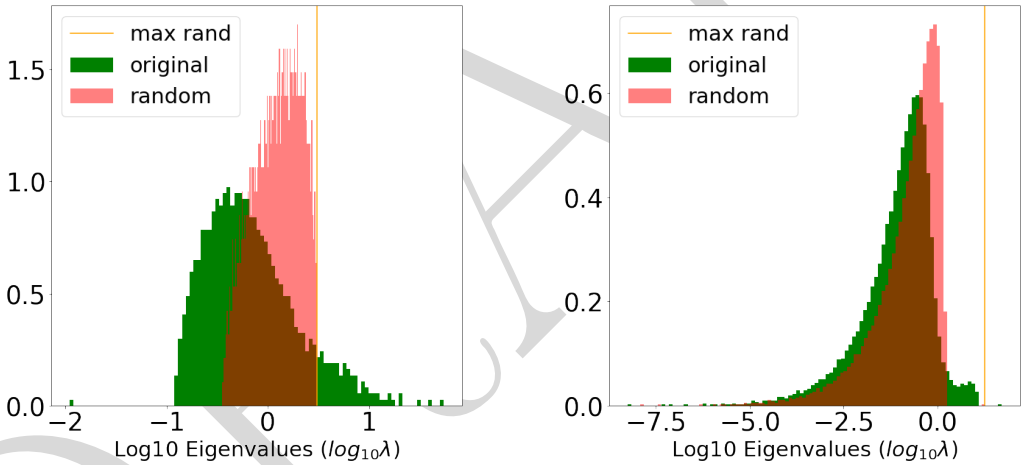
[17] Similarly, when training NNs, frequently the weight matrices [85] or activations [86] may need to be clipped during training to ensure good results.

we reason that the self-organization of correlations may be disrupted by the presence of foreign large eigenvalues – or that failed learning may produce them as a by-product.

Detecting Correlation Traps with `RMT`. RMT suggests that when a matrix $\mathbf{W}$ has unusually large elements W_{ij}, then the ESD will have one or more large eigenvalues λ lying outside the bulk edge λ^{+}_{bulk} of the ESD, as predicted by MP theory. One can detect these so-called Correlation Traps in a weight matrix $\mathbf{W}$ by performing the following:

1. randomize $\mathbf{W}$ element-wise to obtain $\mathbf{W}^{rand}$;
2. compute the ESD for $\mathbf{W}^{rand}$; and
3. look for large eigenvalues $\lambda_{trap} \gg \lambda^{+}_{bulk}$.

`WeightWatcher` looks for Correlation Traps (λ_{trap}) in the ESD of the randomized $\mathbf{W}^{rand}$, that are larger than $\lambda_{trap} > \lambda^{+}_{bulk} + \Delta_{TW}$, where λ^{+}_{bulk} is the MP bulk edge Δ_{TW} are the associated finite-size Tracy Widom (TW) fluctuations. This procedure detects *any* anomaly in the matrix weights that produce spuriously large eigenvalues. It is implemented in `WeightWatcher` (using the randomize option), which was used to generate the plots in Figure 4.



(a) Well-formed ESD

(b) ESD with Correlation Trap

Figure 4: Comparison of a well-formed, Heavy-Tailed ESD (a) to one with a Correlation Trap (b), in the VGG16 model (FC2 layer)

See Figure 4(a), which displays the (log)-ESD of a typical SOTA NN layer $\mathbf{X}$ (green), i.e., on a log-linear scale, along with the (log)-ESD of that layer after randomizing it element-wise (red). The two ESDs differ substantially: the ESD of the original weight matrix $\mathbf{W}$ (green) is very HT, whereas the ESD of the randomized weight matrix $\mathbf{W}^{rand}$ (red) is an MP (and as predicted by the MP `RMT`). The orange line corresponds to the maximum eigenvalue of the randomized ESD. Note that it is at the MP bulk edge of the red plot, indicating that this ESD is not affected by unusually large elements or other weight anomalies. Here, we say that the ESD of $\mathbf{X}$ is HT, and that $\mathbf{W}$ is not HT element-wise. `HTSR` says this layer is well-trained.

Contrast this with Figure 4(b), which displays the (log)-ESD of a NN layer with a Correlation Trap. The ESD of $\mathbf{X}$ (green) is weakly HT, but it looks nothing like the ESD in Figure 4(a).

In fact, it looks very much like the ESD of the randomized weight matrix $\mathbf{W}^{rand}$ (red), except for a small shelf on the right. The orange line again corresponds to the maximum eigenvalue of the randomized ESD, and this is just past this shelf. Relative to the randomized ESD, this line depicts (an) unusually large element(s)—or, equivalently, a rank-1 perturbation of $\mathbf{W}^{rand}$. By a Correlation Trap, we mean that some anomaly in the elements of $\mathbf{W}$ tends to "trap" the ESD of $\mathbf{X}$, concentrating the correlations in $\mathbf{X}$ into the small shelf of density around the orange line. `HTSR` says this layer is not well-trained because it does not have a good PL fit.

3.3.2 Over-Regularization

The second way we identify $\mathbf{W}$ as atypical is when it has an anomalously large variance ($\sigma^2(\mathbf{W})$).

The `SETOL` theory – a single-layer theory of learning – casts the training of a NN layer in terms of how the correlations concentrate into the layer Effective Correlation Space (`ECS`), and becomes exact when the `TRACE-LOG` condition is satisfied. Analogously, the `HTSR` theory – a single layer phenomenology of learning – casts training of an N layer by fitting its ESD to a PL, and noting that the PL exponent α measures the amount of implicit regularization in the layer. Comparing the two approaches, we see that smaller α corresponds to the correlations concentrating into a low-rank `ECS`. In general, and likewise, the more the weight matrix correlations concentrate into a low-rank `ECS`, the better the layer has been regularized. A natural question arises then, namely, can a layer be *Over-Regularized* and can we detect this? and in large, Empirically, we do indeed observe that over the course of training, α decreases, (See Figure 27 (a), Section 6.6,) and that the models predictions are concentrated into the `ECS`, (See Section 6.2). Thus, we also interpret α and `ECS` concentration to be measures of learning itself, meaning that NNs are self-regularizing [25].

Importantly, however, the `HTSR` phenomenology indicates that `Alpha` usually lies in the Fat-Tailed Universality class, such that $\alpha \in [2, 6]$. When $\alpha < 2$, the ESD is Very Heavy Tailed (VHT), and, also, this indicates that $\mathbf{W}$ has an anomalously large variance. That is, $\mathbf{W}$ is atypical. Occasionally, but very infrequently, we do observe $\alpha < 2$, and in large, production quality models (like Llama). Interestingly, we also observe that, frequently, when the `HTSR` $\alpha < 2$, the `SETOL` `TRACE-LOG` condition holds fairly well. This is further explored in Section 6

We have applied the `WeightWatcher` tool to have examined dozens of modern, very large NNs; of particular interest are the so-called Large Language Models (LLMs) that have revolutionized the field of AI. To that end, in Fig. 5, we present the `WeightWatcher` layer `Alpha` metrics for the Falcon-40b and the Llama-65b LLMs. [18]

For the Falcon-40b model, all of the layer `Alpha` range between $\alpha \in [2, 6]$, and therefore lie in the Fat-Tailed Universality class (in Table 1) and are well-fit. In contrast, looking at the Llama-40b layer `Alpha`, very many have $\alpha > 6$, indicating these layers are under-fit, and while a few have $\alpha < 2$, suggesting these are over-fit. Finally, there are more layers with $\alpha \sim 2$ in Llama-65 vs Falcon-40b.

The observations on Llama-2 suggest that the layers with $\alpha \leq 2$ are compensating for the layers with $\alpha > 6$, and yielding suboptimal performance for the Llama-65b architecture. Based on these observations, we hypothesize that, in a multi-layer-perceptron (MLP), when one layer does not or can not learn, then other layers will have to compensate, and will be overloaded with the training data, leading to $\alpha < 2$, and even the `TRACE-LOG` condition $\Delta\lambda_{min} < 0$.

In Section 6.6, we will test this hypothesis. By reducing the trainable parameters in a small MLP, we can simulate the situation seen above in the Llama-65b model, and observe the formation of a Very Heavy Tailed (VHT) ESD in the dominating layer weight matrix. Overloading results from having too few parameters for the complexity of the task. Adding more data increases the load up to the total complexity of the task itself.

[18] Similar results are found for the larger, more modern Llama models, and can be found on the `WeightWatcher` website[66]

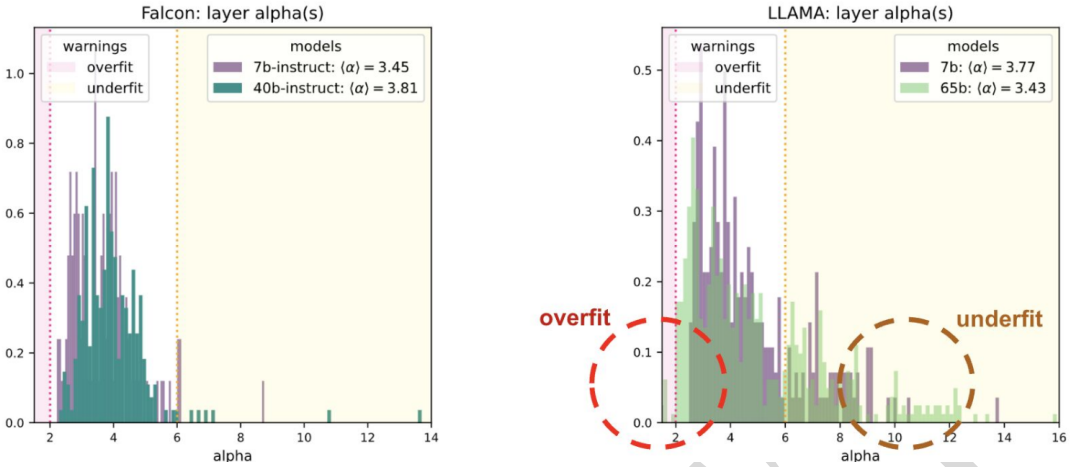


Figure 5: Falcon vs Llama

Moreover, we will also argue that in our experiments, the model enters a kind of glassy meta-stable phase, similar to the kinds of phases predicted by classic `StatMech` theories of learning [8] (described below). Section 6.6 will explore how far we can push the analogy of glassy systems in our experiments to stress test the `SETOL` approach. In particular, we will see effects such as the slowing down of its dynamics, leading to a kind of hysteresis, specific to the under-parameterized regime.

4 Statistical Mechanics of Generalization (SMOG)

In this section, we review the StatMech approach to learning: both to understand how it is usually applied in Statistical Mechanics of Generalization (SMOG) theory; and to understand how our Semi-Empirical approach in SETOL is similar to and different from the traditional approach. We will also obtain an expression for the Generalization Accuracy (or Model Quality $\bar{\mathcal{Q}}^{ST}$) for the classic Student-Teacher (ST) model of the Linear Perceptron (in the AA, and at high-T), as described in [7, 8]. In Section 5, we will generalize this to a Layer Quality metric, $\bar{\mathcal{Q}}$, for a layer in a general Multi-Layer Perceptron (MLP), i.e., $\bar{\mathcal{Q}}^{ST} \rightarrow \bar{\mathcal{Q}}$, so that $\bar{\mathcal{Q}}$ can then be expressed in terms of the ESD of the NN layer.

Outline. Here is an outline of this section.

- **Approaches to the SMOG.** In Section 4.1, we explain the mapping from the StatMech theory of disordered systems to the StatMech theory of NN learning (SMOG); and how our Semi-Empirical approach (SETOL) is similar to and different from the traditional approach.
- **Mathematical Preliminaries.** In Section 4.2, we review mathematical details of StatMech, providing definitions and detailed derivations of quantities and expressions necessary later.
- **Student-Teacher Model.** In Section 4.3, we discuss the setup of the Student-Teacher (ST) model as a general means to estimate the Average Generalization Error empirically. First, in subsection 4.3.1, we describe the ST setup with an operational analogy. Then, in subsection 4.3.2, we derive the (new) result for the ST Model Quality, $\bar{\mathcal{Q}}^{ST}$, using the setup of the classic (ST) model for the Generalization Error (and accuracy) of the Perceptron (in the AA, and at high-T).

Additional information can be found in the Appendix.

- **Symbols and Equations.** In Section A.1, we summarize the important symbols and key results, including the dimensions of the vectors and matrices, different notations for energies, and key equations.
- **Summary of the SMOG.** In Section A.2, we provide a more detailed analysis of the results we derive in Section 4.3.2. In particular, in Section A.2.1, we repeat the derivations of the ST Generalization Error $\bar{\mathcal{E}}^{ST}_{gen}$ and related quantities (from Section 4.2), using more concrete notation to align with [7, 8]; and in Section A.2.2, we use this to derive the matrix-generalization of the ST Annealed Error Potential $\epsilon(R)$ (as well as the normalization for the weight matrices, necessary for later).

4.1 StatMech: the SMOG approach and the SETOL approach

In this subsection, we review the basic StatMech setup necessary to understand SMOG theory as well as SETOL. This theory was developed in the 1980s and 1990s [7, 8, ?, ?, 87].

Traditional SMOG theory. In traditional SMOG theory, one maps the learning process of a NN to the states and energies of a physical system. The mapping is given in Table 2. SMOG theory models the SGD training of a *Perceptron*, on the data, $\mathbf{x}^n$, to learn the optimal weights, $\mathbf{w}$, as a Langevin process.[19] The power of the StatMech approach comes from the fact that the core

[19] Typically, we have no guarantee of the true equilibrium in a high-dim nonconvex landscape; so, when the *Thermodynamic limit* exists, the Langevin process converges or relaxes to the thermodynamic equilibrium.

Statistical Physics	Neural Network Learning
Gaussian field variables	Gaussian i.i.d data $\boldsymbol{\xi}^N \in \mathcal{D}$
State Configuration	Trained / Learned weights $\mathbf{w}$
State Energy Difference	Training and Generalization Errors $\bar{\mathcal{E}}_{train}, \bar{\mathcal{E}}_{gen}$
Temperature	Amount of stochasticity present during training T
Annealed Approximation	Average over the data first
Thermal Average	Expectation w.r.t. the distribution of trained models
Free Energy	Generating function for the error(s) F

Table 2: Mapping between states and energies of a physical system and parameters of the learning process of a neural network.

SETOL Terminology	Explanation
Model Quality $\bar{\mathcal{Q}}$	Generalization accuracy, in the AA and at high-T
Layer Quality $\bar{\mathcal{Q}}$	Layer contribution to the accuracy, in the AA and at high-T
Layer Quality-Squared $\bar{\mathcal{Q}}^2$	Layer Quality squared, used for technical reasons
Quality Generating Function $\Gamma_{\bar{\mathcal{Q}}}, \Gamma_{\bar{\mathcal{Q}}^2}$	Generating function for quality
Annealed Hamiltonian H^{an}	Energy function, for errors or accuracies
Effective Hamiltonian H^{eff}	Exact energy function, but restricted to a low-rank subspace

Table 3: Additional terminology introduced for the `SETOL`. Notice that the Quality Generating Function Γ is simply one minus the Free Energy, $\Gamma := 1 - F$, but it introduced because sign convention for the Free Energy is always decreasing with the error. In contrast, we define the Hamiltonian in terms of the model error or accuracy, depending on the context.

concept of Thermal average corresponds to taking the expectation of a given quantity only *over the set of trained models*, as opposed to uniformly over all possible models (or, in a worst-case sense, over all possible models in a model class). This capability is particularly compelling in light of the `StatMech` capacity to characterize disordered systems with complex non-convex energy landscape (which can even be *glassy*, characterized by a highly non-convex landscape [8, 81, 87], and recognized classically by a slowing down of the training dynamics [88]). Thus, concepts such as training and Generalization Error arise naturally from integrals that are familiar to `StatMech`; and theoretical quantities such as Load, Temperature, and Free Energy also map onto useful and relevant concepts [89]. The Free Energy is of particular interest because it can be used as a generating function to obtain the desired Generalization Error and/or accuracy. We wish to understand how to compute the associated thermodynamic quantities such as the expected value of the various forms of the Average Generalization Error ($\bar{\mathcal{E}}_{gen}$), Partition Function (Z), and the Free Energy (F) and other Generating Functions (Γ).

The Student-Teacher model. We seek to compute and/or estimate the Average Generalization Accuracy for a *fixed* Teacher Perceptron T by averaging over an ensemble of Student S Perceptrons, in the Annealed Approximation (AA), and at High-Temperature (high-T); we call this ST Model Quality, and denote it $\bar{\mathcal{Q}}^{ST}$. We will then, later, in Section 5, we generalize $\bar{\mathcal{Q}}^{ST}$ to an arbitrary

layer in Multi-Layer Perceptron, giving a Layer Quality, i.e., $\bar{\mathcal{Q}}^{ST} \rightarrow \bar{\mathcal{Q}}$. This formulation of the ST problem is different than the classic approach because one usually does not fix the Teacher but, instead, averages over all Teacher vectors $\mathbf{t}$ [8, 87]. This is one of many ways that distinguishes the current approach from previous work. Because of this, we present both a pedagogic derivation of $\bar{\mathcal{Q}}^{ST}$ (for a general NN in Section 4.2, and for the ST model specifically in the Appendix, Section A.2), and we provide a more heuristic approach in Section 4.3.2, assuming the AA and high-T at all times.

Towards a Semi-Empirical Theory. In the `SETOL` approach to `StatMech`, we want a matrix generalization of the ST Model Quality, $\bar{\mathcal{Q}}^{ST}$, for a single Layer Quality $\bar{\mathcal{Q}} \sim \bar{\mathcal{Q}}_L^{NN}$ in an arbitrary Multi-Layer Perceptron (MLP). This matrix generalization is a relatively straightforward extension of the classical (i.e., for a vector Teacher) `SMOG` ST Model Quality (but our `SETOL` approach will use it in a conceptually new way). In our matrix generalization, the Teacher vector $\mathbf{t}$ is replaced by a Teacher matrix $\mathbf{T}$ (i.e., $\mathbf{t} \rightarrow \mathbf{T}$); and, in our `SETOL` approach, $\mathbf{T}$ will be an actual (pre-)trained NN weight matrix (i.e., $\mathbf{T} = \mathbf{W}$). Importantly, this matrix $\mathbf{W}$ is neither a Gaussian Random Matrix, nor is it obtained from Gaussian i.i.d training data. As such, for our `SETOL` theory, we seek an expression that can be parameterized by the Teacher, and in particular by the ESD of the Teacher. This is what makes the basic method Semi-Empirical: even though we do not know the form of the Teacher, we make a modeling assumption that the Teacher has the same limiting spectral distribution as the Student, and hence the same PL fit parameter α. This is all done with the understanding that later we will augment (and hopefully "correct") our mathematical formulations with phenomenological parameters fit from experimental data. To make the Semi-Empirical method a Semi-Empirical *Theory*, we not only seek to parameterize our model; but we also try to understand how to derive `HTSR` empirical metrics, such as `Alpha` and `AlphaHat`, how they arise from this formalism, how they are related to the correlations in the data, and why they are transferable and exhibit Universality. This gives what we call a Semi-Empirical *Theory*.

4.2 Mathematical Preliminaries of Statistical Mechanics

SubSection Roadmap Briefly, in the following subsection, we start by defining an arbitrary NN model, with weights $(\mathbf{w})$ Then, we explain the difference between using real-world $(\mathbf{x})$ and random $(\boldsymbol{\xi})$ data This lets us define an energy error function, $\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w})$, the error the NN makes on the data. We then explain how to take different kinds of *Thermodynamic* averages of the data, including *Sample* and *Thermal average* and the implications, and the difference between computing errors and accuracies. Next, we define the *FreeEnergy* (F) for the error(s), and the *GeneratingFunction* (Γ) for the accuracy and/or quality. From here, we explain the *AnnealedApproximation* (AA) and how to define the *Annealed Hamiltonian*, $H^{an}(\mathbf{w})$, a crucial expression that will be the starting point later for our matrix model. In the AA, $H^{an}(\mathbf{w})$ simplies to $H^{an}_{hT} = \epsilon(\mathbf{w})$, where $\epsilon(\mathbf{w})$ is an Annealed Error Potential that depends only on the weights $\mathbf{w}$. Likewise, we can define the Self-Overlap, $\eta(\mathbf{w}) := 1 - \epsilon(\mathbf{w})$, which is useful for obtaining the Quality. We show how to obtain the *Average Training and Generalization Errors* $\bar{\mathcal{E}}_{train}$, $\bar{\mathcal{E}}_{gen}$ using the `StatMech` formalism, which defines them in terms of partial derivatives of the Free Energy (F). Doing this, we show that in the AA and at high-T they are equivalent, $[\bar{\mathcal{E}}^{ST}_{train}]^{an,hT} = [\bar{\mathcal{E}}^{ST}_{gen}]^{an,hT}$, and can both be expressed as a Thermal Average over all Students, as a function of the Teacher, as $[\bar{\mathcal{E}}^{ST}_{gen}]^{an,hT} = \left\langle H^{an}_{hT}(R) \right\rangle^{\beta}_{\mathbf{s}} = \langle \epsilon(R) \rangle^{\beta}_{\mathbf{s}}$. Note that these averages are obtained by using the Free Energy as a Generating Function. We then explain how to obtain the Model Quality as partial derivatives of a *Generating Function* $(\Gamma_{\bar{\mathcal{Q}}})$. We then discuss the more advanced techniques, the *Large-N Approxmation* and the *SaddlePointApproximation* (SPA), which will be used extensively later.

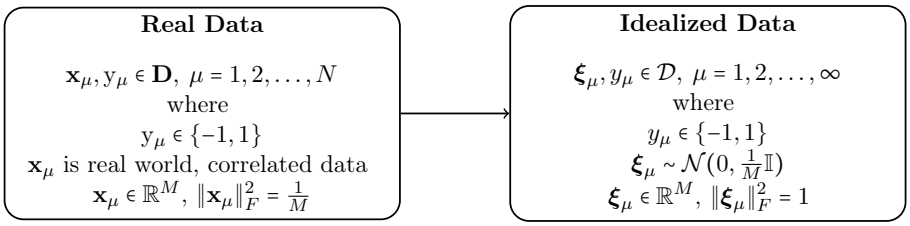


Figure 6: Mapping from a fixed set of $n = N$ real-world, correlated data instances $[\mathbf{x}, \mathrm{y}] \in \mathbf{D}$ to an uncorrelated, random model of idealized data $[\boldsymbol{\xi}, y] \in \mathcal{D}$, drawn from a Gaussian i.i.d. distribution.

Finally, we introduce HCIZ integrals, which will be necessary to evaluate the matrix-generalized form of $\Gamma_{\bar{\mathcal{Q}}}$ to obtain the final result.

In this subsection, we will compare and contrast several types of averages and energies we will encounter.

4.2.1 **Setup** In Section 4.2.1, we will start by describing the basic setup of the problem, including the distinction between the actual training process and how we model the training process.

4.2.2 **Sample Averages, Expected Values, and Thermal Averages** In Section 4.2.2, we will describe Thermal Averages (over the weights $\mathbf{w}$) and Sample Averages (over the data $\mathbf{x}$)—in particular, under the Annealed Approximation (AA) and in the High-Temperature (High-T) limit—showing how they relate to each other and to the notion of Replica Averages.

4.2.3 **Free Energies and Generating Functions** In Section 4.2.3, we will make a connection between these different averaging notions and Free Energies and Generating Functions, showing how they relate to each other.

4.2.4 **The Annealed Approximation (AA) and the High-Temperature Approximation (high-T)** In Section 4.2.4, we explain the Annealed Approximation, the High-Temperature approximation, and the Thermodynamic Large-N limit and the Saddle Point Approximation (SPA). We also introduce the Quality Generating Function $\Gamma_{\bar{\mathcal{Q}}}$

4.2.5 **Average Training and Generalization Errors and their Generating Functions** In Section 4.2.5, we will show how to compute the Average Training and Test/Generalization Errors $\bar{\mathcal{E}}_{train}, \bar{\mathcal{E}}_{gen}$ using the Free Energy as a Generating Function, and how these errors are related to each other in the AA and High-T limit. We will also make connections with the Saddle Point Approximation (SPA).

4.2.6 **The Thermodynamic Large-N limit and the Saddle Point Approximation (SPA)** In Section 4.2.7, we discuss the large-N Thermodynamic limit and the Saddle Point Approximation (SPA). We also the introduce concept of Self-Averaging.

4.2.7 **HCIZ Integrals** Finally, in Section 4.2.8, we will describe how to express the Free Energy as a matrix-generalized Thermal Average over random matrices, called an HCIZ integral.

The various symbols and other important results are summarized in the Appendix A.1

4.2.1 Setup

A basic issue in formulating `SETOL` is that one typically trains one large (expensive) NN, i.e., one does not split the data into training and testing sets. Thus, we want a methodology to

approximate quantities such as the generalization error or generalization accuracy that does not rely on traditional train-test splitting methods. To accomplish this, we idealize the empirical data distribution as Gaussian fields, and we will use `StatMech` to construct quantities (basically, free energies or generating functions) so that we can compute training/testing errors by taking appropriate derivatives of these quantities.

In more detail, we imagine training a NN on n training data instances, $\mathbf{x}_\mu$, which are m-dimensional vectors, with labels y_μ, chosen from a large but finite-size training data set $\mathbf{D}$. The goal of training the NN is to learn the m weights of the vector $\mathbf{w}$ (or, later, a weight matrix $\mathbf{W}$) by running gradient descent to minimize a loss function $\mathcal{L}$ (ℓ_2, cross-entropy, etc.). We want to approximate the actual network's learning dynamics by an analytically tractable ensemble so that we can then obtain analytic expressions for the *Free Energy* and *Generating Function* we then use to compute Thermodynamic averages such as the *Average Generalization Error* ($\bar{\mathcal{E}}_{gen}$) and *Model Quality* ($\bar{\mathcal{Q}}$) (which is our approximation to the *Average Generalization Accuracy*).

Counting Samples and Features: n, m, N, and M. We let the number of training samples be n and the dimension (i.e., number of features) for each sample be m. For simplicity, we also use N (instead of n) and M (instead of m), recognizing that *in later sections* N and M will refer to the dimensions of a layer's weight matrix (i.e., an $N \times M$ matrix). We stress that here, in this subsection, $n = N$ and $m = M$ only hold for our immediate analysis, to avoid extra notation. When we move to matrix-based analyses, we will revisit (and possibly distinguish) N (layer input dimension) and n (training-set size).

Symbol	Definition	Vector	Matrix
n	Total number of free parameters in the model.	n	$N \times M$
m	Number of features per training sample.	m	—
N	Input dimension of a layer's weight matrix.	—	N
M	Output dimension of a layer's weight matrix.	—	M

Table 4: In the *vector* model, lowercase n is the dimension of the weight vector (total parameters), and lowercase m is the number of features per sample. In the *matrix* model, uppercase N and M are the input and output dimensions of the weight matrix, resp., giving $n = N \times M$ free parameters.

Actual and Idealized Data and Energies. Consider having a large set $n = N$ of actual, real-world data,

$$\mathbf{x}_\mu, \mathrm{y}_\mu \in \mathbf{D},\ \mu = 1, \cdots, n, \tag{17}$$

where $\mathbf{x}_\mu$ is an m-dimensional real vector, $\mathbf{x}_\mu \in \mathbb{R}^m$, y_μ is a binary label taking values 1 of -1, denoted $\{-1, 1\}$, and $\mathbf{D}$ denotes the finite-size dataset.

Notice that $\mathbf{x}_\mu \in \mathbb{R}^m$ is normalized such that the Frobenius norm squared is $\frac{1}{m}$:

$$\|\mathbf{x}_\mu\|_F^2 := \sum_{i=1}^{m} \mathbf{x}_{\mu,i}^2 = \frac{1}{m} \tag{18}$$

We call $\mathbf{x}^n$, an n-dimensional sample (of the training data instances $\mathbf{x}$) from $\mathbf{D}$. We may or may not specify the labels for this sample, depending on the context

We associate model errors with a (change in) energy $\Delta E_{\mathcal{L}}$. Smaller energies correspond to smaller errors and therefore better models. For example, for the mean-squared-error (MSE) loss,

one has

$$\Delta E_{\mathcal{L}}(\mathbf{w}, \mathbf{x}_\mu, \mathrm{y}_\mu) \coloneqq (\mathrm{y}_\mu - E_{NN}(\mathbf{w}, \mathbf{x}_\mu))^2, \tag{19}$$

where $E_{NN}(\mathbf{w}, \mathbf{x}_\mu)$ is output prediction of the NN, as in Eqn. 1.

To estimate quantities such as the generalization error or generalization accuracy, we will adopt an approach that involves replacing the real data with Gaussian data and the NN with a parametric model that we will fit with a Semi-Empirical procedure (described later). Following the usual Statistical Mechanics trick, we idealize the inputs as independent Gaussian fields.

$$\mathbf{D} \to \mathcal{D}, \;\; \mathbf{x}_\mu \to \boldsymbol{\xi}_\mu, \;\; \mathrm{y}_\mu \to y_\mu, \tag{20}$$

where we denote the model training and/or test data instances as $(\boldsymbol{\xi}, y)$ such that

$$\boldsymbol{\xi}_\mu, y_\mu \in \mathcal{D}, \; \mu = 1, \cdots, \infty. \tag{21}$$

Here, $\boldsymbol{\xi}_\mu$ is a random vector (i.e., an m-dimensional random variable, $\boldsymbol{\xi}_\mu \in \mathbb{R}^m$), sampled from an i.i.d Gaussian distribution $\mathcal{D}$, and y_μ denotes the (binary) label and/or NN output.

4.2.2 BraKets, Expected Values, and Thermal Averages

Given the setup from Section 4.2.1, we will want to model the average (change in) energy, $\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \mathbf{x}^n)$, averaged over some n-size data set $\mathbf{x}^n$. We can write the Total Data Sample Error as using an overloaded, operator notation

$$\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \mathbf{x}^n, \mathrm{y}^n) \coloneqq \sum_{\mu=1}^{n} \Delta E_{\mathcal{L}}(\mathbf{w}, \mathbf{x}_\mu, \mathrm{y}_\mu), \tag{22}$$

where the boldface $\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w})$ indicates this a sum over the entire set of n pairs $[\mathbf{x}^n, \mathrm{y}^n]$. We should keep in mind that this depends on the specific set of n data pairs $[(\mathbf{x}_\mu, \mathrm{y}_\mu) \in \mathbf{D} | \; \mu = 1, \cdots, n]$, although later we will model the labels y_μ as the output of another NN when describing the Student-Teacher model. For that reason, for now, we will assume that the y_μ is *implicit* in ΔE_{ℓ_2}, will drop the y_μ and y^n symbols, and just write this total error / energy difference as

$$\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \mathbf{x}^n) \coloneqq \sum_{\mu=1}^{n} \Delta E_{\mathcal{L}}(\mathbf{w}, \mathbf{x}_\mu), \tag{23}$$

which is now a function of the entire set of n vectors $[\mathbf{x}^n]$ (where the labels y have been set implicitly). [20] This operator notation will provide useful later in Section 4.3.2 (see Eqn. 83) and in Appendix A.2.

We will not, however, work directly with samples and sample averages. Instead, we will model them. To that end, we need to estimate them with a theoretical approach. For example, we can write the Total Data Sample Error in terms of our random data variables $\boldsymbol{\xi}$ formally as

$$\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n) \coloneqq \sum_{\mu=1}^{n} \Delta E_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}_\mu), \tag{24}$$

but to evaluate this we need to take an integral and/or Expected Value over the data sample $\boldsymbol{\xi}^n$.

[20] In the classic Student Teacher model, the labels y^n represent the Teacher outputs and are effectively treated as either uniform random variables to be averaged over later, or as the outputs of an optimal Teacher. In this work, the Teacher is fixed so we can drop the labels.

Expected Values. We need to compute various sums and integrals, sampling from a model $\mathcal{D}$ for the actual data distribution $\mathbf{D}$, which will frequently (but not always) be defined as more familiar Expected Values. We will denote Expected Values using physics Bra-Ket notion. Importantly, we use the term Expected Value in the physics sense, and BraKets will denote an un-normalized sum or integral; when the quantity is to normalized, we will denote the normalization explicitly. For example, given a function $f(\boldsymbol{\xi})$, we write the BraKet integral as:

$$\langle f(\boldsymbol{\xi})\rangle_{\boldsymbol{\xi}} \coloneqq \int d\mu(\boldsymbol{\xi}) f(\boldsymbol{\xi}), \tag{25}$$

and if we want to express an n-dimensional average over $f()$ then we would express this as

$$\langle f(\boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^n}} \coloneqq \frac{1}{n}\int d\mu(\boldsymbol{\xi}^n) f(\boldsymbol{\xi}^n) \tag{26}$$

where the BraKet $\langle\cdots\rangle_{\overline{\boldsymbol{\xi}^n}}$ deotes the integral over the idealized Gaussian-field data $\boldsymbol{\xi}^n$, but with the convention that the normalization $\frac{1}{n}$ appears inside the Bra-Ket implicitly. If this integral is normalized properly, then this denotes the familiar Expected Value $\mathbb{E}_{\boldsymbol{\xi}}[f(\boldsymbol{\xi})]$. For a more complicated example, consider how to compute Expected Value of the Data Sample Error. That is, we want to model the average Data Sample Error using:

$$\frac{1}{n}\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w}, \mathbf{x}^n) \xrightarrow{\text{Expected Value}} \langle \boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^n}}, \tag{27}$$

In this case, we obtain:

$$\begin{aligned}\langle \boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^n}} \coloneqq& \frac{1}{n}\int d\mu(\boldsymbol{\xi}^n)\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n) \\ =& \frac{1}{n}\int \prod_{\mu=1}^{n} d\boldsymbol{\xi}_\mu P(\boldsymbol{\xi}^n)\sum_{\mu=1}^{n}\Delta E_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}_\mu),\end{aligned} \tag{28}$$

where $P(\boldsymbol{\xi}^n)$ is an n-dimensional probability distribution (i.e., an n-dimensional Gaussian distribution), normalized to 1, and and where the subscript $\overline{\boldsymbol{\xi}^n}$ on the Ket reminds us this is an average or Expected Value of a finite, n-size sample of the ideal data. (This is used in both Sections 4 and 5.) The normalization $\frac{1}{n}$ is included in the defintion to ensure the Bra-Ket is a proper Expected Value. The measure $d\mu(\boldsymbol{\xi}^n)$ signifies a n i.i.d Gaussian vector $\boldsymbol{\xi}$, drawn from an m-dimensional idealized data vectors $\boldsymbol{\xi}$. Also, in some cases, we make use the subscript $\boldsymbol{\xi}^n$ on the Ket as $\langle\cdots\rangle_{\boldsymbol{\xi}^n}$; this represents an integral over the data, but not an average or Expected Value.

Size-Extensivity, Size-Intensivity, and Size-Consistency A key requirement for the Thermodynamic limit in `StatMech` is *Size-Extensivity*: that physically meaningful quantities (i.e, total energies and free energies) scale linearly with the system size, $n = N$. Roughly, extensive quantities scale with system size, intensive ones do not. Along with this, Thermodynamic average quantities should be *Size-Intensive*, meaning that they remain independent of $n = N$ as the system size increases. In our setting, Size-Extensivity and Size-Intensivity underpin the large-N limit, ensuring that macroscopic observables become independent of microscopic fluctuations.

As an example of Size-Extensivity and Size-Intensivity, we write the Expected Value (i.e., the data-average) of Data Sample Error $\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)$ (Eqn. 28) in the large-N limit as

$$\lim_{n\gg 1}\langle \boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^n}} = \lim_{n\gg 1}\frac{1}{n}\int \prod_{\mu=1}^{n} d\boldsymbol{\xi}_\mu P(\boldsymbol{\xi}^n)\sum_{\mu=1}^{n}\Delta E_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}_\mu). \tag{29}$$

Here, the notation ($n \gg 1$) means n grows arbitrarily large, but is not necessarily at the limit point ($n = \infty$). The Total Data Sample Error $\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)$ is Size-Extensive, whereas the average $\langle \mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n) \rangle_{\overline{\boldsymbol{\xi}^n}}$ is Size-Intensive. This limit will be implicit later when taking a Saddle Point Approximation (see below). [21]

The data-averaged error $\langle \mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n) \rangle_{\overline{\boldsymbol{\xi}^n}}$ will appear frequently below. For compatibility with [8], we denote it using the symbol $\epsilon(\mathbf{w})$:

$$\epsilon(\mathbf{w}) := \lim_{n \gg 1} \langle \mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n) \rangle_{\overline{\boldsymbol{\xi}^n}} \quad \text{(Size-Intensive)}. \tag{30}$$

where, by our normalization here, $\epsilon(\mathbf{w}) \in [0, 1]$. The symbol $\epsilon(\mathbf{w})$ is our theoretical estimate of the sample average $\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)$ (Eqn. 31), well-defined for any n. We also call $\epsilon(\mathbf{w})$ the *Annealed Error Potential*, which will be made clear below.

It is also convenient to write *Total Annealed Error Potential* as an Energy,

$$\mathcal{E}(\mathbf{w}) := n\epsilon(\mathbf{w}) \quad \text{(Size-Extensive)}. \tag{31}$$

This will only be useful when the Thermodynamic limit exists, and this can be reasonably expected for the Annealed Approximation (AA), which is the regime in which `SETOL` will be developed.[22] There is a related notion of *Size-Consistency*, often introduced in Quantum Chemistry through the Linked Cluster Theorem [48, 90], which states that the *average* energies and/or free energies ($\bar{F}$) of in particular a correlated system scale with $\mathcal{M}$, the number of "independent interacting components":

$$\begin{aligned} -\beta \bar{F} = \langle \ln Z \rangle_{\overline{\boldsymbol{\xi}^n}} &= \sum_{\mu=1}^{\mathcal{M}} (\text{Connected Components}) \\ &= \sum_{\mu=1}^{\mathcal{M}} \text{Cumulants}(\mu) + \text{higher order terms} \end{aligned} \tag{32}$$

For `SETOL`, below, these connected components will be matrix-generalized cumulants from RMT. For NNs, Size-Consistency appears when scaling the number of features in our matrix model, and it ensure that our layer and model Qualities remain well-behaved as we increase M, just as we do for N. For a simple example, see Appendix A.2 where we derive the expression for the matrix-generalized Annealed Hamiltonian H^{an}. Both Size-Extensivity (in N) and Size-Consistency (in M) are crucial in our `SETOL` analysis: they justify taking the large-N approximation for matrix integrals, and they ensure our resulting the HCIZ integral–a sum of integrated RMT cumulants (below)– scales with the dimension of the Effective Correlation Space (`ECS`), $\mathcal{M} = \tilde{M}$.

From Errors to Accuracies: The Average Generalization Accuracy, the Quality, and the Self-Overlap. We have been discussing various forms of errors. In `SETOL`, we will, however, primarily be concerned with approximating the *Average Generalization Accuracy*, or, more generally, the Quality of a NN model and/or its layers. [23] The average accuracy is simply one minus the error. To represent this, we introduce the *Self-Overlap* η, which is defined generally as

$$\eta(\mathbf{w}) := 1 - \epsilon(\mathbf{w}) \in [0, 1], \tag{33}$$

[21] As we are working within a "physics-level of rigor, we take some liberties in evaluating these large-N limits; and we leave the formal proofs for future work.

[22] We should note that, while our model training and generalization errors are always expressed energies, an energy is not necessarily a model error.

[23] Technically, the Quality will estimate the average *Precision* rather than the Accuracy. This will distinction will be clarified in the Section 4.3.

and which describes the "overlap" between the true and the predicted labels. Unlike here, however, in later sections (4.3.2, 5.1, and Appendix A.3) we will first define a data-dependent Self-Overlap, so that we may obtain $\eta(\mathbf{w}) \coloneqq \langle \eta(\mathbf{w}, \boldsymbol{\xi}) \rangle_{\overline{\boldsymbol{\xi}^n}}$ directly.

Braket Notation. We will use physics Bra-Ket notation, $\langle \cdots \rangle$, to denote different kinds of sums and integrals, with superscripts and subscripts, and for Expected Values (estimated theoretical averages). We use superscripts to denote the kind of integral or average:

$$\text{Thermal } \langle \cdots \rangle^{\beta}, \text{ Annealed } \langle \cdots \rangle^{an}, \text{ high-T } \langle \cdots \rangle^{hT}, \text{ HCIZ } \langle \cdots \rangle^{IZ}, \text{ etc.}$$

We use subscripts to emphasize the dependent variables:

$$\text{weights } \langle \cdots \rangle_{\mathbf{w}}, \langle \cdots \rangle_{\mathbf{s}}, \langle \cdots \rangle_{\mathbf{S}}$$

$$\text{data } \langle \cdots \rangle_{\boldsymbol{\xi}}, \langle \cdots \rangle_{\boldsymbol{\xi}^n}, \langle \cdots \rangle_{\boldsymbol{\xi}^N}, \langle \cdots \rangle_{\overline{\boldsymbol{\xi}^n}}, \langle \cdots \rangle_{\overline{\boldsymbol{\xi}^N}}$$

When averaging over the data, the subscript will appear with a bar (i.e. $\overline{\boldsymbol{\xi}^n}$), but when just integrating over the data, no bar will appear (i.e., $\boldsymbol{\xi}^n$). We also reuse these symbols for other quantities, such as the $Z_n^{an,hT}$, $\bar{\mathcal{E}}_{gen}^{an,hT}$, $H^{an}(\mathbf{w})$, etc, but may mix-and-match subscripts and superscripts for visual clarity.

Sign Conventions. Finally, we discuss the sign conventions used. Since errors decrease with better models, Energies $(\Delta E_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}), \mathcal{E}(\mathbf{w}), \epsilon(\mathbf{w}), \cdots)$ and Free Energies (F) are minimized to obtain better models. Likewise, since accuracies increase with better models, Qualities $(\bar{\mathcal{Q}}, \bar{\mathcal{Q}}^2, \cdots)$, Self-Overlap (η), and Quality Generating Function (Γ) would be maximized to obtain better models. An exception will be Hamiltonians $(H, \mathbf{H})$, which will depend on context.

Thermal Averages (over weights). To evaluate the expectation value of some equilibrium quantity that depends on the weights $\mathbf{w}$ (say $\mathbb{E}_{\mathbf{w}}^{\beta}[f(\mathbf{w})]$), one uses a Thermal Average. By this, we mean a *Boltzmann-weighted average*: given a function $f(\mathbf{w})$, we define the Thermal Average over $\mathbf{w}$ as

$$\langle f(\mathbf{w}) \rangle_{\mathbf{w}}^{\beta} \coloneqq \frac{1}{Z_n} \int d\mu(\mathbf{w}) f(\mathbf{w}) e^{-\beta \mathcal{E}(\mathbf{w})}, \tag{34}$$

where the superscript β denotes Thermal Average, $\beta = \frac{1}{T}$ is an inverse temperature, and Z_n is the normalization term (or Partition function), defined as

$$Z_n \coloneqq \int d\mu(\mathbf{w}) e^{-\beta \mathcal{E}(\mathbf{w})}, \tag{35}$$

defined for the n-size *Data Sample* $[\boldsymbol{\xi}^n]$. In particular, when we want to compute the Thermal Average of the *Total Energy* difference or Error $\mathcal{E}(\mathbf{w})$ over $\mathbf{w}$, we could write

$$\langle \mathcal{E}(\mathbf{w}) \rangle_{\mathbf{w}}^{\beta} \coloneqq \frac{1}{Z_n} \int d\mu(\mathbf{w}) \mathcal{E}(\mathbf{w}) e^{-\beta \mathcal{E}(\mathbf{w})}. \tag{36}$$

Importantly, we will never calculate the average errors directly like this. Instead, we will calculate them from partial derivatives of the Free Energy F (as shown below). Also, we may use $\langle \cdots \rangle_{\boldsymbol{\xi}^n}^{\beta}$ to denote what looks like a Thermal Average over the data; this is not essential and only used once below and can be ignored for this section.

Other Notation: Overbars, Superscripts and Subscripts. As above, we may also occasionally denoted averages using the common notation for expected values, $\mathbb{E}[\cdots]$. See Table 11 and 12 in Appendix A for a list of these and other notational conventions and symbols we use.

When discussing quantities such as the Free Energy (F), training and test errors/eneries ($\mathcal{E}$), the Layer Quality ($\mathcal{Q}$), etc., we will place a bar over the symbol (i.e., $\bar{F}$, $\bar{\mathcal{E}}$, $\bar{\mathcal{Q}}$, etc.) when referring to an average over n (or N, below). Otherwise, we will refer to these quantities as the total (averaged) energy, error, quality, etc.

Finally, when referring to the model (i.e., theoretical) training and generalization errors, we will use the superscript ST for the average Student-Teacher training and generalization errors, $\bar{\mathcal{E}}^{ST}_{train}$ and $\bar{\mathcal{E}}^{ST}_{gen}$, respectively, and the superscript NN for the matrix-generalized NN layer average training and generalization errors, $\bar{\mathcal{E}}^{NN}_{train}$ and $\bar{\mathcal{E}}^{NN}_{gen}$, respectively. When referring to empirical errors, we denoted these as $\bar{\mathcal{E}}^{emp}_{train}$ and $\bar{\mathcal{E}}^{emp}_{gen}$, respectively.

4.2.3 Free Energies and Generating Functions

If one needs an average energy (or error), it is often easier to calculate the associated Free Energy and take corresponding partial derivatives than it is to compute that quantity directly via an expected value or Thermal Average. Generally speaking, a Free Energy, F_n, is defined in terms of a partition function Z_n as

$$\beta F_n := -\ln Z_n. \tag{37}$$

Keep in mind that Z may actually be a function of the data $\boldsymbol{\xi}$ (or some other variables), i.e., $Z(\boldsymbol{\xi})$, but we usually don't write this explicitly. Likewise, while both F_n and Z_n depend explicitly on the system size n, we will only include these subscripts when emphasizing this. Also, F has units of Energy or Temperature, so $\beta F = -\ln Z$ is a kind-of unitless Free Energy. Each model (in single-layer models) and/or layer (in multi-layer models) will have its own Partition Function and associated Generating Functions. We call F and Z *Generating Functions* because they can be used to generate the model errors. That is, given an F and/or Z, we can "generate" the training and generalization errors with the appropriate partial derivatives [9, 10].

From this generating function perspective, i.e., when using a generating function to compute quantities of interest, we can work with other transformations of F. Most notably, we will consider

$$\Gamma = n - F. \tag{38}$$

where n is the number of free parameters ($n = N$ for a vector model, but a matrix model has $N \times M$ free parameters; see Appendix A.2.1). The quantity Γ decreases as the error increases (as opposed to F, which increases), i.e., it increases as the accuracy of quality of the model increases. Thus, we will use it as a generating function for the model quality/accuracy. For average Qualities, one has

$$\bar{\Gamma} := 1 - \bar{F} \tag{39}$$

for the model or layer under consideration (see below).

4.2.4 The Annealed Approximation (AA) and the High-Temperature Approximation (high-T)

In the traditional `SMOG` approach, one models the (typical) generalization behavior of a NN by defining and computing the Expected Value of the Free Energy of the model. The full expected value of the Free Energy, $\beta F_n = -\ln Z_n$, with respect to the (model) data $\boldsymbol{\xi}^n$, is:

$$\mathbb{E}_{\boldsymbol{\xi}^n}[\beta F_n] = \beta \bar{F} := -\langle \ln Z_n \rangle_{\overline{\boldsymbol{\xi}^n}}, \tag{40}$$

where $\langle \ln Z_n \rangle_{\overline{\boldsymbol{\xi}^n}}$ means where we average over the n samples of the data ($\boldsymbol{\xi}^n$, of size n). This is also called the Quenched Free Energy. This is, however, frequently too difficult to analyze, and doing so typically requires a so-called Replica calculation.

The Annealed Approximation (AA) is a way of taking the data-average *first* and greatly simplifies the model under study and its analysis. The standard way to move forward is to follow the methods used in disordered systems theory [8, 5]. The mapping is:

Average over the Data	$\leftrightarrow$	Annealed Approximation	$\leftrightarrow$	Disorder Average
Learning the Weights	$\leftrightarrow$	NN Optimization Algorithm	$\leftrightarrow$	Thermal Average.

The Annealed Approximation (AA). Formally, the AA makes the substitution

$$-\langle \ln Z_n \rangle_{\overline{\boldsymbol{\xi}^n}} \approx -\ln \langle Z_n \rangle_{\overline{\boldsymbol{\xi}^n}}. \tag{41}$$

Here, we are *averaging over the disorder*. We may associate:

$$\begin{aligned} -\langle \ln Z_w(\boldsymbol{\xi}^n) \rangle_{\overline{\boldsymbol{\xi}^n}} &: \text{the (unitless) Quenched Free Energy} \\ -\ln \langle Z_n(\boldsymbol{\xi}^n) \rangle_{\overline{\boldsymbol{\xi}^n}} &: \text{the (unitless) Annealed Free Energy.} \end{aligned}$$

Applying the AA amounts to applying Jensens inequality *as an equality*, and it allows will let us interchange integrals and logarithms when computing the data average:

$$\tfrac{1}{n} \int d\mu(\boldsymbol{\xi}^n) \ln(\cdots) \xrightarrow{\text{AA}} \ln \tfrac{1}{n} \int d\mu(\boldsymbol{\xi}^n)(\cdots) \tag{42}$$

This will allow us to switch the order of the data and the thermal averages, i.e.,

$$\left\langle \langle \cdots \rangle_{\mathbf{w}}^{\beta} \right\rangle_{\overline{\boldsymbol{\xi}^n}} \xrightarrow{\text{AA}} \left\langle \langle \cdots \rangle_{\overline{\boldsymbol{\xi}^n}} \right\rangle_{\mathbf{w}}^{\beta}, \tag{43}$$

greatly simplifying the analysis.

The use of the AA is common in `StatMech`, as it simplifies computations considerably; and it is chosen when it holds exactly (if, say, x is a typical sample from $\mathcal{D}$ and $Z_w(\boldsymbol{\xi})$ has a well-defined mean). In contrast, there are situations in `StatMech` when the average is atypical, and then it one can get different results for the Quenched versus Annealed cases. In a practical sense, one imagines this may occur when the data is very noisy and/or mislabeled, and this requires a special treatment [8].

Annealed Hamiltonian $H^{an}(\mathbf{w})$ and Annealed Parition Function Z^{an}. When we apply the AA (as in Eqn. 42), we average over the data $\boldsymbol{\xi}^n$ first. Doing this will allow us to develop a theory in terms a (Temperature dependent) Annealed Error Potential.

Following [8] (see their Eqn.(2.30)), we will call this average the *Annealed Hamiltonian*, $H^{an}(\mathbf{w})$, which we define as

$$\beta H^{an}(\mathbf{w}) := -\frac{1}{n} \ln \int d\mu(\boldsymbol{\xi}^n) e^{-\beta \boldsymbol{\Delta \mathbf{E}}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)}. \tag{44}$$

The Annealed Hamiltonian is a simple "mean-field-like" Hamiltonian for the problem. This can be

seen by noting that we can express Eqn. 44 as an integral over a single data example $\boldsymbol{\xi}$:

$$\begin{aligned}\beta H^{an}(\mathbf{w}) &= -\frac{1}{n}\ln\int d\mu(\boldsymbol{\xi}^n)e^{-\beta\sum_{\mu=1}^{n}\Delta E_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}_\mu)} \\ &= -\ln\left[\int d\mu(\boldsymbol{\xi}^n)e^{-\beta\sum_{\mu=1}^{n}\Delta E_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}_\mu)}\right]^{\frac{1}{n}} \\ &= -\ln\left[\prod_{\mu=1}^{n}\int d\mu(\boldsymbol{\xi}_\mu)e^{-\beta\Delta E_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}_\mu)}\right]^{\frac{1}{n}} \\ &= -\ln\int d\mu(\boldsymbol{\xi})e^{-\beta\Delta E(\mathbf{w},\boldsymbol{\xi})} \end{aligned} \tag{45}$$

This will be a critical piece needed to generalize the vector-based ST Perceptron model to the matrix-generalized ST MLP model. In BraKet notation, Eqn. 45 can be expressed as

$$\beta H^{an}(\mathbf{w}) := -\tfrac{1}{n}\ln\langle e^{-\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)}\rangle_{\boldsymbol{\xi}^n} = -\ln\langle e^{-\beta\Delta E(\mathbf{w},\boldsymbol{\xi})}\rangle_{\boldsymbol{\xi}}.$$

Using $H^{an}(\mathbf{w})$, we can define the *Annealed Partition Function*, Z_n^{an}, as

$$Z_n^{an} := \int d\mu(\mathbf{w})e^{-n\beta H^{an}(\mathbf{w})} \tag{46}$$

$$\begin{aligned} &= \int d\mu(\mathbf{w})\exp\left[-n\left(-\frac{1}{n}\ln\int d\mu(\boldsymbol{\xi}^n)e^{-\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)}\right)\right] \\ &= \int d\mu(\mathbf{w})\exp\left[\ln\int d\mu(\boldsymbol{\xi}^n)e^{-\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)}\right] \\ &= \int d\mu(\mathbf{w})\int d\mu(\boldsymbol{\xi}^n)e^{-\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)}. \end{aligned} \tag{47}$$

where the lines after the first line follow by substituting Eqn. 44 into Eqn. 46. Note also that the order of the integrals in Eqn. 47 is exactly what we expect using the AA, as in Eqn. 43. Also, analogously to Eqn. 45, we can write Z_n^{an} as the product of the $n=1$ case, $Z_n^{an} = [Z_1^{an}]^n$. Finally, we will only need the high-T version, H_{hT}^{an}, of $H^{an}(\mathbf{w})$, and this will take a very simple form. [24]

The High-Temperature (High-T) Annealed Hamiltonian ($H_{hT}^{an}(\mathbf{w}) = \epsilon(\mathbf{w})$) and Partition Function ($Z_n^{an,hT}$). In addition to the AA, we will be evaluating our models at at high-T. Notably, the Annealed Hamiltonian $H^{an}(\mathbf{w})$ in Eqn. 44 is a non-linear function of β; by taking the high-T approximation, we can remove this dependence and obtain the simpler expression that $H^{an}(\mathbf{w}) = \epsilon(\mathbf{w})$. This greatly simplifes both the Partition Function, i.e., $Z_n^{an,hT}$, and subsequent results (below).

To obtain the high-T result, we can write the Taylor expansion for $e^{\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)}$ and keep the first two terms:

$$e^{-n\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)} \simeq \underbrace{1-\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)}_{high-T} + (\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n))^2 + \cdots. \tag{48}$$

Let us now express $H^{an}(\mathbf{w})$ directly in terms of $\epsilon(\mathbf{w}) = \langle\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^n}}$ (see Eqn. 30) as a Thermal Average at high-T. charlesBbecause the 1/n appears outside the ln, not in the exp To do

[24] We will derive expressions for $\epsilon(R)$ and $\bar{\mathcal{E}}_{gen}^{ST}(R)$ in Eqn. 94 and Eqn. 95, respectively, using relatively simple arguments. In Appendix A.2.1 we present a more detailed derivation of $H^{an}(R)$ and $H_{hT}^{an}(R) = \epsilon(R)$ in Appendix A.2.2 we show that this derivations generalizes to the matrix-generalized case, $H^{an}(\mathbf{R})$. This more detailed derivation is important for our `SETOL` setup because it lets us define the normalization necessary for the `TRACE-LOG` condition.

so, let's take a high-T expansion of Eqn. 44 by expanding the exponential to first order in β, to obtain

$$\begin{aligned}\beta H^{an}_{hT}(\mathbf{w}) &= -\tfrac{1}{n}\ln\int d\mu(\boldsymbol{\xi}^n)[1-\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)] \qquad (49)\\ &\approx -\frac{1}{n}\int d\mu(\boldsymbol{\xi}^n)\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)\\ &= \langle\beta\boldsymbol{\Delta}\mathbf{E}_{\mathcal{L}}(\mathbf{w},\boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^n}}\\ &= \beta(n\epsilon(\mathbf{w}))\\ &= \beta\epsilon(\mathbf{w}). \qquad (50)\end{aligned}$$

Here, we have used the AA, the property that $\ln(1+y)\approx y$, for $|y|\ll 1$, and the fact that $\epsilon(\mathbf{w})$ takes the form given in Eqn. 30. This gives $H^{an}_{hT}(\mathbf{w}) := \epsilon(\mathbf{w})$, which is no longer a non-linear function of β. Moving forward, we will assume we are taking the high-T limit like this.

Given Eqn. 50, we can now express the Annealed Partition Function at high-T directly in terms of the Annelaed (i.e., data-averaged) error $\epsilon(\mathbf{w})$:

$$\begin{aligned}Z^{an,hT}_n &:= \int d\mu(\mathbf{w})e^{-n\beta H^{an}_{hT}(\mathbf{w})}\\ &= \int d\mu(\mathbf{w})e^{-n\beta\epsilon(\mathbf{w})} \qquad (51)\end{aligned}$$

If we assume that at high-T we can make this approximation, since we only care about the small β results. This will prove very useful later when working with HCIZ integrals.

4.2.5 Average Training and Generalization Errors and their Generating Functions

Here, we show how to derive the Average Training Error $\bar{\mathcal{E}}_{train}$ and the Average Generalization Error $\bar{\mathcal{E}}_{gen}$, in the Annealed Approximation (AA), and at high-Temperature (high-T), using the Free Energy F and/or the Partition Function Z as a generating function. In particular, we show that, in the AA and at high-T, these errors are (approximately) equal and equal to the Thermal Average of the Annealed Error Potential, $\bar{\mathcal{E}}^{an,hT}_{train} \approx \bar{\mathcal{E}}^{an,hT}_{gen} \approx \langle\epsilon(\mathbf{w})\rangle^{\beta}_{\mathbf{w}}$.

Generating Functions for the Errors: the `StatMech` way. In our theory, after applying the AA, we obtain expressions where the random model data $\boldsymbol{\xi}^n$ has been integrated out. This leaves formal quantities that depend only on the weights $\mathbf{w}$, which are the variables being learned during training. Since we are left with a distribution over $\mathbf{w}$, we define the training error not explicitly as an average over the training data itself, but instead in terms of how the Free Energy, $\beta F := -\ln Z_n$, varies with β, i.e., the amount of stochasticity in the model weights.

Following [9, 10], we define the *Average Training Error*, in the AA, by by differentiating $\ln Z^{an}_n$ with respect to β:

$$\bar{\mathcal{E}}^{an}_{train} := -\frac{1}{n}\frac{\partial(\ln Z^{an}_n)}{\partial\beta} = -\frac{1}{n}\frac{1}{Z_n}\frac{\partial Z^{an}_n}{\partial\beta}. \qquad (52)$$

This error captures how the model predictions will vary with changes in the learned weights $\mathbf{w}$, which implicitly describes how the changes will vary with the training data $\boldsymbol{\xi}^n$. Similarly, we define the *Average Generalization Error*, in the AA, by differentiating $\ln Z^{an}_n$ with respect to n, the number of data points. Following

$$\bar{\mathcal{E}}^{an}_{gen} := -\frac{1}{\beta}\frac{\partial(\ln Z^{an}_n)}{\partial n} + \frac{1}{\beta}\ln z(\beta) = -\frac{1}{\beta}\frac{1}{Z_n}\frac{\partial Z^{an}_n}{\partial n} + \frac{1}{\beta}\ln z(\beta), \qquad (53)$$

where $z(\beta)$ is a constant normalization term that depends only on β (which, moving forward, we ignore, as it only shifts the scale). This error captures how the model's predictions will change as more data points are introduced.

In the Thermodynamic limit ($n \gg 1$), these two definitions of the error become equivalent at High-T, and they equal to the Thermal Average of the Annealed Error Potential:

$$\bar{\mathcal{E}}^{an,hT}_{train} = \bar{\mathcal{E}}^{an,hT}_{gen} = \langle \epsilon(\mathbf{w}) \rangle^{\beta}_{\mathbf{w}} \ , n \gg 1. \tag{54}$$

To see this, substitute Eqn. 51 into Eqn. 52, and take the partial derivative w.r.t β, to obtain

$$\begin{aligned} \bar{\mathcal{E}}^{an,hT}_{gen} :=& \frac{1}{n} \frac{\partial(-\ln Z^{an,hT}_n)}{\partial \beta} \\ =& -\frac{1}{n} \frac{\partial}{\partial \beta} \ln \int d\mu(\mathbf{w}) e^{-\beta n \epsilon(\mathbf{w})} \\ =& \frac{\frac{1}{n} \int d\mu(\mathbf{w}) n\epsilon(\mathbf{w}) e^{-\beta n \epsilon(\mathbf{w})}}{\int d\mu(\mathbf{w}) e^{-\beta n \epsilon(\mathbf{w})}} \\ =& \langle \epsilon(\mathbf{w}) \rangle^{\beta}_{\mathbf{w}}. \end{aligned} \tag{55}$$

Likeswise, if we substitute Eqn. 51 into Eqn. 53, and take the partial derivative w.r.t. n, we obtain

$$\begin{aligned} \bar{\mathcal{E}}^{an,hT}_{gen} :=& \frac{1}{\beta} \frac{\partial(-\ln Z^{an,hT}_n)}{\partial n} \\ =& -\frac{1}{n} \frac{\partial}{\partial \beta} \ln \int d\mu(\mathbf{w}) e^{-\beta n \epsilon(\mathbf{w})} \\ =& \frac{\frac{1}{\beta} \int d\mu(\mathbf{w}) n\epsilon(\mathbf{w}) e^{-\beta n \epsilon(\mathbf{w})}}{\int d\mu(\mathbf{w}) e^{-\beta n \epsilon(\mathbf{w})}} \\ =& \langle \epsilon(\mathbf{w}) \rangle^{\beta}_{\mathbf{w}}. \end{aligned} \tag{56}$$

Notice that both of these results arise because of the simple expression that appears in the exponent of $Z^{an,hT}_n$, namely because $-n\beta H^{an}_{hT}(\mathbf{w}) = n\beta\epsilon(\mathbf{w})$.

This equivalence reflects the fact that when the system is large enough, adding a new data example to the training distribution is formally equivalent to adding noise, making the two errors indistinguishable. This approach allows us to define both training and generalization errors in terms of fundamental thermodynamic quantities, providing a simplified formal framework suitable for empirical adjustment later.

Also, note that the model data variables $\boldsymbol{\xi}$ do not enter the calculation because we integrated them out before the calculation of Thermal Average. (This illustrates the difference between taking an annealed versus a quenched average.) Also, our sign convention is consistent with a model of NN training that *minimizes* the total loss ($\mathcal{L}$) and/or ST error, and, therefore minimizes Free Energies as well.

More generally, we see that we can use the Partition Function, Z, and/or the Free Energy, $\beta F := -\ln Z$, as a Generating Function to obtain any desired Thermodymanic average by taking the appropriate partial derivative of the correpsonding form of $\ln Z$. In Appendix **??** we show how to obtain $\bar{\mathcal{E}}^{an}_{train}$ and $\bar{\mathcal{E}}^{an}_{gen}$ obtain explicitly in this way using Z^{an}.

4.2.6 The Quality ($\bar{\mathcal{Q}}$) and its Generating Function ($\Gamma_{\bar{\mathcal{Q}}}$)

Here, we explain how to define what we call the Quality $\bar{\mathcal{Q}}$, which is defined as the Thermal Average of the Self-Overlap, $\bar{\mathcal{Q}} := \langle \eta(\mathbf{w}) \rangle^{\beta}_{\mathbf{w}}$, and which can be obtained from the associated *Quality Generating Function* $\Gamma_{\bar{\mathcal{Q}}}$. See Table 5 for a quick summary.

Symbol	Definition & Formula		Eq. #
ϵ	Annealed Error Potential	$\epsilon = \lim_{n\to\infty} \frac{1}{n}\langle \Delta E \rangle_{\text{data}}$	30
η	Self-Overlap (average accuracy):	$\eta = 1 - \epsilon$	33
$\bar{F}$	Average Free Energy:	$\beta \bar{F} = -\frac{1}{n} \ln Z_n$	40
$\bar{Q}$	Model Quality (average self-overlap):	$\bar{Q} = \langle \eta \rangle_{\mathbf{w}}^{\beta}$	59
$\Gamma_{\bar{Q}}$	Quality-Generating Function:	$\Gamma_{\bar{Q}} = 1 - \bar{F}$	39

Table 5: Summary of the main intensive (average, per-parameter) quantities. Here n is the number of free parameters. The average model Quality $\bar{Q}$ is the model's average accuracy (one minus the error), and the Quality-Generating function $\Gamma_{\bar{Q}}$ plays the same role as the free energy $\bar{F}$ but with an opposite sign convention.

For our purposes below, we define the Model Quality (as in Eqn. 7) as our approximation to the Average Generalization Accuracy for our model. We denote the Model Quality for the ST Perceptron model, $\bar{\mathcal{Q}}^{ST}$, and for a general NN, $\bar{\mathcal{Q}}^{ST}$, such that

$$\bar{\mathcal{Q}}^{ST} := 1 - \bar{\mathcal{E}}_{gen}^{ST} \tag{57}$$

$$\bar{\mathcal{Q}}^{NN} := 1 - \bar{\mathcal{E}}_{gen}^{NN} \tag{58}$$

In this work, however, the Quality will always be defined at high-T, and so we may write

$$\bar{\mathcal{Q}}^{ST} = 1 - [\bar{\mathcal{E}}_{train}^{ST}]^{an,hT} = 1 - [\bar{\mathcal{E}}_{gen}^{ST}]^{an,hT} \tag{59}$$

$$\bar{\mathcal{Q}}^{NN} = 1 - [\bar{\mathcal{E}}_{train}^{NN}]^{an,hT} = 1 - [\bar{\mathcal{E}}_{gen}^{NN}]^{an,hT} \tag{60}$$

We also define a Layer Quality, simply denoted $\bar{\mathcal{Q}}$, which will describe the contributions an individual layer makes to the overall Model Quality $\bar{\mathcal{Q}}^{NN}$. To obtain the Layer Quality, we define an accuracy–or Quality–Generating Function, denoted $\Gamma_{\bar{\mathcal{Q}}}$, which is analogous to a layer Free Energy, but with the opposite sign convention.

Generally speaking, the Quality Generating Function $\Gamma_{\bar{\mathcal{Q}}}$ is defined in the AA, and at high-T and is given as

$$\beta \Gamma_{\bar{\mathcal{Q}}} := \ln \int d\mu(\mathbf{w}) e^{n\beta\eta(\mathbf{w})} \tag{61}$$

For example, for the single-layer ST Perceptron, $\Gamma_{\bar{\mathcal{Q}}^{ST}} := n - F^{ST}$ (where n here is also the number of free parameters for this model, and is in units of energy or error). The term $\Gamma_{\bar{\mathcal{Q}}^{ST}}$ is directly related to the *Total* Free Energy F^{ST}, which can be seen by substituting Eqn. 33 for $\eta(\mathbf{w})$ in Eqn. 61:

$$\begin{aligned}
\beta \Gamma_{\bar{\mathcal{Q}}^{ST}} &= \ln \int d\mu(\mathbf{w}) e^{n\beta(1-\epsilon(\mathbf{w}))} \\
&= \ln \int d\mu(\mathbf{w}) e^{n\beta} e^{-n\beta\epsilon(\mathbf{w})} \\
&= \ln \left(\int d\mu(\mathbf{w}) e^{n\beta} \right) + \ln \left(\int d\mu(\mathbf{w}) e^{-n\beta\epsilon(\mathbf{w})} \right) \\
&= \ln e^{n\beta} + \ln \left(\int d\mu(\mathbf{w}) e^{-n\beta\epsilon(\mathbf{w})} \right) \\
&= \beta n - \beta F^{ST}
\end{aligned} \tag{62}$$

Dividing by n, we can also recover the more general relation for the *Average* Free Energy, $\bar{\Gamma}_{\bar{\mathcal{Q}}} = 1 - \bar{F}$. (Eqn. 39). For the matrix case we do something similar; see Appendix A.3.

Likewise, one can show that the Quality $\bar{\mathcal{Q}}$ (again, always in the AA and at high-T) can be identified as the Thermal Average of the (data-averaged or Annealed) Self-Overlap,

$$\bar{\mathcal{Q}} = \left\langle \langle \eta(\mathbf{w}, \boldsymbol{\xi}) \rangle_{\overline{\boldsymbol{\xi}^n}} \right\rangle_{\mathbf{w}}^{\beta} = \langle \eta(\mathbf{w}) \rangle_{\mathbf{w}}^{\beta} \tag{63}$$

We can then obtain $\bar{\mathcal{Q}}$ by taking the appropriate partial derivative of its Generating Function, Γ_Q.

For technical reasons, however, we will actually define and use a Generating Function for the Average Layer Quality-Squared $\bar{\mathcal{Q}}^2$, denoted $\beta \mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$. In analogy with Eqns. 52 and 53, and at high-T and large-N (explained below), we can obtain $\bar{\mathcal{Q}}^2$ (see Section 5 Eqn. 125) as

$$\bar{\mathcal{Q}}^2 \coloneqq \lim_{N \gg 1} \frac{1}{\beta} \frac{\partial}{\partial N} \beta \mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} \underset{\text{high-}T}{\approx} \lim_{N \gg 1} \frac{1}{N} \frac{\partial}{\partial \beta} \beta \mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} \tag{64}$$

See Section 5 and Appendix A.3 for more details.

4.2.7 The Thermodynamic Large-N limit and the Saddle Point Approximation (SPA)

To evaluate $\bar{\mathcal{Q}}^2$, we take a Large-N limit, which is in the Thermodynamic limit, which we assume going forward.

We note that, technically, the Thermodynamic limit and the Large-N limit are different. In particular, the Thermodynamic limit typically refers to the case where all Thermodynamic averages remain Size-Intensive as the system size increases; and, while this does refer to $n = N$ growing large, frequently for NNs there is an additional constraint that the ratio $(m/n = M/N)$, i.e., the load, remains constant [8, 91]. For the ST model of the Perceptron, however, the Thermodynamic averages we need to compute not depend on m, so this constraint is not necessary. Moreover, later, when we form our matrix generalization of the ST model, we will not enforce that N/M remain constant but, instead, the final result will be Size-Consistent.

Here, the Thermodynamic limit is the large-N limit of the model, as $n = N$ grows arbitrarily large, i.e.m $n \gg 1$. To express the average Free Energy $\bar{F}$ in the large-N limit, we can write

$$-\beta \bar{F} = \lim_{n \gg 1} \frac{1}{n} \ln \int d\mu(\mathbf{w}) e^{-n\beta\epsilon(\mathbf{w})}. \tag{65}$$

When this large-N approximation is well behaved, then the total energy $\mathcal{E}(\mathbf{w})$ is extensive, i.e., when $\mathcal{E}(\mathbf{w}) = n\epsilon(\mathbf{w})$; and, consequently, the total Free Energy is also extensive, i.e., $F = n\bar{F}$. (This is a cornerstone of statistical mechanics, as it allows for meaningful macroscopic predictions from microscopic interactions.)

Self-Averaging. The existence of the limit signifies that the system is *Self-Averaging*, meaning that the macroscopic properties are independent of fluctuations, etc. This also implies that the relevant averages (i.e., training and generalization errors) are the same for almost all samples, or *realizations of the disorder*. Additionally, the Annealed and the Quenched averages, $\ln\langle Z_n \rangle_{\overline{\boldsymbol{\xi}^n}}$ and $\langle \ln Z_n \rangle_{\overline{\boldsymbol{\xi}^n}}$, respectively, become sharply peaked, and

$$\langle \ln Z_n \rangle_{\overline{\boldsymbol{\xi}^n}} \approx \ln \langle Z_n \rangle_{\overline{\boldsymbol{\xi}^n}}, \quad \text{as } n \to \infty. \tag{66}$$

For a NN, Self-Averaging implies that the weights $\mathbf{w}$ are *typical* of the distribution, and therefore the NN can generalize to similar but unseen test examples.

Saddle Point Approximation (SPA). When the Thermodynamic limit exists, we can approximate the asymptotic behavior of integrals we will encounter (e.g., the Free Energy F and/or Partition Function Z) in the large-N limit by using the Saddle Point Approximation (SPA). In the case of the F and/or Z, we assume that we can apply the SPA:

$$\int d\mu(\mathbf{w}) e^{-n\beta\epsilon(\mathbf{w})} \approx \sqrt{\frac{2\pi}{n\epsilon(\mathbf{w}^*)}} e^{-n\beta\epsilon(\mathbf{w}^*)}, \tag{67}$$

where $\mathbf{w}^*$ satisfies the saddle point equations:

$$\epsilon(\mathbf{w}^*) := \frac{\partial}{\partial \mathbf{w}} \epsilon(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^*} = 0 \tag{68}$$

$$\epsilon(\mathbf{w}^*) := \frac{\partial^2}{\partial^2 \mathbf{w}} \epsilon(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^*} > 0. \tag{69}$$

To apply the SPA rigorously, we expect that $\epsilon(\mathbf{w})$ decays exponentially, whereas we will be working with Heavy-Tailed (HT)/Power-Law (PL) distributions. At first glance, this might seem problematic. These distributions, however, can also be modelled as Truncated Power Laws (TPL) because of the finite-size effects. This suggests we can apply the SPA formally, assuming (but not checking) that the finite-size effects will satisfy the requirements we need. The SPA will play a central role when we evaluate the matrix-generalized form of our model and the HCIZ integrals that arise.

When the limit fails: atypical behavior. When the Thermodynamic limit fails to exist, the Free Energy will contain additional, non-extensive terms, i.e., $F = n\bar{F}_{ex} + n^{1+x}\bar{F}_{non-ex}$, $x > 0$. [25] In this case, the AA may fail, the SPA may not apply, and the system may fail to be Self-Averaging. This causes the system to behave in an atypical way, possibly converging to a meta-stable and/or glassy phase. Indeed, when the weights $\mathbf{w}$ are *atypical*, they may describe the training data well, but they would fail to describe the test data well; in this sense, we say the model is *overfit* to the training data. We will not explicitly consider a model that is non-extensive; however, we will present empirical results where we suspect the model is overfit (in Section 6.5) and, additionally, where we observe glassy behavior, which we refer to as a *Hysteresis Effect* (in Section 6.6).

Remark. Additional results are provided in Appendix A.2. In particular, in Appendix A.2.1, we use the ideas from this section to derive the full non-linear vector form of the Annealed Hamiltonian $H^{an}(\mathbf{w})$ (Eqn. 44) for the Linear Perceptron, in the AA, and then we express it at High-T, i.e., $H^{an}_{hT}(R) = \epsilon(R)$ (Eqn. 49). Then, in Appendix A.2.1, we derive the matrix generalization $H^{an}(R) \to H^{an}(\mathbf{R})$ of these quantities. This derivation is necessary to obtain the normalization on the weight matrices $\mathbf{W}$ necessary for the final results in Section 5.

4.2.8 HCIZ Integrals

To generalize the Linear ST Perceptron (in the AA, and at high-T) from Perceptron vectors to MLP matrices, we need to generalize the thermal average over the m-dimensional Perceptron weight vectors ($\mathbf{w} = \mathbf{s}$) to an integral over NN Student $N \times M$ weight matrices ($\mathbf{W} = \mathbf{S}$). The resulting Partition Function and Free Energy will be expressed with what is called an HCIZ integral. (See Tanaka [82, 83], Gallucio et al. [40] and/or Parisi [92], and also Eqn. 127 in Section 5.) Analogously to Eqn. **??**, we will define a *Layer Quality Generating Function*, $\beta\mathbf{\Gamma}^{IZ}_{\mathcal{Q}^2}$, for the Layer Quality

[25] When dealing with matrix integrals (below), $F \sim NMF_0 + \cdots$, when there are $N \times M$ degrees of freedom [92], but we will only be concerned with the large-N limit.

(squared). This will take the form of an HCIZ integral,

$$\beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2} \coloneqq \lim_{N\gg 1} \ln \underbrace{\int d\mu(\mathbf{A}) \exp[N\beta \operatorname{Tr}[\mathbf{A}_2\mathbf{X}_2]]}_{\text{HCIZ Integral}} \approx N\beta \operatorname{Tr}[\mathcal{G}_A(\lambda)], \tag{70}$$

that is evaluated in the "large-N limit" (here, meaning as $N \gg 1$, but not at the limit $N = \infty$). Here, $\mathbf{A}$ and $\mathbf{X}$ are $N \times N$ Hermetian matrices, $d\mu(\mathbf{A})$ is a measure over all random (Orthogonal) matrices (see Eqn. 252), and $\mathcal{G}_A()$ is a Norm Generating Function (different from Γ, above), defined in Eqn. 130 in Section 5. In applying this, we will actually write $\operatorname{Tr}[\mathbf{AX}] = \frac{1}{N}\operatorname{Tr}[\mathbf{AWW}^\top] = \frac{1}{N}\operatorname{Tr}[\mathbf{W}^\top\mathbf{AW}]$, where $\mathbf{W}$ is an $N \times M$ layer weight matrix, and $\mathbf{A} = \mathbf{A}_2 \coloneqq \frac{1}{N}\mathbf{SS}^\top$ is a layer Correlation matrix, and, here, $\mathbf{X} = \mathbf{X}_2 = \frac{1}{N}\mathbf{WW}^\top$). Moreover, to evaluate this, we will need to restrict $\mathbf{A}$ (and $\mathbf{X}$) to the lower-rank Effective Correlation Space (ECS), i.e., $\mathbf{A} \to \tilde{\mathbf{A}}$.

Notice this looks similar to Saddle Point Approximation (SPA), but where the more complicated function $\mathbb{G}_{\mathbf{A}}()$ now appears. Also, $\mathbb{G}_{\mathbf{A}}()$ here depends on only the limiting form of the ESD of $\mathbf{A}$, $\rho^{\infty}_{A}(\lambda)$, and depends on the M normalized eigenvalues λ_i of $\mathbf{X}$.

To evaluate this, we will form the large-N limit, using a result from Tanaka [82, 83], but extended slightly (in Appendix A.6) to include the inverse-Temperature β explicitly. We can write

$$\beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2,N\gg 1} \coloneqq \lim_{N\gg 1} \beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2}, \tag{71}$$

with the final result

$$\beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2,N\gg 1} \coloneqq N\beta \sum_{i=1}^{\tilde{M}} \int_{\lambda^{ECS}_{min}}^{\tilde{\lambda}_i} dz R(z), \tag{72}$$

where $R(z)$ is a complex function, the R-transform of the ESD $\rho(\lambda)$ of the Teacher, and $\tilde{\lambda}_i$ are the eigenvalues of Teacher Correlation Matrix $\tilde{\mathbf{X}}$, restricted to the ECS. For more details, see Section 5 and Appendicies A.4 and A.6.

Branch Cuts and Phase Behavior. Free energies (F, Γ) often exhibit *branch cuts* when expressed as analytic functions of complex parameters (i.e., temperature, coupling constants, or eigenvalue cutoffs), and arise from singularities in the underlying integral representations of the partition function Z, When a branch cut occurs, this demarcates non-analytic behavior and this indicates the onset of a *phase transition* where macroscopic observables such as correlation lengths and/or variance-like quantities may diverge or change character abruptly.

In the context of our HCIZ-based construction, integrating a complex function like the $R(z)$-transform of a Heavy-Tailed ESD can produce precisely this phenomenon. For example, if $R(z)$ has a square-root term, i.e., $(\sqrt{z-c})$, then it will have a branch cut at $z = c$, and the Generating Function (i.e., Effective Free Energy) $\beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2,N\gg 1}$ will be non-analytic and we must choose the appropriate, physically meaningful branch, i.e., $(z > c)$, corresponding to the ECS. We argue that this cut signifies a *phase boundary*—an abrupt change in the system's correlation structure and corresponds to an emerging singularity in the Layer Quality. Even though we perform only a single exact RG step , (rather than fully iterating a renormalization flow), the appearance of a branch cut in $\beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2,N\gg 1}$ wil encode non-trivial *phase-like* behavior in the SETOL Heavy-Tailed matrix model.

4.3 Student -Teacher Model

In this subsection, we present a unified view of the *Student-Teacher* (ST) model from both a practical and a theoretical perspective. From the practical (*operational*) side, imagine a real-world

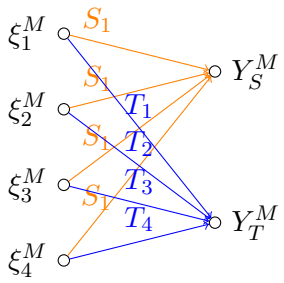


Figure 7: Pictorial representation Student and Teacher Perceptrons.

practitioner training a large-scale neural network (NN); call this NN the Teacher. We wish to assess the Teacher's *true* accuracy on ground-truth labels in consistently reproducing its own (possibly imperfect) outputs, one can train another network –the Student–to mimic the Teacher's predictions. By comparing Student and Teacher outputs, one can approximate the Teacher's generalization performance even without an explicit hold-out set. Notably, if the Teacher perfectly interpolates its training data, the Student's error directly estimates the Teacher's *true* Generalization Accuracy. Otherwise, it captures the Teacher's *Precision* in reproducing its own noisy or biased labels.

From the theoretical side, we seek a succinct, analytic formulation of the ST Average Generalization Error, denoted $\bar{\mathcal{E}}_{gen}^{ST}$. We work in the Annealed Approximation (Å)–a simplification often valid when networks are sufficiently large and can nearly interpolate their training data. Under this approximation, one obtains closed-form expressions for the Student-Teacher *Overlap* (R) and thus for the Teacher's overall error or accuracy. These results lay the foundation for our *Semi-Empirical* approach, in which we supplement this theoretical form with empirical measurements (e.g., from the trained weight matrices) to account for real-world correlations in the data and the model's internal structure.

4.3.1 **Operational Setup** In subsection 4.3.1 we explain how to set up the classic Student-Teacher model in an *operational* manner. In particular emphasize the difference between *true accuracy* (vs. ground-truth labels) and Precision (vs. the Teacher's own labels). We also the discuss the difference between our Quality metrics and the *GeneralizationGap*.

4.3.2 **Theoretical Student-Teacher Average Generalization Error** ($\bar{\mathcal{E}}_{gen}^{ST}$) In subsection 4.3.2 we outline how to derive $\bar{\mathcal{E}}_{gen}^{ST}$ using the AA. We introduce the key expressions that will serve as the starting point for our extended Semi-Empirical theory in subsequent subsubsection.

4.3.1 Operational Setup

Here, describe the basic setup of the classic Student-Teacher model, taking an operational view from the perspective of a practitioner training real-world Student and Teacher models. Specifically, we present the Annealed Approximation (AA) in a practical light, and use it explain the difference between computing the *EmpiricalGeneralization Error*, $\bar{\mathcal{E}}_{gen}^{emp}$, for the *True Accuracy* and for the *Precision* of a Teacher model. Table **??** summarizes some of the key notation we will be using in this section.

Test Error of the Teacher We start by describing how to obtain a simple formal expression for the empirical test errors of the Teacher, first for the True Accuracy.

Symbol	Meaning
T	Teacher model (fixed network)
S	Student model (trained to mimic T)
$\mathcal{D}$	True dataset of (x, y) pairs
$\widetilde{\mathcal{D}}$	Gaussian-field idealization of $\mathcal{D}$
ε_T	True generalization error of T (unknown)
$\widehat{\varepsilon}$	Empirical test-set error estimate
ε_{ST}	Student–Teacher average gen. error
R	Overlap $R = \frac{1}{N} S^\top T$
$\epsilon(R)$	Annealed potential (e.g. $1 - R$ for linear, high-T)

Table 6: Key notation for the Student–Teacher operational setup.

Let us say we have a model, called Teacher (T), which maps some *actual* (i.e., correlated) data $\mathbf{x}_\mu \in \mathbf{D}$ to some known or *true* labels (y_μ^{true}) (where, y_μ^{true} is, say, an N-dimensional vector of binary labels). We might say that y_μ^{true} represents the *Ground Truth* for the problem. Operationally, we train the Teacher T to reproduce or at least approximate the true labels y_μ^{true}.

$$T : \mathbf{x}_\mu \to \mathrm{y}_\mu^T \approx \mathrm{y}_\mu^{true}. \tag{73}$$

If T reproduces the true labels exactly, we might say that the Teacher has been trained to *Interpolation*, and, therefore, $\mathrm{y}_\mu^T = \mathrm{y}_\mu^{true}$. Indeed, most models today are trained to *Interpolation*, and we don't need to necessarily worry about the difference between the true and the predicted Teacher labels. Formally, however, and to understand better the Å, it is beneficial to discuss the implication of this distinction.

Following Eqn. 1, lets say the Teacher outputs are specified by an Energy function E_{NN}^T

$$\mathrm{y}_\mu^T = E_{NN}(T, \mathbf{x}_\mu) \tag{74}$$

[26] so that we may write the *Empirical* Average Training Error $\bar{\mathcal{E}}_{train}^{emp}$ as

$$\bar{\mathcal{E}}_{train}^{emp} := \frac{1}{N^{train}} \sum_{\mu=1}^{N^{train}} \mathcal{L}[\mathrm{y}_\mu^{train}, E_{NN}(T, \mathbf{x}_\mu^{train})]. \tag{75}$$

Ideally, we seek the *True* Average Generalization Error of the Teacher, denoted $\bar{\mathcal{E}}_{gen}^T$. Of course, this is unknowable, but in practice, we estimate $\bar{\mathcal{E}}_{gen}^T$ by measuring the error of the Teacher predictions on some test (or hold-out) set $(\mathbf{x}_\mu^{test}, \mathrm{y}^{test})$. We call this the *Empirical Average Generalization Error*, $\bar{\mathcal{E}}_{gen}^{emp}$, and write

$$\bar{\mathcal{E}}_{gen}^T \approx \bar{\mathcal{E}}_{gen}^{emp} := \frac{1}{N^{test}} \sum_{\mu=1}^{N^{test}} \mathcal{L}[\mathrm{y}_\mu^{test}, E_{NN}(T, \mathbf{x}_\mu^{test})]. \tag{76}$$

To measure the error, the loss function $\mathcal{L}$ may be a L1 (ℓ_1) or L2 (ℓ_2) loss; whereas for training a model, it is usually something like a cross-entropy loss, but this detail does matter later.

If we don't have a hold-out set, however, we can still estimate $\bar{\mathcal{E}}_{gen}^T$ using the Student-Teacher approach.

[26] Do not confuse the Energy/output function E_{NN} with the energies $\mathbf{\Delta E}$ defined below to represent the ST error function(s). We refer to outputs of $E_{NN}(\mathbf{t}, \mathbf{x}_\mu)$, when applied to a data point $\mathbf{x}_\mu$, as energies because they are effectively un-normalized probabilities for the class outputs (for labels $\mathrm{y}_\mu = 1$ or -1).

Estimating the Student-Teacher Error: Accuracy vs. Precision Imagine we train a Student model S (with the same architecture as the Teacher T) on the dataset $\mathcal{D}$, using T's outputs as targets:

$$S : \mathcal{D} \to y_\mu^S \approx y_\mu^T,$$

where $y_\mu^T = E_{NN}(T, x_\mu)$ are the teacher's predicted labels or energies.

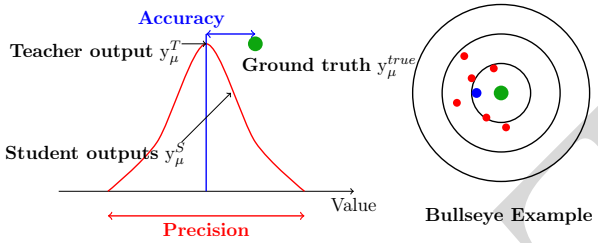


Figure 8: Precision vs. Accuracy

Figure 8 illustrates the following two scenarios side by side:

- **Accuracy (Interpolation) regime:** If T perfectly interpolates the true labels ($y_\mu^T = y_\mu^{\text{true}}$), then $\|y_\mu^S - y_\mu^T\| = \|y_\mu^S - y_\mu^{\text{true}}\|$ measures how well the student recovers the Ground Truth labels—i.e., the Student's *training accuracy*.
- **Precision (Noisy teacher) regime:** If T itself makes mistakes ($y_\mu^T \neq y_\mu^{\text{true}}$), then $\|y_\mu^S - y_\mu^T\|$ estimates how faithfully S reproduces T (its precision), rather than true accuracy.

By analyzing the Student–Teacher overlap $R = \frac{1}{N}\sum_\mu s^T t$, one can show that the average generalization error of T depends simply on $1 - R$ under the annealed approximation—even when T was trained on noisy labels. In practice, we exploit this fact by training a large ensemble of students to estimate R and hence recover the teacher's true error.

The Student-Teacher model also explains why NNs can generalize even when trained to Interpolation on noisy data (which has been a source of confusion [93]). In this model, the Generalization Error $\bar{\mathcal{E}}_{gen}^T$ is a simple function of the overlap R between the Teacher T and the Students S, i.e., $\bar{\mathcal{E}}_{gen}^T \sim \langle 1 - \epsilon(R) \rangle_{\mathbf{s}}^\beta$. So even if the Teacher T is trained on noisy data, as long as there are Students S with significant overlap R with the Teacher, the Teacher Generalization Error $\bar{\mathcal{E}}_{gen}^T$ can be considerably small. For more details, see [91]

Learning the Student Moving forward, we will always assume the Teacher is trained to Interpolation because this actually corresponds to the Annealed Approximation, whereas if the Teacher makes errors, we may need to consided Quenched averages, explained below.

Imagine now that in order to estimate the empirical Average Generalization Error, $\bar{\mathcal{E}}_{gen}^{emp}$, by training a very large number of Students, and computing the average ST error on some test set. Let us break the data set into training $\mathbf{x}_\mu^{train}$ and test $\mathbf{x}_\mu^{test}$ examples, train models on the training data (that is, find the optimal model weights), and evaluate the S and T models on the test data.

The Student learning task can be written as in Eqn. (2) as the following optimization problem over the training data:

$$\underset{\{S'\}}{\text{argmin}} \sum_{\mu=1}^{N^{train}} \mathcal{L}[E_{NN}(S', \mathbf{x}_\mu^{train}), E_{NN}(T, \mathbf{x}_\mu^{train})], \tag{77}$$

If the Teacher is trained to Interpolation, then the optimization problem in Eqn. **??** is training a Student to reproduce the Ground Truth labels, so that $\mathrm{y}_\mu^S \sim y_\mu^{true}$ for both the training and test sets.

$$\underset{\{S'\}}{\text{argmin}} \sum_{\mu=1}^{N^{train}} \mathcal{L}[E_{NN}(S', \mathbf{x}_\mu^{train}), \mathrm{y}_\mu^{true}], \tag{78}$$

But if not, then the Student is reproducing the possibly incorrect Teacher labels, and, importantly, the Student S now depends explicitly on how the Teacher was trained. That is, we should denote that the learned Student explicitly depending on T, i.e. $S \to S[T]$. This will be important below.

The Average Generalization Error In either case, however, we may still estimate the Empirical Average Generalization Error by replacing the test predictions in Eqn. 76 with the student predictions $y_\mu^{test} \to \mathrm{y}_\mu^S$, and then averaging directly over the test data $\mathbf{x}_\mu^{test}$ for all possible or available test examples.

If we have a very large number of suitable Students (say, drawn from some random distribution), then we can try to estimate the Average Generalization Error of the Teacher, i.e., $\mathcal{E}_{gen}^T \approx \bar{\mathcal{E}}_{gen}^{emp}$. $\bar{\mathcal{E}}_{gen}^{emp}$ is given by an average loss, the average is over all possible Students N_S, and then over all N^{test} test data points $\mathbf{x}_\mu^{test} \in \mathbf{D}$

$$\begin{aligned} \bar{\mathcal{E}}_{gen}^{emp} &= \frac{1}{N^{test}} \sum_{\mu=1}^{N^{test}} \frac{1}{N_S} \sum_S \mathcal{L}[E_{NN}(S, \mathbf{x}_\mu^{test}), E_{NN}(T, \mathbf{x}_\mu^{test})] \\ &= \frac{1}{N_S} \sum_S \frac{1}{N^{test}} \sum_{\mu=1}^{N^{test}} \mathcal{L}[E_{NN}(S, \mathbf{x}_\mu^{test}), E_{NN}(T, \mathbf{x}_\mu^{test})], \end{aligned} \tag{79}$$

where (ideally) N^{test} is extremely large.

In Bra-Ket notation, we may also write

$$\begin{aligned} \bar{\mathcal{E}}_{gen}^{emp} &= \langle \langle \mathbf{\Delta E}_{\mathcal{L}}(S, T, \mathbf{x}) \rangle_S \rangle_{\mathbf{x}^{test}} \\ &= \langle \langle \mathbf{\Delta E}_{\mathcal{L}}(S, T, \mathbf{x}) \rangle_{\mathbf{x}^{test}} \rangle_S \end{aligned} \tag{80}$$

where $\mathbf{\Delta E}_{\mathcal{L}}(S, T, \mathbf{x}) \coloneqq \mathcal{L}[E_{NN}(S, \mathbf{x}_\mu^{test}), E_{NN}(T, \mathbf{x}_\mu^{test})]$. For the empirical estimate, it does not matter what order we take the sums in, but we are not estimating the the True Average Generalization Error of the Teacher, $\bar{\mathcal{E}}_{gen}^T$, unless T is trained to Interpolation. For the theoretical estimate, however, the order can be important, and this also depends on if T is trained to Interpolation or not. [27]

Annealed vs. Quenched Averages Recall that in the `StatMech` approach to computing errors, we do not break the data into training and test, but, instead, to obtain the Average Generalization Error, $\bar{\mathcal{E}}_{gen}$, use the Generating Function approach. In doing this, we need to compute both the Thermal Average over the model weights (S, T), and also take the data average over the entire available model data set $\mathcal{D}$ And the order can matter.

In the case where the Teacher is trained to Interpolation, may may train the Student independently of *when* the Teacher. But if Teacher is *not* trained to Interpolation, then formally we

[27] This approach can be likened to the Bootstrap method [94] used for error estimation. However, the Bootstrap method predominantly emphasizes variations in the input data $\mathbf{x}^n \in \mathbf{D}$, while in this context, we are essentially bootstrapping over the students S.

must train the Teacher first to obtain target predictions for the Student. That is, the Student formally depends on the Teacher, $S[T]$. The empirical errors in T would then formally depend on the specific instantiation of the data $\boldsymbol{\xi}^n \in \mathcal{D}$, and therefore, conceptually imagine that we must first average over the data before averaging over the weights. Training to Interpolation correpsonds conceptually to working with a model in the Annealed Approximation, whereas not correpsonds to Quenched case.

Practically, when the Teacher is not trained to Interpolation, we may need to resample the training data and training an ensemble of models to compensate for anomalies in training data (bad labels, noise, etc.) that may cause the underlying model to overfit to the training data. Theoretically, within `SMOG`, this is equivalent to *quenching* the system to the data (a term analogous to quickly cooling a physical system, frezing in any defects). In contrast, when one *anneals* a physical system, one heats up and cools it down slowly, and repeatedly, thereby removing any defects (of data anomalies for NNs, or material defects in a physical system).

In `StatMech`, one can perform a so-called quenched average using a replica calculation, effectively removing the dependence on test and/or training data from the final estimate for $\bar{\mathcal{E}}^{emp}_{gen}$. However, the theoretical quenched result may differ significantly from the annealed case when the underlying model is unrealizable [8]. This may occur when the training data is very noisy and/or the model architecture is such that it can not correctly predict all the training labels. In such cases, the model will always have some finite, non-zero Average Training Error, $\bar{\mathcal{E}}_{train} > 0$, even in the large-$N$ limit of infinite data. In such a case, this indicates a highly complex error landscape with many local minima separated by extremely high barriers, and a slowing down of the dynamics.[28]

While it is commonplace to train ensembles and/or use cross-validation when training small models (as the above discussion assumes), this could be extremely expensive and impractical in modern ML, e.g., for very large models like LLMs [**?**]. For such massive NNs, one needs a theory that can detect anomalies in training directly from observations during and/or after training. This is a hallmark of the `SETOL` approach, and it distinguishes `SETOL` from the classic `StatMech` approach.

Generalization Gap vs. Model Quality We should distinguish between what is typically done in the `SLT` literature versus the `StatMech` approaches. In `SLT`, one is typically interested in modeling the *Generalization Gap*. The Generalization Gap quantifies the difference between a models performance on training data versus unseen test data:

$$\mathcal{E}^{emp}_{gap} := \mathcal{E}_{train}[\mathbf{x}^{train}] - \mathcal{E}_{gen}[\mathbf{x}^{test}] \tag{81}$$

In contrast, in `StatMech` approaches, one considers the Model Generalization Accuracy directly, which is sometimes called the Model Quality in the `SLT` literature. Model quality is an indication of the models accuracy, precision, recall, or any other relevant metric based on the task at hand.

While related, in developing an analytic theory, the Generalization Gap and the Model Quality (or Model Generalization Accuracy) require conceptually different approaches. This is because the Generalization Gap depends on a specific realization of the training data, whereas our Model Generalization Accuracy will be formulated on a random training data set (and then corrected later with empirical data). In this sense, any theory of the Generalization Gap requires a formalism where the predicted model error is Quenched to the training data, which is not what we want. In contrast, the Model Generalization Accuracy will be formulated using the Annealed Approximation (AA), and is therefore both conceptually and mathematically simpler. Notably, the `HTSR` model Quality metric trend well when compared to the Model Generalization Accuracy, whereas methods from the `SLT` literature tend to only describe the Generalization Gap well.[28]

[28] In modern ML parlance, one might say the model can not be evaluated at interpolation, although in practice such a model might have a zero empirical Training Error since it may overfit the specific training data.

4.3.2 Theoretical Student-Teacher Average Generalization Error ($\bar{\mathcal{E}}_{gen}^{ST}$),

Here, we seek a simple, formal expression for the Student-Teacher Average Generalization Error, $\bar{\mathcal{E}}_{gen}^{ST}$, that can be used as the starting point for our extended Semi-Empirical theory.

The Idealized Data To develop a Semi-Empirical theory of the Teacher Generalization Error, $\mathcal{E}_{gen}^{T}$, instead of training and evaluating a NN model using real data ($\mathbf{x}$), we seek a simple, analytical expression with parameters that can be fit to empirical measurements. So in addition to using a model for our NN, we must specify a idealized model for the data. In a real NN, the data $\mathbf{x}$ is correlated, and, in fact, very strongly correlated; and this is reflected in the layer weight matrices. However, to be tractable, our starting theoretical expressions use uncorrelated (i.i.d) data. Formally, we must replace the correlated data with some uncorrelated, random model of the data, i.e., $\mathbf{x} \to \boldsymbol{\xi}$. As described in Figure 6, our Data Model is a standard Gaussian $N(0, \sigma^2\mathbb{I})$ model for the input data

$$\mathbf{x}_\mu \to \boldsymbol{\xi}_\mu, \;\; \boldsymbol{\xi}_\mu \in N(0, \sigma^2\mathbb{I}), \tag{82}$$

where $N(0, \sigma^2\mathbb{I})$ denotes a Gaussian distribution with zero mean and variance $\sigma^2 = \frac{1}{M}$, $\boldsymbol{\xi}_\mu$ is normalized such that $\|\boldsymbol{\xi}_\mu\|_F^2 = 1$ for all N data vectors.

We make this distinction between Actual and Model Data to emphasize that, later, we will use our so-called Semi-Empirical procedure to account for the real correlations in the actual data phenomenologically by taking some analytical parameter of the theory and fitting it to the real world observations, here, on the ESD of the NN weight matrices.

The ST Error Model and the Annealed Potential $\epsilon(\mathbf{s}, \mathbf{t})$. We now model Teacher error $\bar{\mathcal{E}}_{gen}^{T}$ with the *Average ST Generalization Error* $\bar{\mathcal{E}}_{gen}^{ST}$, which is obtained by *first* computing the ST error function $\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi})$ over the set of *all* possible N input examples $\boldsymbol{\xi}$. Define the data-dependent ST test error function –or Energy difference– as

$$\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \coloneqq \sum_{\mu=1}^{N} \mathcal{L}[E_{NN}(\mathbf{s}, \boldsymbol{\xi}_\mu), E_{NN}(\mathbf{t}, \boldsymbol{\xi}_\mu)]. \tag{83}$$

where $\mathcal{L}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi})$ is simply the ℓ_2 loss. This measures the error between the Student and the Teacher; it is zero when their predictions are identical, ($\mathrm{y}_\mu^S = \mathrm{y}_\mu^T$) when ($\boldsymbol{\xi} = \boldsymbol{\xi}_\mu$), and is nonzero otherwise.

We aim to derive a simple expression for the Average ST Generalization Error, $\bar{\mathcal{E}}_{gen}^{ST}$, and to do this, we define the Annealed Error Potential for the data-averaged ST Generalization Error $\epsilon(\mathbf{s}, \mathbf{t})$, as in Eqn. 30, as:

$$\epsilon(\mathbf{s}, \mathbf{t}) = \langle \mathbf{\Delta E}_{\mathcal{L}}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}^N} \coloneqq \frac{1}{N} \int d\mu(\boldsymbol{\xi}^N) \mathbf{\Delta E}_{\mathcal{L}}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \tag{84}$$

The measure $d\mu(\boldsymbol{\xi}^N)$ will end up being a Gaussian measure over N samples (see Appendix A.2), and the intent is to evaluate it in the large-N limit, thereby sampling all possible inputs in the model space, $\boldsymbol{\xi} \in \mathcal{D}$.

As in Eqn. **??** (Section 4.2), by applying the AA, we can rewrite the Average ST Generalization Error, $\bar{\mathcal{E}}_{gen}^{ST}$: first, a simple average over all the possible inputs $\boldsymbol{\xi}$; and, second, then as a Thermal average over all Students S, in the AA, and at high-T

$$\bar{\mathcal{E}}_{gen}^{ST} \coloneqq \langle \epsilon(\mathbf{s}, \mathbf{t}) \rangle_{\mathbf{s}}^{\beta}. \tag{85}$$

(Recall that in this regime $\bar{\mathcal{E}}_{gen}^{ST} = \bar{\mathcal{E}}_{gen}^{an,hT}$.)

In the classic `StatMech` approach, the average $\langle \cdots \rangle_{\mathbf{s}}^{\beta}$ is a Thermal Average in the canonical ensemble with β fixed, as explained in Section 4.2. Here, we will do something similar, as the Student average $\langle \cdots \rangle_{\mathbf{s}}^{\beta}$ will be computed from the associated generating function $\beta \mathbf{\Gamma}_{\bar{Q}^2}^{IZ}$ for the matrix-generalized case (i.e., an HCIZ integral defined over all students, and in the large-N limit).)

Recall that above, the empirical estimate for $\bar{\mathcal{E}}_{gen}^{emp}$ depended on a specific instantiation of the model for the training data $\mathbf{x}_{\mu}^{train}$, i.e $\bar{\mathcal{E}}_{gen}^{ST}$ is Quenched to the training data. For that reason, for the final result, we needed to take a second, quenched average over all possible data sets. Here, we do not need to consider this and always work in the Annealed Approximation (AA). This is because we incorporate the specific effects of the real-world training data $(\mathbf{x}^n)$ after we derive our formal expressions by fitting the model parameters to empirical data. The final expression for $\bar{\mathcal{E}}_{gen}^{ST}$, derived below, will be generalized to $\bar{\mathcal{E}}_{gen}^{NN}$, matrix-generalization of the classic `StatMech` formula for the Linear Perceptron, in the Annealed and High-T approximations. (see Appendix A.2).

The Annealed Potential as a function of the overlap ($\epsilon(R)$). We want an expression for the data average of the ST test error, from Eqn. 84, generalized from Perceptron vectors to NN layer weight matrices. For the Perceptron, one obtains different expressions for the ST error function, depending on the type of activation function $h(x)$ in Eqn. **??**; The simplest are the Linear and Boolean Perceptrons, and for both (and with ℓ_2 loss), $\epsilon(\mathbf{s}, \mathbf{t})$ is simply a function of the ST overlap R [8]. This gives $\epsilon(\mathbf{s}, \mathbf{t}) \to \epsilon(R)$, where

$$R = \frac{1}{N} \mathbf{s}^{\top} \mathbf{t} = \frac{1}{N} \sum_{i=1}^{M} s_i t_i, \tag{86}$$

which is simply the dot product between the M-dimensional Student $\mathbf{s}$ and Teacher $\mathbf{t}$ weight vectors, and normalized by the number of training instances N. For a Linear Perceptron [8],[29] with activation function $h(x) = x$, the error function is

$$\epsilon(R) = 1 - R. \tag{87}$$

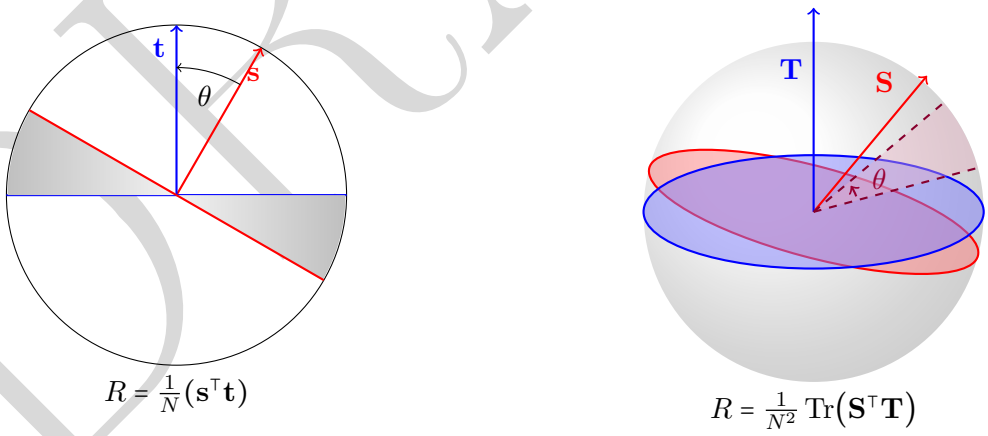


Figure 9: Comparison of 2D and 3D representations of the vector and matrix Student–Teacher overlap R. **Left:** $R = \frac{1}{N}\mathbf{s}^{\top}\mathbf{t}$. **Right:** $R = \frac{1}{N^2}\,\mathrm{Tr}\left(\mathbf{S}^{\top}\mathbf{T}\right)$ with conic sections on the sphere (red $\mathbf{S}$, blue $\mathbf{T}$), plus a purple wedge for the angle.

[29] In the classic approach for the ST model, the theory examined different expressions $\epsilon(R)$. For example, one can consider the Boolean Perceptron [8, 95], with activation function $h(x) = \mathrm{sgn}(x)$, i.e., the Heaviside step function. Then, the error is $\epsilon(R) = 1 - \frac{1}{\pi}\arccos(R)$. In both cases, perfect learning occurs when $R = 1$ [8].

Derivation of the ST error ($\epsilon(R)$) for the Linear Perceptron. To derive Eqn. 87, define the data-dependent ST error (Eqn. 83) in terms of an ℓ_2 loss function

$$\begin{aligned}\mathbf{\Delta E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n) =& \frac{1}{2}(\mathbf{y}^S-\mathbf{y}^T)^\top(\mathbf{y}^S-\mathbf{y}^T)\\ =& \frac{1}{2}\left[(\mathbf{y}^S)^\top\mathbf{y}^S - 2(\mathbf{y}^S)^\top\mathbf{y}^T + (\mathbf{y}^T)^\top\mathbf{y}^T\right]\\ =& N-(\mathbf{y}^S)^\top(\mathbf{y}^T)\\ =& N(1-\eta(\mathbf{s},\mathbf{t},\boldsymbol{\xi})),\end{aligned} \tag{88}$$

where we define $\eta(\mathbf{s},\mathbf{t},\boldsymbol{\xi}) \coloneqq \frac{1}{N}\mathbf{y}_S^\top\mathbf{y}_T$, called the *data-dependent Self-Overlap*; we expand this below. The expression $\eta(\mathbf{s},\mathbf{t},\boldsymbol{\xi})$ is analogous to the ST overlap R, but before the data has been integrated out. It is convenient to work directly with the Self-Overlap $\eta(\mathbf{s},\mathbf{t},\boldsymbol{\xi})$ because it will appear later in Eqn. 104 (in Section 5), when formulating the matrix-generalized overlap operator $\mathbf{R}$.

In defining $\eta(\mathbf{s},\mathbf{t},\boldsymbol{\xi})$, we replace the labels $(\mathbf{y}_S,\mathbf{y}_T)$ with the Energy functions E_{NN} that generate them, giving an expression in terms of the weights $(\mathbf{s},\mathbf{t})$ and the Gaussian data variables $(\boldsymbol{\xi})$. We will then integrate out the data variables, leaving an expression just in terms of the weights. Using the E_{NN} Energy generating or output function (Eqn. 1, **??**, 74), we can replace the label vectors $\mathbf{y}^T,\mathbf{y}^S$ as

$$\mathbf{y}^S = \mathbf{s}^\top\boldsymbol{\xi},\;\; \mathbf{y}^T = \mathbf{t}^\top\boldsymbol{\xi}, \tag{89}$$

which gives

$$\eta(\mathbf{s},\mathbf{t},\boldsymbol{\xi}) = \tfrac{1}{N}(\mathbf{s}^\top\boldsymbol{\xi})^\top(\mathbf{t}^\top\boldsymbol{\xi}) = \tfrac{1}{N}\boldsymbol{\xi}^\top(\mathbf{s}^\top\mathbf{t})\boldsymbol{\xi}. \tag{90}$$

or, more simply, after integrating over the data, we have the *data-independent Self-Overlap*

$$\eta(\mathbf{s},\mathbf{t}) = \langle\eta(\mathbf{s},\mathbf{t},\boldsymbol{\xi})\rangle_{\overline{\boldsymbol{\xi}^N}} = \tfrac{1}{N}(\mathbf{s}^\top\mathbf{t}) \tag{91}$$

We want to evaluate this as an Annealed Error Potential $\epsilon(R)$ for the data-averaged ST test error, as in Eqn. 84. To do this, we need to compute the average or Expected Value over all N possible input data vectors $\boldsymbol{\xi}$ (i.e., $d\mu(\boldsymbol{\xi}^N) = \mathcal{D}\boldsymbol{\xi}^N P(\boldsymbol{\xi}^N)$).

$$\langle\mathbf{\Delta E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^N}} = \int d\mu(\boldsymbol{\xi}^N)(1-\eta(\boldsymbol{\xi})) \tag{92}$$

$$\begin{aligned}=& \int d\mu(\boldsymbol{\xi}^N)(1-\boldsymbol{\xi}^\top\tfrac{1}{N}\mathbf{s}^\top\mathbf{t}\boldsymbol{\xi})\\ =& \int d\mu(\boldsymbol{\xi}^N) - \int d\mu(\boldsymbol{\xi}^N)\boldsymbol{\xi}^\top\tfrac{1}{N}\mathbf{s}^\top\mathbf{t}\boldsymbol{\xi}\end{aligned} \tag{93}$$

$$\begin{aligned}=& 1-\int d\mu(\boldsymbol{\xi}^N)\boldsymbol{\xi}^\top\tfrac{1}{N}\mathbf{s}^\top\mathbf{t}\boldsymbol{\xi}\\ =& 1-\int d\mu(\boldsymbol{\xi})\boldsymbol{\xi}^\top R\boldsymbol{\xi}\\ =& 1-R\int d\mu(\boldsymbol{\xi})\boldsymbol{\xi}^\top\boldsymbol{\xi},\end{aligned}$$

where this holds because the elements of $\boldsymbol{\xi}$ are i.i.d. Since $d\mu(\boldsymbol{\xi}^N)$ is a measure over a (multi-variate) Gaussian distribution, The data vectors $\boldsymbol{\xi}$ are normalized (See Section A.2.1) such that the second term on the R.H.S. is unity and we recover (i.e.,Eqn. 30)

$$\epsilon(R) = \langle\mathbf{\Delta E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^N}} = 1-R. \tag{94}$$

In traditional `StatMech` (e.g., [8]), one is interested in how the *Total Model Generalization Accuracy* $\mathcal{E}_{gen}(R)$ depends on R. With these simple error functions, Eqn. 85 reduces to a function over R, and the Average ST Generalization Error $\mathcal{E}^{ST}_{gen}(R)$ is then obtained by averaging over the Students

$$\bar{\mathcal{E}}^{ST}_{gen}(R) = \langle \epsilon(R) \rangle^{\beta}_{\mathbf{s}} = \left\langle 1 - \langle \eta(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \rangle_{\overline{\boldsymbol{\xi}^N}} \right\rangle^{\beta}_{\mathbf{s}} = \langle 1 - \eta(\mathbf{s}, \mathbf{t}) \rangle^{\beta}_{\mathbf{s}} = \left\langle 1 - \tfrac{1}{N} \mathbf{s}^{\mathsf{T}} \mathbf{t} \right\rangle^{\beta}_{\mathbf{s}} = \langle (1 - R) \rangle^{\beta}_{\mathbf{s}} \,, \tag{95}$$

where $\langle \cdots \rangle^{\beta}_{\mathbf{s}}$ is a Thermal Average over the Student weight vector $\mathbf{s}$.

The Model Quality for the ST Perceptron, $\bar{\mathcal{Q}}^{ST}$ is just the Average Generalization Accuracy, so we can write

$$\bar{\mathcal{Q}}^{ST} := 1 - \bar{\mathcal{E}}^{ST}_{gen}(R) = \langle 1 - \epsilon(R) \rangle^{\beta}_{\mathbf{s}} = \left\langle \langle \eta(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \rangle_{\overline{\boldsymbol{\xi}^N}} \right\rangle^{\beta}_{\mathbf{s}} = \langle \eta(\mathbf{s}, \mathbf{t}) \rangle^{\beta}_{\mathbf{s}} = \left\langle \tfrac{1}{N} \mathbf{s}^{\mathsf{T}} \mathbf{t} \right\rangle^{\beta}_{\mathbf{s}} = \langle R \rangle^{\beta}_{\mathbf{s}} \,. \tag{96}$$

Eqn. 96 is the starting point for deriving a `SemiEmpirical` theory for the `WeightWatcher` quality metrics (`Alpha`,`AlphaHat`); see Section 5.1. To generalize this expression, we will start with the Self-Overlap $\eta(\mathbf{S}, \mathbf{T}, \boldsymbol{\xi})$ for a Multi-Layer Perceptron (MLP3) in Section 5.

Before doing this, however, we note that we can obtain this expression for $\mathcal{E}^{ST}_{gen}$ by defining the Annealed Hamiltonian $H^{an}_{hT}(R)$, at high-Temperature, as in Section 4.2, Eqn. 50. Indeed, it is really $H^{an}_{hT}(R) = \epsilon(R)$ that we must generalize to the matrix case, which we do (using a technique similar to a Replica calculation, but still in the AA). For more details, see Appendix A.2. (In particular, doing this allows us to define the normalization needed later for the `TRACE-LOG` condition).

Quantity	Traditional SMOG	Linear Perceptron in Traditional SMOG	Matrix Generalization for SETOL
Total (Idealized) Data Error	$\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)$ (31)	$\mathbf{\Delta E}_{\mathcal{L}}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi})$ (88)	$\mathbf{\Delta E}_{\ell_2}(\mathbf{S}, \mathbf{T}, \boldsymbol{\xi})$ (99)
Annealed Hamiltonian (Data-Averaged Error)	$H_{hT}^{an} = \epsilon(\mathbf{w})$ (30) (AA, at high-T)	$H_{hT}^{an}(R) = \epsilon(\mathbf{s}, \mathbf{t}) = 1 - R$ (94) (and at large-N)	$H_{hT}^{an}(\mathbf{R}) = \mathbf{I}_M - \mathbf{R}$ (202) (only for a layer)
Self-Overlap	$\eta(\mathbf{w}) = 1 - \epsilon(\mathbf{w})$	$\eta(\mathbf{s}, \mathbf{t}) = \frac{1}{N}\mathbf{s}^\intercal \mathbf{t}$ (91)	$\eta(\mathbf{S}, \mathbf{T}) = \frac{1}{N}\mathbf{S}^\intercal \mathbf{T}$ (104)
Model Quality in terms of Layer Quality	$\bar{\mathcal{Q}} := 1 - \bar{\mathcal{E}}_{gen}$	$\bar{\mathcal{Q}}^{ST} := 1 - \bar{\mathcal{E}}_{gen}^{ST}$ (59)	$\bar{\mathcal{Q}}^{NN} := 1 - \bar{\mathcal{E}}_{gen}^{NN}$ (59) $\bar{\mathcal{Q}}^{NN} := \prod_L \bar{\mathcal{Q}}_L^{NN}$

Table 7: Summary of key quantities compared across traditional SMOG models, the Student-Teacher (ST) Linear Perceptron–in the Annealed Approximation (AA) and at high-Temperature (high-T) and at large-N, and the matrix-generalized forms as the starting point to frame SETOL. The Total Data Model Error represents the difference between the model and its labels for the ST model between the Student and Teacher predictions. The Annealed Hamiltonian is the Energy function for this Error after it is averaged over the model for the training data (an n or N-dimensional i.i.d. Gaussian model, i.e., $\boldsymbol{\xi}^n$ or $\boldsymbol{\xi}^N$). In the AA, the Annealed Hamiltonian is equal to the Annealed Error Potential. For the ST model, this is one minus the overlap, $(1 - R)$; for the SETOL, this is the (M-dimensional) identity minus the overlap operator/matrix, $\mathbf{I}_M - \mathbf{R}$. The Self-Overlap $\eta(\cdots)$ is used to describe the Accuracy (as opposed to the Error) for both the ST model and its matrix-generalized form. Finally, the different forms of the Quality are defined. Generally speaking, the Quality $\bar{\mathcal{Q}}$ is an approximation to some measure of 1 minus the Average Generalization Error, $(\bar{\mathcal{Q}} := 1 - \bar{\mathcal{E}}_{gen})$ (in the AA, at high-T, at large-N, and with whatever else approximations are applied). For the ST model, having just 1 layer, the Model Quality and the Layer Quality are the same, and denoted $\bar{\mathcal{Q}}^{ST}$. For SETOL, the Model Quality $\bar{\mathcal{Q}}^{NN}$ is a product of individual Layer Qualities $\bar{\mathcal{Q}}_L^{NN}$. (Note that the final SETOL Layer Quality $\bar{\mathcal{Q}}$ is defined in terms of the Layer Quality-Squared $\bar{\mathcal{Q}}^2$, and the starting point for this is expressed with the Layer Quality-Squared Hamiltonian $\mathbf{H}_{\bar{\mathcal{Q}}^2} = \mathbf{R}^\intercal \mathbf{R}$).

5 Semi-Empirical Theory of the HTSR Phenomenology

In this section, we present the main technical elements of our Semi-Empirical Theory of Deep Learning (SETOL). Our goal is to explain and, where possible, derive the HTSR PL metrics Alpha (α) and AlphaHat $\hat{\alpha}$ from first principles, and, in doing so, also present the TRACE-LOG condition and newly proposed WeightWatcher DetX metric. To do this, we introduce a Matrix Generalization of the Student-Teacher model for a Linear Perceptron (See Section 4.3.2), adapted here for a (3-layer) Multi-Layer Perceptron (MLP3). We seek a theory for the Layer Quality $\bar{\mathcal{Q}} = \bar{\mathcal{Q}}_L^{NN}$ of a NN, where this Layer Quality now corresponds to the (approximate) contribution each layer makes to the total generalization accuracy, or total Quality $\bar{\mathcal{Q}}^{NN}$. For technical reasons, we actually seek a formal expression(s) for the Layer Quality-Squared, $\bar{\mathcal{Q}}^2 \approx \bar{\mathcal{Q}}^2$. We say that the SETOL is Semi-Empirical because the final result $\bar{\mathcal{Q}}^2$ is expressed directly in terms of the empirically observable spectral properties of the Teacher layer weight matrix $\mathbf{T} = \mathbf{W}$.

5.1 **Matrix Generalization of the ST Model.** Section 5.1 generalizes classical StatMech vector-based ST model of Section 4.3 to obtain a Layer Quality for a single layer in a NN. It starts by first formulating the learning problem for the NN generalization accuracy or quality, $\bar{\mathcal{Q}}^{NN}$, of a 3-layer MLP (MLP3). We then replace vectors with $N \times M$ matrices $\mathbf{s}, \mathbf{t} \rightarrow \mathbf{S}, \mathbf{T}$, and obtain and expression for the NN Self-Overlap $\eta(\mathbf{S}, \mathbf{T}, \boldsymbol{\xi})$, which then gives a matrix-generalized overlap operator $\mathbf{R} := \langle \eta(\mathbf{S}, \mathbf{T}, \boldsymbol{\xi}) \rangle_{\overline{\boldsymbol{\xi}^N}} = \frac{1}{N}\mathbf{S}^\intercal\mathbf{T}$. This can be related to a single-layer matrix-generalization of the ST Annealed Hamiltonian, presented in Appendix A.2, $H_{hT}^{an} := M - \mathbf{R}$, where, importantly, the scalar overlap R is now a matrix $\mathbf{R}$ of M adjustable parameters.

5.2 **The Layer Quality-Squared $\bar{\mathcal{Q}}^2$** Section 5.2 presents the expression for NN Layer Quality-Squared $\bar{\mathcal{Q}}^2$. Following the ST analogy, we define a Thermal Average over possible Student weight matrices $\mathbf{S}$ for the matrix overlap, giving $\bar{\mathcal{Q}}_L^{NN} := \left\langle H_{hT}^{an} \right\rangle_{\mathbf{S}}^{\beta} = \langle \mathbf{R} \rangle_{\mathbf{S}}^{\beta}$, For technical reasons, however, we actually seek the (approximate) Layer Quality-Squared, $\bar{\mathcal{Q}}^2 \approx \bar{\mathcal{Q}}^2$, defined as $\bar{\mathcal{Q}}^2 := \langle \mathbf{R}^\intercal\mathbf{R} \rangle_{\mathbf{S}}^{\beta}$. To evaluate $\bar{\mathcal{Q}}^2$, rather than sampling all random Student matrices $\mathbf{S}$ directly, we switch measures to the Student correlation matrices $\mathbf{A}_2 = \frac{1}{N}\mathbf{S}\mathbf{S}^\intercal$. Importantly, we argue that the measures $d\mu(\mathbf{A}_1) \leftrightarrow d\mu(\mathbf{A}_2)$, can be interchanged for our purposes, making them effectively equivalent. This reparametrization leads us to an integral of the HCIZ type (as in Eqn. 70) which, as shown by Tanaka [82, 83].

Then, we introduce the Effective Correlation Space (ECS), and two key approxmations, the TRACE-LOG Condition and the Independent Fluctuation Approximation (IFA). The TRACE-LOG condition states that the determinant of the (effective) Student correlation matrix is unity, $\det\left(\tilde{\mathbf{A}}\right) = 1$. Critcially, this condition can be tested empirically by assuming the (effective) Teacher correlation matrix also follows the TRACE-LOG condition, $\det\left(\tilde{\mathbf{X}}\right) = 1$, and this is a key result of this work. Finally, we impose the IFA (described below) because it is necessary for the final result.

5.3 **The Large-N limit** Section 5.3 presents the core result, (as in Eqn. 15), a closed-form or semi-analytic expressions for the Layer Quality-Squared $\bar{\mathcal{Q}}^2$ formed in the large-N limit. Restricted to the ECS, and under the TRACE-LOG condition and the IFA, our HCIZ integral for $\bar{\mathcal{Q}}^2$ becomes tractable at large-N, givin an expression that can be parameterized in terms of $\tilde{M}$ eigenvalues $\tilde{\lambda}$ of the Teacher correlation matrix restricted to the ECS $\tilde{\mathbf{X}}$. In doing this, the $\tilde{M}$ Teacher eigenvalues are treated as experimental observables, and become the effective Semi-Empirical parameters (i.e, α, λ_{max}) of our SETOL.

5.4 **Modeling the Heavy-Tailed R-transform** Section 5.4 presents several models of different R-transforms. Evaluating $\bar{\mathcal{Q}}^2$ requires evaluating selecting an R-transform $R(z)$ for the Teacher ESD, and also ensure that it is analytic and single-valued on the domain of interest– the `ECS` and/or tail of the ESD. We examine four possible models for $R(z)$: *(i)* the *Bulk+Spikes* (BS), *(ii)* the *Free Cauchy* (FC) model, *(iii)* the *Inverse Wishart* (IW) model, and *(iv)* the Levy-Wigner (LW) model. First, as a trivial case, the tail of ESD can be treated as a collection of spikes, and the ESD is simply a sum of Dirac delta functions; in this case, $\bar{\mathcal{Q}}^2$ becomes a "Tail Norm", the Frobenius Norm of the PL tail. When the layer is Ideal, i.e., $\alpha \sim 2$ and $\det\left(\tilde{\mathbf{X}}\right) \sim 1$, one can use *Free Cauchy* (FC) model. and the resulting Layer Quality $\bar{\mathcal{Q}}$ is approximately the Spectral Norm λ_{max}. Since λ_{max} increases with decreasing α, the FC model yields the `HTSR Alpha` metric. One can also use *Inverse Wishart* (IW) model. As required, the IW R-transform contains a *branch cut* in the complex plane which aligns with the start of the `ECS` Power Law tail. Finally, Using the Levy-Wigner (LW) model, one can model cases where $\alpha \leq 2$ and derive the `HTSR AlphaHat` metric.

These core elements form a bridge between well-established empirical properties of large-scale NNs and a tractable ST-based theory. In the subsequent sections, we formalize the key steps: *(i)* setting up the matrix-based ST problem, *(ii)* defining our HCIZ integrals over restricted correlation matrices (`ECS`), and *(iii)* analyzing the resulting *Layer Quality* (or *Quality-Squared*) expressions in the large-N limit.

5.1 Multi-Layer Setup: MLP3

In this section, we describe the matrix generalization of the ST model of the Linear Perceptron; and, in particular, a matrix-generalized version of the key quantities we derived in Section 4.3.2.

A simple model. Consider a simple NN with three layers (two hidden and an output), i.e., a three-layer Multi-Layer Perceptron, denoted as the *MLP3* model. (This is a *very* simple model of a modern NN with hundreds of layers and complex internal structure.)

Ignoring the bias terms, *Without Loss of Generality*, (WLOG), the NN outputs $E_{1\mu}, E_{2\mu}, E_{3\mu}$ for each layer, as defined in Eqn. 1, are given by:

$$\begin{aligned} E_{1\mu} &:= \frac{1}{\sqrt{N_1}} h(\mathbf{W}_1^\top \boldsymbol{\xi}_\mu), \\ E_{2\mu} &:= \frac{1}{\sqrt{N_2}} h(\mathbf{W}_2^\top \mathbf{Z}_{1\mu}), \\ y_\mu := E_{3\mu} &= \frac{1}{\sqrt{N_3}} h(\mathbf{W}_3^\top \mathbf{Z}_{2\mu}), \end{aligned} \tag{97}$$

where h is a general function or functional, denoting either a non-linear activation or a more complex layer structure, such as a CNN or an RNN. We can consider $h(\cdot)$ to be an (unspecified) activation function.

Let us specify the ST error, or Energy difference, specifically in terms of the L2 or RMSE loss:

$$\mathbf{\Delta E}_{\ell_2}(\mathbf{S}, \mathbf{T}, \boldsymbol{\xi}) = \frac{1}{2} \sum_{\mu=1}^{N} (\mathrm{y}_\mu^S - \mathrm{y}_\mu^T)^2. \tag{98}$$

We now start to develop the self-overlap η using $\mathbf{S} - \mathbf{T}$:

5.1.1 Data-Dependent Multi-Layer ST Self-Overlap ($\eta(\mathbf{S},\mathbf{T})$)

It is convenient to rewrite $\Delta\hat{\mathbf{E}}$ in Eqn. 98 as:

$$\boldsymbol{\Delta\mathbf{E}}_{\ell_2}(\mathbf{S},\mathbf{T},\boldsymbol{\xi}) \coloneqq \frac{1}{2}\,\mathrm{Tr}\left[(\mathrm{y}_\mu^S-\mathrm{y}_\mu^T)^\intercal(\mathrm{y}_\mu^S-\mathrm{y}_\mu^T)\right] = N - \mathrm{Tr}\left[(\mathbf{y}^S)^\intercal\,\mathbf{y}^T\right] = N(1-\eta([\mathbf{S}_l,\mathbf{T}_l],\boldsymbol{\xi})) \tag{99}$$

where the Self-Overlap $\eta([\mathbf{S}_l,\mathbf{T}_l],\boldsymbol{\xi})$ is of the same form as the (vector) Linear Perceptron (in Eqn. 88). Note that $\eta([\mathbf{S}_l,\mathbf{T}_l],\boldsymbol{\xi})$ depends on the (model) data $\boldsymbol{\xi}$ because we have not evaluated the expected value $\langle\cdots\rangle_{\boldsymbol{\xi}}$ yet.

Using the general expression from Eqn. 97 for the action of the NN on the input data $\boldsymbol{\xi}$, we can write the formal expression of the ST error for the simple MLP3 model as:

$$\begin{aligned}
&\langle\eta([\mathbf{S}_l,\mathbf{T}_l],\boldsymbol{\xi})\rangle_{\overline{\boldsymbol{\xi}^N}} \qquad\qquad (100)\\
&\quad= \langle\eta(\mathbf{S}_1,\mathbf{S}_2,\mathbf{S}_3,\mathbf{T}_1,\mathbf{T}_2,\mathbf{T}_3,\boldsymbol{\xi})\rangle_{\overline{\boldsymbol{\xi}^N}}\\
&\quad\coloneqq \mathrm{Tr}\left[\frac{1}{\sqrt{N_3}}h\left(\mathbf{S}_3^\intercal\frac{1}{\sqrt{N_2}}h\left(\mathbf{S}_2^\intercal\frac{1}{\sqrt{N_1}}h\left(\mathbf{S}_1^\intercal\boldsymbol{\xi}\right)\right)\right)^\intercal\times\frac{1}{\sqrt{N_3}}h\left(\mathbf{T}_3^\intercal\frac{1}{\sqrt{N_2}}h\left(\mathbf{T}_2^\intercal\frac{1}{\sqrt{N_1}}h\left(\mathbf{T}_1^\intercal\boldsymbol{\xi}\right)\right)\right)\right].
\end{aligned}$$

So far, we have not used any particular assumption on the form of the NN or the data, other than that the layer structure used to write the explicit expression for the form eventually needed, $\eta(\mathbf{S},\mathbf{T})$, a single layer Self-Overlap. As a next step, we show which assumptions are needed in order to reformulate the setup as an effectively a single layer linear model for a NN.

5.1.2 A Single Layer Matrix Model

Following others in the literature [96], and for simplicity, one can restrict to the simplifying case that the function $h(x)$ is the identity function. To evaluate Eqn. 100, there are three possibilities. First, we can multiply all the matrices together, and treat a multi-layer NN effectively as a single layer. Under this assumption, Eqn.100 simplifies to

$$\langle\eta([\mathbf{S}_l,\mathbf{T}_l],\boldsymbol{\xi})\rangle_{\overline{\boldsymbol{\xi}^N}} = \tfrac{1}{N_3\,N_2\,N_1}\,\mathrm{Tr}\left[\boldsymbol{\xi}^\intercal\mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\mathbf{T}_3^\intercal\mathbf{T}_2^\intercal\mathbf{T}_1^\intercal\boldsymbol{\xi}\right]. \tag{101}$$

While this is possible, it would not lead to layer-by-layer insights (as `HTSR`-based approaches do). Second, we could attempt to expand Eqn. 101 into inter- and intra-layer terms, which we could readily do if the S and T matrices were square and the same shape, and then apply Wick's theorem:

$$\eta([\mathbf{S}_l,\mathbf{T}_l]) \approx \prod_{l=1}^{L}\tfrac{1}{N_l}\langle\eta(\mathbf{S}_l,\mathbf{T}_l,\boldsymbol{\xi})\rangle_{\overline{\boldsymbol{\xi}^N}} = \prod_{l=1}^{L}\tfrac{1}{N_l}\,\mathrm{Tr}\left[\boldsymbol{\xi}^\intercal\mathbf{S}_l^\intercal\mathbf{T}_l\boldsymbol{\xi}\right) + \text{intra-layer cross terms}. \tag{102}$$

However, these matrices are not square, and we don't know how to express the intra-layer cross terms. Finally, we can simply assume that the individual layers are statistically independent, in which case we can treat each layer independently. By ignoring the intra-layer cross-terms, let us write the single-layer self-overlap $\eta(\mathbf{S}_l,\mathbf{T}_l,\boldsymbol{\xi})$ as:

$$\eta(\mathbf{S}_l,\mathbf{T}_l,\boldsymbol{\xi}) = \langle\eta(\mathbf{S}_l,\mathbf{T}_l,\boldsymbol{\xi})\rangle_{\overline{\boldsymbol{\xi}^N}} \to \tfrac{1}{N}\,\mathrm{Tr}\left[\boldsymbol{\xi}^\intercal\mathbf{S}_l^\intercal\mathbf{T}_l\boldsymbol{\xi}\right]. \tag{103}$$

This third approach is the one we will adopt. Moving forward, we will drop the layer subscript, l, and we will consider a `SETOL` as a single-layer theory.

5.1.3 The Matrix-Generalized ST Overlap ($\eta(\mathbf{S},\mathbf{T})$)

We can now relate the Self-Overlap for the NN layer $\eta(\boldsymbol{\xi})_l$ in Eqn. 103 as a matrix-generalized form of the ST Annealed Error Potential $\epsilon(R) = 1 - R$. Since we can interpret the Trace as an expected value over the model data $\boldsymbol{\xi}^N$, this gives the desired

$$\eta(\mathbf{S},\mathbf{T}) \coloneqq \tfrac{1}{N}\,\mathrm{Tr}\left[\mathbf{S}^\intercal\mathbf{T}\right] \tag{104}$$

This is the matrix generalized form of Eqn. 96 for the Layer Quality.

5.2 Quality Metrics of an Individual Layer as an HCIZ Integral

In this subsection, we describe how to generalize the (Thermal) average over the Students $\langle\cdots\rangle^\beta_{\mathbf{s}}$ to an integral over random Student matrices, $\langle\cdots\rangle^\beta_{\mathbf{S}}$, called an HCIZ integral.

5.2.1 A Generating Function Approach to Average Quality-Squared of a Layer

For our matrix generalization, we need to express the Layer Quality $\bar{\mathcal{Q}}$ in terms of the data-averaged Self-Overlap in Eqn. **??** for a individual layer.

- **Student-Teacher Overlap R** For the vector-based Perceptron ST model, the data-averaged Self-Overlap appears in the expression for the Layer Quality in Eqn. 96, and is just the ST vector overlap $R = \frac{1}{N}\mathbf{s}^\intercal\mathbf{t}$. For our matrix generalization, we can define

$$\mathbf{R} = \tfrac{1}{N}\mathbf{S}^T\mathbf{T}. \tag{105}$$

  For the vector-based ST model, the ST Quality $\bar{\mathcal{Q}}^{ST}$ in Eqn. 96 is expressed as the Thermal Average $\bar{\mathcal{Q}}^{ST}(R) = \langle R\rangle^\beta_{\mathbf{s}}$. Here, we want something similar.

- **Model and Layer Qualities $\bar{\mathcal{Q}}^{NN}$, $\bar{\mathcal{Q}}^{NN}_L$** For our matrix generalization, the Model Quality $\bar{\mathcal{Q}}^{NN}$, as explained in Subsection 2.3, Eqn. 7, will be a product of individual NN Layer Qualities $\bar{\mathcal{Q}}^{NN}_L$, and, as in Eqn. 59 approximates the total NN Average Generalization Accuracy $(1-\bar{\mathcal{E}}^{NN}_{gen})$:

$$\bar{\mathcal{Q}}^{NN} \coloneqq \prod_L \bar{\mathcal{Q}}^{NN}_L \approx 1 - \bar{\mathcal{E}}^{NN}_{gen} \tag{106}$$

  The individual $\bar{\mathcal{Q}}^{NN}_L$ expresses the contribution that layer makes to the approximate total NN Average Generalization Accuracy.

- **Layer Quality-Squared $\bar{\mathcal{Q}}^2$** For technical convenience, however, rather than compute the NN Layer Quality $\bar{\mathcal{Q}}^{NN}_L$ directly, we will work with the *Average Layer Quality-Squared*, defined as

$$\bar{\mathcal{Q}}^2 \coloneqq \left\langle \mathbf{R}^\intercal\mathbf{R}\right\rangle^\beta_{\mathbf{S}} \tag{107}$$

  where $\langle\cdots\rangle^\beta_{\mathbf{S}}$ is now a Thermal Average over Student weight matrices $\mathbf{S}$– an HCIZ integral.

  This choice means that final Layer Quality $\bar{\mathcal{Q}}$ we use approximates what would be the matrix-generalized NN Layer Quality $\bar{\mathcal{Q}}^{NN}_L$ (above) as

$$\begin{aligned}\bar{\mathcal{Q}} \coloneqq \sqrt{\bar{\mathcal{Q}}^2} &= \sqrt{\langle\mathbf{R}^\intercal\mathbf{R}\rangle^\beta_{\mathbf{S}}} \\ &\approx \left\langle\sqrt{\mathbf{R}^\intercal\mathbf{R}}\right\rangle^\beta_{\mathbf{S}} \\ &\approx \langle\mathbf{R}\rangle^\beta_{\mathbf{S}} \\ &= \bar{\mathcal{Q}}^{NN}_L\end{aligned} \tag{108}$$

- **Overlap Squared** The Overlap operator (squared) $\mathrm{Tr}\left[\mathbf{R}^\intercal\mathbf{R}\right]$ is defined in terms of Eqn. 113 such that we can

$$\mathrm{Tr}\left[\mathbf{R}^\intercal\mathbf{R}\right] \coloneqq \frac{1}{N^2}\mathrm{Tr}\left[\mathbf{T}^\intercal\mathbf{S}\mathbf{S}^\intercal\mathbf{T}\right] = \frac{1}{N}\mathrm{Tr}\left[\mathbf{T}^\intercal\mathbf{A}_2\mathbf{T}\right]. \tag{109}$$

  This choice places $\bar{\mathcal{Q}}^2$ in the form of the HCIZ integral, as in Eqn. 70. See Appendix A.6 for a detailed discussion of why we choose $\mathbf{A} = \mathbf{A}_2$ here.

- **Generating Function** For the vector-based ST model, we could compute $\bar{\mathcal{Q}}^{ST}$ using a generating function, $\beta\mathbf{\Gamma}^{ST}_{\bar{\mathcal{Q}}}$, For our matrix generalization, we compute $\bar{\mathcal{Q}}$ from a *Layer Quality-Squared Generating Function* $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$, given as

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} \coloneqq \ln\int d\mu(\mathbf{S})\exp\left[N\beta\,\mathrm{Tr}\left[\mathbf{R}^\intercal\mathbf{R}\right]\right]. \tag{110}$$

  See Appendix A.3 for the derivation of Eqn. 110 (and recall the discussion in Section 4.2) . .

We can not evaluate Eqn. 110 directly; but we will be able to evaluate it if we transform it into an HCIZ integral (as in Eqn. 70). To do this, however, requires a trick which will allow us to work with different forms of the Student $\mathbf{A}$ (and Teacher $\mathbf{X}$) Correlation matrices.

From weight matrices to Correlation matrices. To evaluate $\bar{\mathcal{Q}}^2$ in terms of derivatives of Eqn. 110, we need to introduce the change of measure:

$$d\mu(\mathbf{S}) \to d\mu(\mathbf{A}), \tag{111}$$

and then restrict $d\mu(\mathbf{A})$ to resemble just the generalizing eigencomponents of the Teacher correlation matrix $\mathbf{X}$. We have two choices for the Student Correlation Matrix $\mathbf{A}$, call them $\mathbf{A}_1$ and $\mathbf{A}_2$, defined as:

$$\mathbf{A}_1 \coloneqq \frac{1}{N}\mathbf{S}^\intercal\mathbf{S} \quad (\text{which is } M\times M) \tag{112}$$

$$\mathbf{A}_2 \coloneqq \frac{1}{N}\mathbf{S}\mathbf{S}^\intercal \quad (\text{which is } N\times N). \tag{113}$$

Note that $\frac{1}{N}$ is the correct scaling on each of these. Eqn. 112 is consistent with our definition of the layer Correlation Matrix, and we use it as the starting point below to derive the Volume-Preserving `TRACE-LOG` condition (Appendix A.4). Eqn. 113 is consistent with Tanaka, which requires $\mathbf{A}$ be $N\times N$, but the we still need the a *Duality of Measures* to rederive this (Appendix A.6).[82, 83]

Duality of measures. For either form of $\mathbf{A}$, the measure $d\mu(\mathbf{A})$ is the same because we will restrict the measures to the `ECS` space of non-zero eigenvalues ($\lambda_i \gg 0$). We note that $\mathbf{A}_1$ and $\mathbf{A}_2$ have the same eigenvalues λ_i, or ESD, upto the additional zero eigenvalues ($\lambda_i = 0$) in the null space of $\mathbf{A}_2$. Consequently, both forms of $\mathbf{A}$ have the same $N-M$ non-zero eigenvalues, the same non-zero part of the ESD ($\rho(\lambda) \gg 0$), and the same Trace ($\mathrm{Tr}\left[\mathbf{A}_1\right] = \mathrm{Tr}\left[\mathbf{A}_2\right]$). In the large-N approximation, the ESD of (either form of) $\mathbf{A}$, $\rho^\infty_{\mathbf{A}}(\lambda)$, becomes continuous (and bounded), but remains zero in the null space. Consequently, when integrating over the eigenvalues λ, one can interchange $\mathbf{A}_1$ with $\mathbf{A}_2$, such that

$$\int d\mu(\mathbf{A})\,[\cdots] \leftrightarrow \int d\lambda\,[\cdots]\,\rho^\infty_{\mathbf{A}_1}(\lambda) \leftrightarrow \int d\lambda\,[\cdots]\,\rho^\infty_{\mathbf{A}_2}(\lambda) \tag{114}$$

This equivalence will be essential both to derive the `TRACE-LOG` condition (Section A.4), and to (re)derive the core result by Tanaka (Section A.6). Additionally and WLOG, we may occasionally denote the Student Correlation matrix as $\tilde{\mathbf{A}}$ instead of explicilty using $\mathbf{A}_1$ and $\mathbf{A}_2$.

5.2.2 Evaluating the Average Quality (Squared) Generating Function

We can write the generating function $\beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2}$ in Eqn. 110 in terms of $\mathbf{A}_2$, giving, as in Eqn. 12,

$$\beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2} = \ln \int d\mu(\mathbf{S}) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^T\mathbf{A}_2\mathbf{T}])}. \tag{115}$$

To recast Eqn. 115 as an HCIZ integral, as in Eqn. 70, we must perform change of measure, from $N \times M$ Student weight matrices $\mathbf{S}$ to $N \times N$ Student Correlation matrices $\mathbf{A}$, as in Eqn. 111.

When we perform the change of measure on Eqn. 115, we obtain the following expression:

$$\begin{aligned}\beta\mathbf{\Gamma}^{IZ}_{\tilde{\mathcal{Q}}^2} &\approx \ln \int d\mu(\mathbf{A}) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal\mathbf{A}_2\mathbf{T}]} e^{\frac{N}{2}\ln(\det(\mathbf{A}_1))} \\ &= \ln \left\langle e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal\mathbf{A}_2\mathbf{T}]} e^{\frac{N}{2}\ln(\det(\mathbf{A}_1))} \right\rangle_{\mathbf{A}}\end{aligned} \tag{116}$$

where the latter expresses the former in Bra-Ket notation. This expression is derived in Appendix A.4 (see Eqn. 245): it contains the original overlap term that depends $\mathbf{A}_2$ as well as a new term from the transformation that depends on $\det(\mathbf{A}_1)$ that is not yet defined.

5.2.3 The Effective Correlation Space (ECS)

Currently, Eqn. 116 is a "formal" expression, and we have not identified and/or justified its realm of applicability. Fortunately, we have empirical evidence to suggest that this can be made "physically" meaningful, by "restricting" the integral to the tail of the ESD.

Prior work on HTSR theory indicates that the generalizing parts of a layer weight matrix concentrate into $\rho^{emp}_{tail}(\lambda)$, the tail of the ESD, as the ESD becomes more Power Law (PL), and layer PL exponent $\alpha \to 2$ (from above) [62, 63, 25, 26, 28]. This suggests the integral in Eqn. 9 should instead average over a low rank subspace spanned *only by* the generalizing eigen-components of the student layer correlation matrix, $\mathbf{A}_1 = \frac{1}{N}\mathbf{S}^\intercal\mathbf{S}$ (or, equivalently, $\mathbf{A}_2$). We call this subspace the Effective Correlation Space (ECS).

Let $\tilde{\mathbf{A}}$ be the matrix spanned by the largest $\tilde{M}$ eigen-components $\mathbf{A}$ (in either form, $\mathbf{A}_1$ or $\mathbf{A}_2$).

$$\tilde{\mathbf{A}} := \mathbf{P}_{ecs}\mathbf{A}, \;\; \mathbf{P}_{ecs} := \sum_{i=1}^{\tilde{M}} |\lambda_i\rangle\langle\lambda_i| \tag{117}$$

where $\mathbf{P}_{ecs}$ is the projection operator onto the subspace spanned by the eigenvector $|\lambda_i\rangle$ associated with the eigenvalue λ_i of $\mathbf{A}_1$ or, equivalently, $\mathbf{A}_2$. We denote the corresponding Student correlation matrices with tilde, as follows:

$$\begin{aligned}\mathbf{A}_1 \to \tilde{\mathbf{A}}_1 \;\text{ such that }\; d\mu(\mathbf{A}_1) \to d\mu(\tilde{\mathbf{A}}_1) \\ \mathbf{A}_2 \to \tilde{\mathbf{A}}_2 \;\text{ such that }\; d\mu(\mathbf{A}_2) \to d\mu(\tilde{\mathbf{A}}_2)\end{aligned} \tag{118}$$

where now the matrices $\tilde{\mathbf{A}}_1$ and $\tilde{\mathbf{A}}_2$ are restricted to the ECS, and the measure $d\mu(\tilde{\mathbf{A}}_1)$ is similarly restricted. See Appendix A.4 for a more detailed discussion of this.

When an ESD is *Fat-Tailed* the ECS is at least as large if not larger than the PL tail of the ESD. For an ESD with $\alpha \geq 2$, The generalizing eigen-components, i.e.$|\lambda_i\rangle$, will mostly concentrate into the tail of the ESD, but some will remain in the bulk, so the ECS will be larger than the tail. In this case that $\tilde{M} \geq M^{tail}$, and $\rho_{ECS}(\lambda) \supseteq \rho_{tail}(\lambda)$, as depicted in Figure XXX. When $\alpha = 2$, the layer is Ideal (as in Section 3.1), in that all of the generalizing eigen-components to have now concentrated completely into the tail, so that $\tilde{M} = M^{tail}$, and $\rho_{ECS}(\lambda) = \rho_{tail}(\lambda)$. (When $\alpha < 2$,

this suggests that the layer is overfit, and the layer may have a Correlation Trap and/or frequently also has many near-zero eigenvalues.)

This leads to the following Model Selection Rule (MSR) for the ECS:

> When transforming the measure $d\mu(\mathbf{S}) \to d\mu(\tilde{\mathbf{A}})$, we invoke an eigenvalue cutoff rule that prescribes how to replace $\mathbf{A}$ with a low-rank effective matrix $\mathbf{A} \to \tilde{\mathbf{A}}$, where the cutoff $\tilde{\lambda} \geq \lambda_{min}^{ECS}$ is chosen so that the ECS at least contains the PL tail and, importantly, such that $\det(\mathbf{A}) = \det(\mathbf{A}_1) = \det(\mathbf{A}_2)$ is well defined.

Formally, this means that when we evaluate the Quality (squared) $\bar{\mathcal{Q}}^2$, Generating Function $\beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2}^{IZ}$ or other relevant averages, we restrict the measure (i.e., integral or sum) to the eigencomponents in the tail of the ESD of $\mathbf{A}$ (or $\mathbf{X}$, when appropriate) starting with λ_{min}^{ECS}. To our knowledge, this proposed MSR is completely novel.[30]

Restricted to the ECS, we now replace Eqn. 116 with:

$$\begin{aligned}\beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2}^{IZ} &= \ln \int d\mu(\tilde{\mathbf{A}}) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal \tilde{\mathbf{A}}_2 \mathbf{T}]} e^{\frac{N}{2}\ln(\det(\tilde{\mathbf{A}}_1))} \\ &= \ln \left\langle e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal \tilde{\mathbf{A}}_2 \mathbf{T}]} e^{\frac{N}{2}\ln(\det(\tilde{\mathbf{A}}_1))} \right\rangle_{\tilde{\mathbf{A}}} \end{aligned} \tag{119}$$

where we have used the formal Duality of Measures, $d\mu(\tilde{\mathbf{A}}) = d\mu(\tilde{\mathbf{A}}_1) = d\mu(\tilde{\mathbf{A}}_2)$. [31]

5.2.4 Two Simplifying Assumptions: the IFA and TRACE-LOG Condition

It now remains how to define the cutoff for the ECS space. To this, we make the following assumptions.

- **The Independent Fluctuation Assumption (IFA).** This condition states that the two terms appearing in the exponential of Eqn. 119 are statistically independent:

$$\beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2}^{IZ} \approx \ln \left\langle e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal \tilde{\mathbf{A}}_2 \mathbf{T}]} \right\rangle_{\tilde{\mathbf{A}}} \left\langle e^{\frac{N}{2}\ln(\det(\tilde{\mathbf{A}}_1))} \right\rangle_{\tilde{\mathbf{A}}}. \tag{120}$$

- **The Trace Log Condition (TRACE-LOG).** This condition states that the determinant of the Student (and Teacher) Correlation matrix is unity, such that:

$$\det\left(\tilde{\mathbf{A}}\right) = 1 \quad \text{or} \quad \mathrm{Tr}\left[\ln \tilde{\mathbf{A}}\right] = 0, \tag{121}$$

  (and with $\mathbf{A}$ additionally normalized to M, as explained in the Appendix, Section A.2.1) so that when we replace the measure over Student layer weight matrices $d\mu(\mathbf{S})$ with a measure over all Student Correlation matrices $d\mu(\mathbf{A})$, *restricted to the ECS*, the second term in Eqn. 120 becomes unity $\langle \exp\left[\frac{N}{2}\mathrm{Tr}\left[\ln[det\ \mathbf{A}_1]\right]\right]\rangle_{\tilde{\mathbf{A}}} = 1$, and vanishes in the final expression for $\beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2}^{IZ}$

At this point, the IFA is made purely for mathematical convenience, i.e., we have not demonstrated it empirically, but it is not implausible as a statistical modeling assumption. On the other hand, we can test the TRACE-LOG condtion empirically; this is a critical part of our SETOL approach.

[30] It is, however, similar in spirit to cut-offs used in some Quantum Field Theories to make the theories physically meaningful.

[31] Also, while we could replace $\mathbf{T} \to \tilde{\mathbf{T}}$, but to simplify the notation we do not do this.

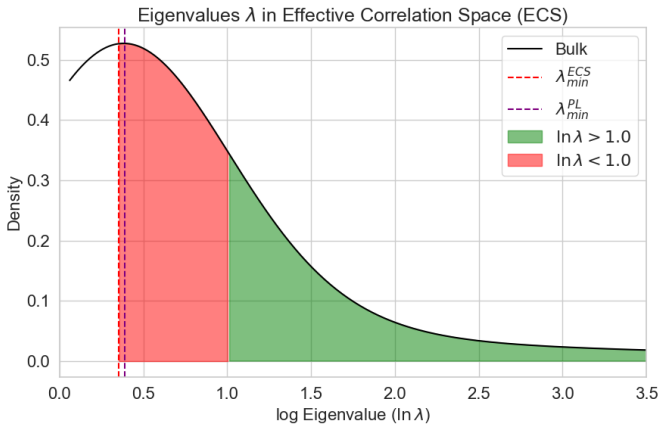


Figure 10: The image depicts a typical Empirical Spectral Density (ESD) of a layer correlation matrix $\mathbf{X}$, with `WeightWatcher` Power-Law (PL) exponent $\alpha = 2.0$.The green and red shaded regions depict eigenvalues λ in the Effective Correlation Space (`ECS`) of $\tilde{\mathbf{X}}$, defined by $\lambda > \lambda_{min}^{ECS}$. The x-axis displays the eigenvalues on the log scale, $\ln \lambda$. The vertical red line is at the start of the PL tail (λ_{min}^{PL}). The purple, vertical line is at the start of the `ECS` tail (λ_{min}^{ECS}). The green shaded region depicts those eigenvalues where $\ln \lambda > 1.0$, whereas the red shaded region depicts those eigenvalues where $\ln \lambda < 1.0$. The `ECS` is defined such that the volume-preserving `TRACE-LOG` condition is best satisfied, i.e $\sum \ln \lambda = 0$ for $\lambda \geq \lambda_{min}^{ECS}$.

Notably, when taking the large-N approximation of $\beta \mathbf{\Gamma}_{\mathcal{Q}^2}^{IZ}$, we effectively do this independently in two steps. First, we apply a Saddle Point Approximation (SPA) to the second term in Eqn. 120, leading to the `TRACE-LOG` condition (see Appendix A.4). Most importantly, we can test the `TRACE-LOG` condtion empirically, and this is another critical part and justification of our `SETOL` approach. Second, when applying the result by Tanaka (Eqn. 72), we are applying an SPA to the result, but assuming we are restricted to the `ECS`.

Empirical Tests of the `TRACE-LOG` Condtion and the `ECS` Since we require that the students resemble the fixed Teacher, a reasonable estimate for the average of the Student Correlation matrix $\tilde{\mathbf{A}}$ (either $\tilde{\mathbf{A}}_1$ or $\tilde{\mathbf{A}}_2$) (in the `ECS`) is the point estimate provided by the actual (known) fixed Teacher correlation matrix $\tilde{\mathbf{X}}$, so that

$$\left\langle \det \tilde{\mathbf{A}}_1 \right\rangle_{\tilde{\mathbf{A}}} = \left\langle \det \tilde{\mathbf{A}}_2 \right\rangle_{\tilde{\mathbf{A}}} \simeq \det \tilde{\mathbf{X}} = \prod_t \lambda_t = 1 \quad \forall \lambda_t \in \rho_{tail}^{emp}(\lambda), \tag{122}$$

for all eigenvalues λ_t in $\rho_{tail}^{emp}(\lambda)$ the tail of the ESD of $\tilde{\mathbf{X}}$, i.e., $\lambda_t \geq \lambda_{min}^{ECS}$. How should one choose λ_{min}^{ECS} in this expression? We already know that most NNs layers have Fat-Tailed ESDs, but if the Teacher ESD is simply *Random-Like* or even Bulk+Spikes, then the expected value $\langle \det(\mathbf{A}_1) \rangle$ itself of the determinant of a random full rank Student Correlation matrix $\mathbf{A}_1$ might be well defined and easy to estimate, and we might not even need to define the lower rank `ECS`. Indeed, in these cases, the correlated eigencomponents, if they exist, may be buried in the bulk region of the ESD and not readily identifiable. But because $\rho_{tail}^{emp}(\lambda)$ is Fat-Tailed and Power Law (PL), this poses some difficulty, which is why need to first define the Effective Correlation Space (`ECS`).

Because the Teacher ESD is most likely MHT and PL, if we choose λ_{min}^{ECS} too small, and the tail extends too far into the bulk region of the ESD, then for all practical purpose $\det(\tilde{\mathbf{X}}) \ll 1$. On the other hand, if we choose λ_{min}^{ECS} too large, then we only capture very large eigenvalues, and for

all practical purposes $\det\left(\tilde{\mathbf{X}}\right) \gg 1$. Therefore, if we set the scale of $\tilde{\mathbf{X}}$ appropriately, we can choose a λ_{min}^{ECS} such that $\det\left(\tilde{\mathbf{X}}\right) = 1$. In this case, by choosing λ_{min}^{ECS} appropriately, we can estimate the expected value of $\langle \det\left(\mathbf{A}_1\right)\rangle$ with a empirical point estimate over the Teacher Correlation matrix, which is unity.

$$|\det \tilde{\mathbf{X}}| \simeq 1; \quad \mathrm{Tr}\left[\ln \tilde{\mathbf{X}}\right] = \ln|\det \tilde{\mathbf{X}}| \simeq 0. \tag{123}$$

This expression now be used in a practical calculation to define a low-rank subspace that both allows us to evaluate the HCIZ integral, and to identify, in principle, generalizing components of the layer. We also refer to Eqn. 123 the `TRACE-LOG` condition, which is technicaly its empirical form. For example, Figure 5.2.4 depicts the eigenvalues in the `ECS` for a typical ESD with PL $\alpha = 2.0$, (normalized such that $\mathrm{Tr}\left[\mathbf{X}\right] = \|\mathbf{W}\|_F = N$). Notice that the start of the PL tail, (λ_{min}^{PL}), is very close to the start of the `ECS` tail. $(\lambda_{min}^{ECS}$, i.e. $\Delta\lambda_{min} := \lambda_{min}^{ECS} - \lambda_{min}^{PL} \approx 0$. Also, notice that while there are many large eigenvalues, $\ln\lambda > 1.0$, there are numerous small eigenvalues as well, $\ln\lambda < 1.0$, such that the $\sum \ln\lambda \approx 0$ for $\lambda \geq \lambda_{min}^{ECS}$. Additional plots like Figure 5.2.4, generated with `WeightWatcher`, can be found in Section 6, in Figure 21 as well as plots of $\Delta\lambda_{min}$ vs. the `WeightWatcher` PL α for several real-world examples, in Figures 22 and 23.

5.3 Evaluating the Layer Quality ($\bar{\mathcal{Q}}$) in the Large-N Limit

To generate the Average Quality, $\bar{\mathcal{Q}}^2$, we first take the large-N limit of $\beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2}^{IZ}$

$$\beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2,N\gg 1}^{IZ} := \lim_{N\gg 1} \beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2}^{IZ} = \lim_{N\gg 1} \ln\left\langle \exp N\beta\,\mathrm{Tr}\left[\tfrac{1}{N}\mathbf{T}^\top \tilde{\mathbf{A}}_2\,\mathbf{T}\right]\right\rangle_{\tilde{\mathbf{A}}} \tag{124}$$

and then take the appropriate partial derivative , analogously to as we did for $\bar{\mathcal{E}}_{gen}^{ST}$; see Section A.3 for more details. This gives (as in Eqn. 64)

$$\bar{\mathcal{Q}}^2 := \frac{1}{\beta}\frac{\partial}{\partial N}\beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2,N\gg 1}^{IZ} \tag{125}$$

$$\underset{\text{high-}T}{\approx} \frac{1}{N}\frac{\partial}{\partial \beta}\beta\mathbf{\Gamma}_{\bar{\mathcal{Q}}^2,N\gg 1}^{IZ} \tag{126}$$

Notice that since we are at high-Temperature, it doesn't matter which partial derivative we take, and we expect both results to be yield the same expression.

This HCIZ integral in Eqn. 124 can be evaluated (i.e in the large-N limit) using a result by Tanaka —provided the matrix $\tilde{\mathbf{A}}_2$ is low rank, which holds when the `TRACE-LOG` condition is satisfied. Thus, moving forward, we will assume an Effective Correlation Space (`ECS`) of rank $\tilde{M}$, where λ_{min}^{ECS} is the M^{th}-largest eigenvalue of $\tilde{\mathbf{X}}$, and defines the start of the `ECS` (and whatever branchcut is necssary to integrate $R(z)$).

Tanaka's result for the `ECS` can be expressed as:

$$\lim_{N\gg 1}\frac{1}{N}\ln\left\langle \exp\left(\beta\,\mathrm{Tr}\left[\mathbf{T}^\top\tilde{\mathbf{A}}_2\mathbf{T}\right]\right)\right\rangle_{\tilde{\mathbf{A}}} = \beta\sum_{i=1}^{\tilde{M}}\mathcal{G}(\tilde{\lambda}_i), \tag{127}$$

where the sum now only includes the eigenvalues of $\tilde{\mathbf{X}}$ (in the `ECS`), $\beta = \frac{1}{T}$ is the Inverse-Temperature, and $\tilde{\lambda}$ is an eigenvalue of $\tilde{\mathbf{X}}$, the Teacher Correlation matrix projected into the `ECS` space. $\tilde{\mathbf{A}}_2$ is the $N\times N$ form of the Student Correlation matrix, with $N-M$ non-zero eigenvalues, and $\mathbf{T}$ is the $N\times M$ Teacher weight matrix (also effectively projected into the `ECS`, i.e. $\mathbf{T} = \tilde{\mathbf{T}}$ here). $\mathcal{G}(\lambda_i)$ is

the Norm Generating Function, defined below. [32] This gives

$$\beta \mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2, N \gg 1} = N\beta \sum_{i=1}^{\tilde{M}} \mathcal{G}(\tilde{\lambda}_i), \tag{128}$$

This gives a final expression for the Average Layer Quality (Squared) $\bar{\mathcal{Q}}^2$ as

$$\bar{\mathcal{Q}}^2 = \sum_{i=1}^{\tilde{M}} \mathcal{G}(\tilde{\lambda}_i), \tag{129}$$

Note that $\bar{\mathcal{Q}}^2$ is independent of N and β, and, indeed, Eqn. 125 is an equality.

The average Quality (squared) can be expressed as a sum over Generating Functions $\mathcal{G}(\lambda)$, which depend only the statistical properties of the actual Teacher Correlation matrix $\tilde{\mathbf{X}}$ (projected into the ECS). Each term in the sum, $\mathcal{G}(\tilde{\lambda}_i)$, takes the form

$$\mathcal{G}(\lambda) := \int_0^{\lambda} R_{\mathbf{A}}(z)dz \xrightarrow{\texttt{ECS}} \int_{\lambda_{min}^{ECS}}^{\lambda} R_{\tilde{\mathbf{A}}}(z)dz \tag{130}$$

where $R_{\tilde{\mathbf{A}}}(\tilde{\lambda})$ is the R-transform from RMT, and λ_{min}^{ECS} is the lower bound of the ECS spectrum. Importantly, the R-transform for a Heavy-Tailed ESD may have a branchcut at or near the start of the ECS (as explained in Section 5.4), so restricting the integrand to start at λ_{min}^{ECS} is critical.

Since we expect the best Students matrices to resemble the actual Teacher matrices, we expect the Student correlation matrix $\tilde{\mathbf{A}}$ to have similar spectral properties to our actual empirical correlation matrices $\tilde{\mathbf{X}}$. That is, from the perspective of HTSR theory and the classification into 5+1 Phases of Training [25], we expect all the $\tilde{\mathbf{A}}$ to be in the same phase as $\tilde{\mathbf{X}}$ (and, in addition, to have the same PL exponent value). That is,

We expect the R-transform of $\tilde{\mathbf{A}}$ to have the same functional form as the R-transform of $\tilde{\mathbf{X}}$.

If our (Teacher) NN weight matrix exhibits a HT PL, then the tail the ESD ($\rho_{tail}(\lambda)$) of the Student and Teacher will both take the limiting form of a PL, with the same empirical variance σ^2 and (critically) the same PL exponent α:

$$\rho_{tail}[\tilde{\mathbf{A}}](\lambda) \sim \rho_{tail}[\tilde{\mathbf{X}}](\lambda) \sim \lambda^{-\alpha}. \tag{131}$$

Up until this point, our derivation of $\bar{\mathcal{Q}}^2$ only depends on the TRACE-LOG condition, irrespective of the exact functional form of $R(z)$, therefore the SETOL approach tested by examining how well the TRACE-LOG condition holds for the layers in very well performing models. We do this in Section 6.3.

5.4 Modeling the R-Transform

In this section, we explain how to select the R-transform $R(z)$ and evaluate the Norm Generating Function $\mathcal{G}(\lambda)$ under different modeling assumptions. To apply SETOL, the model satisfy the TRACE-LOG condition–which occurs during the case of Ideal Learning. For most cases of NN models, the ESD are HT; and this, in practice, one usually would select $R(x)$ that reflects this. But to be

[32] We use the notation $\langle \cdots \rangle_{\tilde{\mathbf{A}}}$ for expected value and placed $\frac{1}{N}$ on the L.H.S. to help the reader compare this to the original expressions in [82, 83]. Also, in [82, 83],$\beta = 1|2$, but, in fact, if one inserts $-\beta$ as an inverse-Temperature into the final expression, it simply factors out.

more general, we formally extend the theory to allow the practioners to both model the ESD as just a collection of discrete spikes, and to even correct for Correlation Traps and/or Very Heavy-Tailed (VHT) ESDs. Most importantly, we derive expressions for both the `WeightWatcher Alpha` and `AlphaHat` metrics, valid for the case Ideal Learning where $\alpha = 2$, and formally extended for other cases.

The goal of this section is to obtain formal expressions for the Layer Quality, $\bar{\mathcal{Q}}$ which is given in terms of what we somewhat imprecisely refer to as a Norm Generating Function. In many cases, however, the resulting approximate expressions for $\bar{\mathcal{Q}}$ do take the form well-known norms, including the Frobenius norm, the Spectral norm, and the Shatten norm. The models for $R(z)$ we use are presented in Table 8 ,and the final expression for the Layer Quality $\bar{\mathcal{Q}}$ for each model are given in Table 9.

5.4.1 Elementary Random Matrix Theory

We begin with some useful notions definitions from Random Matrix Theory. Using the ESD $\rho(\lambda)$, defined as

$$\rho(\lambda) := \frac{1}{N}\sum_i \delta(\lambda - \lambda_i), \tag{132}$$

we can express the *Greens Function* (or *Cauchy-Stieltjes* transform $C(z)$) by[33]

$$G(z) = C(z) := \int \mathrm{d}\lambda \frac{\rho(\lambda)}{z - \lambda}. \tag{133}$$

From $G(z)$, we can recover the ESD, $\rho(\lambda)$, using the inversion relation

$$\rho(\lambda) = \lim_{\epsilon \to 0+} \frac{1}{\pi}\mathrm{IM}(G(\lambda + i\epsilon)), \tag{134}$$

where IM is the imaginary part of $G(z)$, and where the $\lim_{\epsilon \to 0+}$ means to take the limit approaching from the upper half of the complex plane. The R-transform, $R(z)$, can be defined using the Blue function $B(z)$

$$R(z) := B(z) - \frac{1}{z}, \tag{135}$$

where the Blue function $B(z)$ [38] is the functional inverse of the Greens Function $G(z)$,[34] satisfying

$$B[G(z)] = z. \tag{136}$$

By specifying the $R(z)$ transform, we specify the complete ESD, $\rho(\lambda)$. Here, we are actually only interested in the tail of $\rho(\lambda)$. That is, we can given $R(z) = R(z)_{tail} + R(z)_{bulk}$, we only need $R(z) \approx R(z)_{tail}$.

5.4.2 Known R-transforms and Analytic (Formal) Models

There only a few known analytic results for the explicit R-transform $R(z)$. The ones we need are in Table 8. Below, we review some of them, explaining what ESD they correspond to, and what the resulting Norm Generating Function $G(\lambda)$ would be if applied as a model $R(x)$ here. 1;95;0c

[33] Please notice our naming and sign convention in Eqn. 133. We equate the Greens Function $G(z)$ with the (positive)Cauchy-Stieltjes transform, $G(z) = C(z)$. Other works may use the opposite sign convention, $G(z) = -C(z)$.

[34] The Blue function was first introduced by Zee [38] to model, among other things, spectral broadening in quantum systems. Briefly, given a deterministic Hamiltonian matrix $\mathbf{H}_0$, with eigenvalues λ_i^0, one can model the spectral broadening of λ_i^0 by adding a random matrix $\mathbf{H}_1$ to $\mathbf{H}_0$: $\mathbf{H} = \mathbf{H}_0 + \mathbf{H}_1$. The resulting eigenvalues of $\mathbf{H}$ now contain some level of randomness, σ, i.e., $\lambda = \lambda^0 + \sigma$. To model the ESD of $\mathbf{H}$, one then specifies the invididual R-transforms for $\mathbf{H}_0$ and $\mathbf{H}_1$; the full ESD of $\mathbf{H}$ can then be reconstructed by adding the two R-transforms together $R(z) = R_0(z) + R_1(z)$. Zee also notes that $R(z)$ is the same as the self-energy $\Sigma(z)$ from quantum many body theory [38].

Model	**HTSR Universality Class**	$R(z)$
Discrete	Bulk+Spikes, MHT, HT	$\frac{1}{\tilde{M}}\sum_{i=1}^{\tilde{M}}\lambda_i$
Wishart Models		
Multiplicative-Wishart	HT/VHT	$\frac{\epsilon\phi z^2}{2-\epsilon\phi^2 z^2}$
Inverse Wishart	HT/VHT	$\frac{\kappa-\sqrt{\kappa(\kappa-2z)}}{z}$
Lévy Wigner (LW)		
Free Cauchy (FC) ($\alpha_l=1$)	HT $\alpha=2$	$a+i\gamma$
General Lévy ($\alpha_l\neq 1$)	VHT $\alpha<2$	$a+bz^{\alpha-2}$

Table 8: Known R-transforms for random matrix ensembles relevant to modeling heavy-tailed spectral densities (eigenvalues or singular values squared). The *Multiplicative-Wishart* model has two real, non-zero parameters, ϵ and ϕ; for more details, see [**?**]. For the *Inverse Wishart*, as given by Bun [97], $\kappa=\frac{1}{2}(Q-1)$ where, $q=\frac{1}{Q}=\frac{M}{N}\leq 1$. The *Lévy-Wigner* (LW) model describes Wigner-like square random matrices (as opposed to Wishart-like or Correlation Matrices), where the elements are drawn from a Lévy-Stable distribution. The resulting LW ESD is Heavy-Tailed Power Law, and characterized by the Lévy exponent α_l. The LW $R(z)$ is parameterized by a (real) shift parameter a, a complex phase factor b (that depends on 3 real parameters α_l,β, and γ), and, of course, α_l. The Free Cauchy (FC) model is a special case of the IW model, corresponding to the Lévy $\alpha_l=1$, and the `HTSR` $\alpha=2$. We will extend the LW models to the rectangular case for our modeling purposes here by making the association $\alpha=\alpha_l+1$ for $\alpha\leq 2$. (Also, wwe take thew variance $\sigma=1$ for all models.) Generally speaking, the Lévy $R(z)$ are more complicated; for more details, see [98, 99, 100].

5.4.3 Discrete Model: Bulk+Spikes, MHT, HT

Here, we consider modeling the tail ESD, $\rho_{tail}(\lambda)$, as a collection of discrete spikes λ_{spike}, where $\lambda_{spike}\geq\lambda_{min}^{ECS}$. This could be for modeling an ESD in the Bulk+Spikes `HTSR` Universality class, or it could be applied to an MHT or HT ESD as a discrete distributoon with no inherent structure; its a modeling choice. Here, R-transform for the `ECS` is sum of Dirac delta functions. This lets us compute the Layer Quality $\bar{\mathcal{Q}}$ in closed form in terms of the Teacher weight matrix $\mathbf{T}=\mathbf{W}$ as a Tail norm, the Frobenius norm of the tail eigenvectors.

Let the tail of the ESD have $\tilde{M}=M^{tail}$ eigenvalues that define the `ECS`, i.e.,

$$\rho_{tail}(\lambda)=\sum_{i=1}^{\tilde{M}}\delta(\lambda-\tilde{\lambda}_i). \tag{137}$$

where the $\tilde{\lambda}_i$ are normalized by $\frac{1}{M}$.

The Greens Function $G(z)$ is then

$$G(z)=\int d\lambda\frac{\rho_{tail}(\lambda)}{z-\lambda}=\sum_{i=1}^{\tilde{M}}\int d\lambda\frac{\delta(\lambda-\tilde{\lambda}_i)}{z-\lambda}=\sum_{i=1}^{\tilde{M}}\frac{1}{z-\tilde{\lambda}_i}, \tag{138}$$

and the Blue function for each individual term i is $\frac{1}{z-\tilde{\lambda}_i}$, i.e., $B(z)=\tilde{\lambda}_i+\frac{1}{z}$. Now, using the additive property of the R-transform, we can express the total $R(z)$ as the sum of the R-transforms for the

individual terms i, giving

$$R(z) = \sum_{i=1}^{\tilde{M}} \left(\left(\tilde{\lambda}_i + \frac{1}{z} \right) - \frac{1}{z} \right) = \sum_{i=1}^{\tilde{M}} \tilde{\lambda}_i. \tag{139}$$

This gives the Norm Generating Function $\mathcal{G}(\lambda)$ as

$$\begin{aligned} \mathcal{G}(\lambda) &= \int_{\lambda_{min}^{ECS}}^{\lambda} \sum_{i=1}^{\tilde{M}} \tilde{\lambda}_i d\lambda \\ &= \sum_{i=1}^{\tilde{M}} \tilde{\lambda}_i \int_{\lambda_{min}^{ECS}}^{\lambda} 1 d\lambda \\ &= \left(\sum_{i=1}^{\tilde{M}} \tilde{\lambda}_i \right) \left(\lambda - \lambda_{min}^{ECS} \right) \end{aligned} \tag{140}$$

Seeing that λ_{min}^{ECS} is usually quite small, we make the approximation

$$\mathcal{G}(\lambda) \approx \left(\sum_{i=1}^{\tilde{M}} \tilde{\lambda}_i \right) (\lambda), \tag{141}$$

which gives the Quality-Squared approximately as

$$\bar{\mathcal{Q}}^2 = \sum_{i=1}^{\tilde{M}} \mathcal{G}(\tilde{\lambda}_i) \approx \left(\sum_{i=1}^{\tilde{M}} \tilde{\lambda}_i \right)^2. \tag{142}$$

We now see that we can define $\bar{\mathcal{Q}} := \sqrt{\bar{\mathcal{Q}}^2} = \sum_{i=1}^{\tilde{M}} \tilde{\lambda}_i$ is what we call a Tail Norm, the Frobenius norm of the tail eigenvalues.

5.4.4 Free Cauchy Model ($\alpha = 2$)

For the *Free Cauchy* (FC) model the R-transform is a constant

$$R(z)[\mathrm{FC}] = a + i\gamma, \tag{143}$$

where the shift parameter $a > 0$ translates the FC spectrum (i.e. so it overlaps with the the Power Law tail of the ESD), and the scale parameter $\gamma > 0$ sets the tail width. We do not attempt to fit a and γ to a real-world ESD here but rather simply use this model formally,

Because $R(z)$ is independent of z, the integral is straightforward:

$$\mathcal{G}(\lambda)[\mathrm{FC}] = \int_{\lambda_{min}^{ECS}}^{\lambda} R(z)[\mathrm{FC}] dz = (a + i\gamma)(\lambda - \lambda_{min}^{ECS}). \tag{144}$$

As above with the IW model, taking the modulus removes the phase factor in the $\mathcal{G}(\tilde{\lambda}_i)$, giving,

$$\left| \mathcal{G}(\lambda)[\mathrm{FC}] \right| = \sqrt{a^2 + \gamma^2} (\lambda - \lambda_{min}^{ECS}). \tag{145}$$

To obtain a formal expression for `Alpha`, we make two approximations. First, we take the lower cut-off to be near zero ($\lambda_{min}^{ECS} \approx 0$).

$$\left| \mathcal{G}(\lambda)[\mathrm{FC}] \right|_{\lambda_{min}^{ECS} \approx 0} \approx \lambda \sqrt{a^2 + \gamma^2} \tag{146}$$

Second, we assume the Layer Quality-Squared is dominated by the largest term in the sum:

$$\bar{\mathcal{Q}}^2_{\text{FC}} = \sum_{i=1}^{\tilde{M}} \left|\mathcal{G}(\lambda)[\text{FC}]\right|_{\lambda_{min}^{ECS} \approx 0} = \sum_{i}^{\tilde{M}} \tilde{\lambda}_i \sqrt{a^2 + \gamma^2} \approx \lambda_{max} \sqrt{a^2 + \gamma^2} \tag{147}$$

For the Heavy-Tailed ESDs, we genrally find that as λ_{max} increses, the `HTSR` α decreases. Juat for illustrative purposes, If we take $\log_{10} \lambda_{\text{max}} \sim \frac{1}{\alpha}$ (assuming $\log_{10} \lambda_{\text{max}} > 1$), and take the square root of both sides, we obtain the desired result for the log Quality, namely, that it scales inversely with α

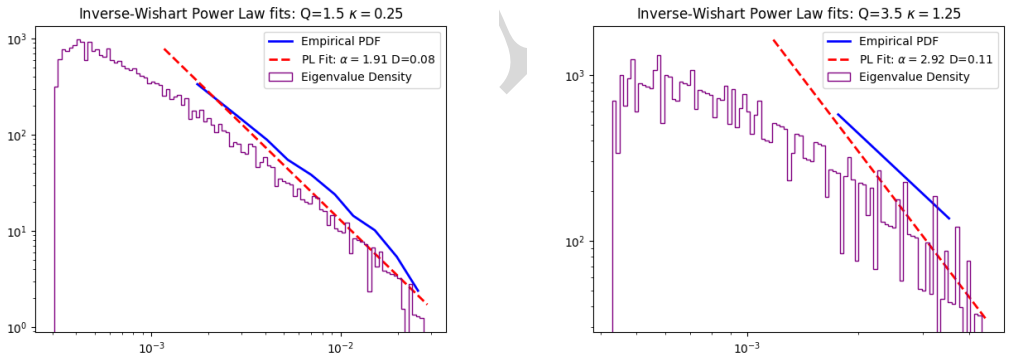
$$\log_{10} \bar{\mathcal{Q}}_{\text{FC}} \sim \frac{1}{\alpha} \tag{148}$$

If we extend this formally beyond the Ideal case to $\alpha \geq 2$, we now can exlain why the `HTSR` α makes a good Layer Quality metric in deep networks, as smaller α suggests higher quality layers and therefore a higher overall model quality and/or accuracy.

5.4.5 Inverse-Wishart Model of Ideal Learning

For a more realistic model of a layer ESD, here, we consider the Inverse-Wishart (IW) model. The IW model treats the ESD of $\mathbf{X}^{-1}$ when the ESD of $\mathbf{X}$ itself is MP. As a parametiric model, it can be quite effective at treating VHT and HT (or Fat-Tailed) ESDs when the far tail decays very rapidly, like a TPL, and/or for $\alpha \leq 4$. To do this, one simply considers the parameters κ as an adjustable paremeter. [I need to think this section through a bit more. I think we may need more than 1 adjustable parameter, and the branch cut may or may not matter depending on those choices] It is an excellent model for the ESD when $\alpha = 2.0$ (and $Q = 2$). Using this model, we can derive an expression for the `HTSR` `AlphaHat` Layer Quality metric, $\hat{\alpha} := \alpha \log_{10} \lambda_{max}$) as a leading order term in the final expression for $\log_{10} \bar{\mathcal{Q}}^2$.

This model

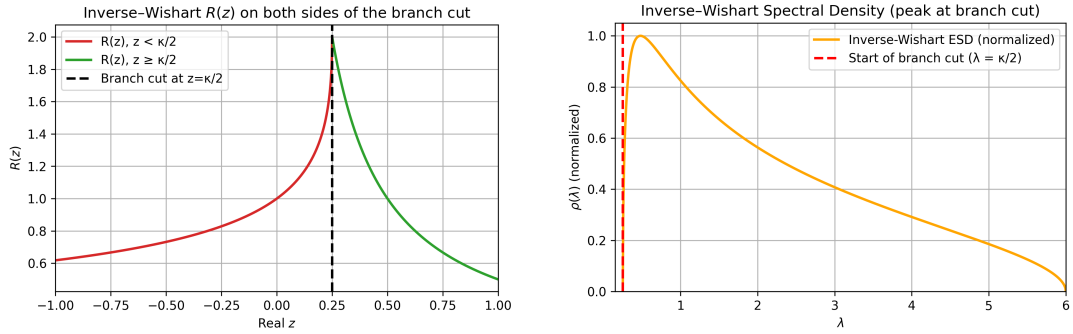


(a) IW Distribution, $\kappa = 0.25$

(b) IW Distribution, $\kappa = 1.25$

Figure 11: Example Inverse-Wishart (IW) distributions for $\kappa = 0.25$ and $\kappa = 1.25$, along with Power Law (PL) fits of the generated distribution. Plots on Log-Log scale.

In Figure 11, we fit some typical layer ESDs to an IW distribution. When $\kappa = 0.25$, the fitted $\alpha = 1.91$, and the fit is a reasonably accurate model of the underlying Power Law distribution and the `WeightWatcher` PL fit. For $\kappa = 1.25$, the fitted $\alpha = 2.92$ is larger, but the fit is not as good as a model. Generally speaking, α scales with κ, but the free cumulants scale inversely with κ. So smaller α will give larger free cumulants and therefore a larger $\bar{\mathcal{Q}}^2$. Importantly, as seen in



(a) IW $R(z)$, real z.

(b) Branch cut at $z = \kappa/2$

Figure 12: (a) The function $R(z)$ of the Inverse Wishart model, with a singularity at $z = \kappa/2$. (b) The branch cut in the empirical spectral density, corresponding to the tail for $\kappa = 0.5$.

Figure 11(a), for $\alpha = 2.0$, the IW model (with $\kappa = 0.25$) is an effective, simple model to illustrate the `SETOL` case of Ideal Learning.

Lets consider $R(z)$ for the Inverse-Wishart model, denoted $R(z)[IW]$. To integrate this function, we require that it be analytic. At first glance, it may seem that that $R(z)[IW]$ is not analytic because it has a pole at $z = 0$ and because the square-root term $\sqrt{\kappa(\kappa - 2z)}$ creates branch cut at and $z = \kappa/2$ (and $z = \infty$). Figure 12 presents this in two ways: Figure 12(a) shows the R-transform $R(z)[IZ]$ for real z, highlighting its singular behavior and the location of the branch cut at $z = \kappa/2$; and Figure 12(b) shows the corresponding branch cut in the ESD of the Inverse Wishart model (for $\kappa = 0.5$). We select the branch cut starting at $z = \kappa/2$ and ending at $z = \infty$, which allows us to at least formally defined the integral along the physically meaningful part of the ESD:

$$\mathcal{G}(\lambda)[\text{IW}] := \int_{\lambda_{min}^{ECS}}^{\lambda} R(z)[\text{IW}]dz, \tag{149}$$

noting that we expect $\lambda_{min}^{ECS} \geq \kappa/2$.

It turns out, however, that due to the branch cut in $R(z)[\text{IW}]$, the function $\mathcal{G}(\lambda)[\text{IW}]$ is not analytic in the domain we need. To correct for this, we will instead model the Layer Quality-Squared using the modulus of $\mathcal{G}(\lambda)[\text{IW}]$,

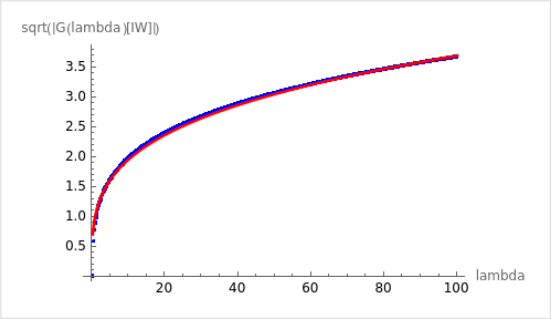
$$|\mathcal{G}(\lambda)[\text{IW}]| := \sqrt{\mathcal{G}(\lambda)[\text{IW}]^{*}\mathcal{G}(\lambda)[\text{IW}]} \tag{150}$$

where $\mathcal{G}(\lambda)[\text{IW}]^{*}$ is the complex conjugate of $\mathcal{G}(\lambda)[\text{IW}]$. This is somewhat involved, so we present the full calculation in Appendix A.7

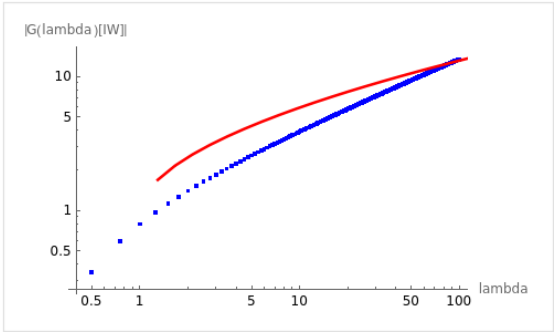
Figure 13 plots $|\mathcal{G}(\lambda)[\text{IW}]|$ on Lin-lin and Log-log plots, on the range $\lambda \in (0.25, 100)$, and fits the it to a PL. The fit follows the general trend of the function, but it is not terribly accurate. Still, from this plot, we can see that the Layer Quality-Squared has the same general trend as `AlphaHat` and/or a Shatten Norm.

5.4.6 The Multiplicative-Wishart (MW) model

The Multiplicative-Wishart (MW) model has two real, adjustable parameters ϵ, ϕ. It treates $\mathbf{X}$ as resulting from product of random matrices, and is good for modeling an ESD with a Very Heavy Tail (VHT). It has been used previously to model the heavy tail of the Hessian matrix in NNs[**?**] This model would work better for fitting HT ESDs that decay slower than a PL or TPL. We note that unlike the IW, the MW model does not have a branch cut at the ECS, but it does have 2



(a) $|\mathcal{G}(\tilde{\lambda}_i)[IW]|$ Lin-lin plot and PL fit



(b) $|\mathcal{G}(\tilde{\lambda}_i)[IW]|$ Log-log plot and PL fit

Figure 13: Behavior of $\mathcal{G}(\lambda)[\text{IW}]$ for the Inverse Wishart (IW) model, with a Power Law (PL) fit (red), $|\mathcal{G}(\lambda)[\text{IW}]| \approx 1.138\lambda^{0.539}$. (a) Lin-lin plot. (b) Log-log plot.

poles, compicating the integration of the R-transform. We will not consider this model further here and instead leave this for future work.

5.4.7 Levy-Wigner Models and the `AlphaHat` Metric

eHere, we consider General Levy-Wigner (LW) model.. We show how to obtain the `WeightWatcher AlphaHat` metric by modeling VHT ESDs with an approximation to a Levy distribution, suitable for cases where the `HTSR` $\alpha \leq 2$.

The `AlphaHat` metric has been developed to adjust for *Scale* anomalies that arise from issues like Correlation Traps, rank collapse, and overfitting, which can in many cases make `Alpha` smaller than expected, and even smaller than 2. We model these VHT ESDs *as if* they follows a Levy-Stable distribution, generated from a Levy-Wigner random matrix. The Levy-Wigner (LW) model treats $\mathbf{X}$ as if it were a Wigner matrix (and not actually a rectnalgyle correlation matrix), s we must decide how to map the `HTSR` α to the Lévy α_l; for this, we select $\alpha_l = \alpha - 1$. (See in Table 8.) The LW model is defined in terms of 3 paremeters: a is a shift parameter, and b is a complex phase factor depending on 2 real factors, β and γ. Strictly the ESD for an LW model, $\rho_{LW}(\lambda)$, is defined by its characteristic function (i.e., the Fourier Transform of $\rho_{LW}(\lambda)$), but we when the ESD is VHT, $\rho_{LW}(\lambda) \sim \lambda^{-\alpha_l - 1}$, and when $\alpha_l \approx 1$, the ESD resembles a PL HT ESD with the `HTSR` $\alpha \approx 2$.

Let us model the R-transform of our Very Heavy-Tailed (VHT) ESDs as

$$\begin{aligned} R(z)[\text{LW}] &= bz^{\alpha_l - 1}, \ \alpha_l \in (0,1) \\ &= bz^{\alpha-2}, \ \alpha \in (0,2), \end{aligned} \tag{151}$$

where b is an unspecified constant (possibly negative and/or complex). This is not a particualry good model for VHT ESDs in practice; it is simply an approximate model we choose to obtain a formal expresion.

Integrating $R(z)[\text{LW}]$, and (as above) taking the approximation $\lambda_{min}^{ECS} \approx 0$, we obtain

$$\mathcal{G}(\lambda)[\text{LW}] \approx \tfrac{b}{\alpha-2}\lambda^{\alpha-2}. \tag{152}$$

For simplicity, if we now choose $b = \alpha_l = \alpha - 2$, then $\bar{\mathcal{Q}}^2_{\text{LW}}$ takes the form of a Shatten Norm

$$\bar{\mathcal{Q}}^2_{\text{LW}} = \frac{1}{\tilde{M}} \sum_i^{\tilde{M}} \lambda^{\alpha-1}. \tag{153}$$

Taking the logarithm of $\bar{\mathcal{Q}}^2_{\text{HT}}$, we obtain

$$\log \bar{\mathcal{Q}}^2[\text{LW}] \approx (\alpha - 1) \log \lambda \tag{154}$$

As with the Free Cauchy (FC), let us approxmate $\bar{\mathcal{Q}}^2$ by the largest term in the sum over $\mathcal{G}(\tilde{\lambda}_i)$, and then let $\lambda = \lambda_{max}$, giving

$$\hat{\alpha} = \log_{10} \bar{\mathcal{Q}}^2_{\text{LW}} \approx (\alpha - 1) \log \lambda_{max}. \tag{155}$$

We present this as a formal example, noting that is slightly different from the result for the IW model, Eqn. **??**. We do not claim this is a valid empirical model, as we have not attempted to fit a real-world ESD to Levy-stable distribtion. We leave this to a future study, noting, however, there has been some work doing such fits [101].

Ideally, we would like to have an rigorous expression for $R(z)$ not just in the case of Ideal Learning but also for the entire Fat-Tailed Universality class. This is non-trivial to obtain and we will attempt this in a future work. Fow now, we will take a different approach, and evaluate $R(z)$ explicitly using numerical techniques.

Scaling insight. For the VHT Universality class, we may expect λ_{max} to be fairly large such that $\lambda_{max} > 1$ and therefore $\log_{10} \lambda_{max} > 0$ This implies that $\log_{10} \bar{\mathcal{Q}}^2$ decreases as α (and thus α_l) decreases, even though $\log_{10} \lambda_{max}$ simultaneously *increases* in the VHT class. This opposing interplay explains why correlation traps produce small α yet yield deceptively moderate quality scores.

Historical remark. An earlier study on `AlphaHat`,reported the opposite trend, with `AlphaHat` decreasing with increasing model quality. This is because we employed an un–normalised covariance, suggesting that $\lambda_{max} < 1$ and $\log_{10} \lambda_{max} < 0$ (although we have not rigorously checked this). The present M normalisation clarifies that the VHT regime is meaningful *only* when $\alpha < 2$, indicating the layer amd therefore the model is overfit (in some unspecified way).

5.4.8 Summary of models.

Table 9 consolidates the derived expressions for the log Layer Quality, $\log \bar{\mathcal{Q}}$, across multiple R–transform models, each capturing distinct heavy-tail characteristics of the ESD within the `SETOL` framework. The Discrete (Spikes) model yields a Frobenius-like tail norm by summing the effective eigenvalues in the `ECS`, directly tying quality to tail magnitude (cf. Eqn. 142). The Free Cauchy (FC) model links $\log \bar{\mathcal{Q}}$ inversely to the `HTSR` metric `Alpha`, demonstrating that smaller α values enhance quality (cf. Eqn. 148). The Lévy-Wigner (LW) model connects $\log \bar{\mathcal{Q}}$ to `AlphaHat` through the term $(\alpha - 1) \log \lambda_{\text{max}}$, revealing how very heavy tails depress layer quality (cf. Eqn. 155). The Inverse Wishart (IW) model defines the `ECS` boundary via a branch cut at $\kappa/2$ and approximates $\log \bar{\mathcal{Q}}$ as $0.5 \log \lambda_{\text{max}}$, providing a robust metric for Ideal Learning scenarios (cf. Figure 13). Each model thus offers a unique lens on layer quality, enabling direct mapping between theoretical R–transform assumptions and empirical `WeightWatcher` metrics.

Practical Implications for Neural Network Analysis These models empower researchers to tailor layer-quality assessments to specific `HTSR` universality classes, boosting the predictive power of `WeightWatcher` metrics like `Alpha` and `AlphaHat`. The Discrete model excels when the ESD exhibits clear eigenvalue spikes, delivering a sharp quality signal for Bulk+Spikes layers. The

[35] See Appendix A.7, Eq. (A.7.25): $|G(\lambda)| \approx 1.138\, \lambda^{0.539}$.

Model	Tail Norm	WW Metric	$\log \mathcal{Q}$ **(Log Quality)**
Bulk+Spikes (BS)	Frobenius Norm	N/A	$\log\left(\sum_{i=1}^{\tilde{M}} \tilde{\lambda}_i\right)$
Free Cauchy (FC)	Spectral Norm	`Alpha` α	$\log \lambda_{max} \sim \frac{1}{\alpha}$
Lévy Wigner (LW)	Shatten Norm	`AlphaHat` $\hat{\alpha}$	$(\alpha - 1) \log \lambda_{\max}$ + const.
Inverse Wishart (IW)[35]	ECS Boundary*	`Alpha` α	$0.5 \log \lambda_{\max}$ + const.

Table 9: Closed-form or leading-order expressions for the log Layer Quality $\log \mathcal{Q}$ derived from the integrated R–transform for each core tail-model, simplified to show the relation to the `WeightWatcher` `Alpha` and `AlphaHat` metrics. For each model, we interpret the final result. Bulk+Space (BS): Sums the $\tilde{\lambda}_i$ in the `ECS`, giving a Frobenius Tail norm. Free Cauchy (FC): Yields the Spectral Norm, λ_{max}. Since this scales as $\frac{1}{\alpha}$, this explaines the `HTSR` `Alpha` metric as it shows why smaller α yields higher $\mathcal{Q}$. Lévy Wigner (LW): Yields a Shatten Norm, which can be approximated by `AlphaHat`, $\hat{\alpha}$, Being for the VHT Universality class, it implies that heavier tails ($\alpha \in (1, 2)$) depress Layer Quality. Inverse Wishart (IW): Also gives `Alpha`, and, importably, a branch cut that defines the `ECS` Boundary *(not a Tail norm).

FC model supports Ideal Learning by linking quality directly to $1/\alpha$, which simplifies cross-layer comparisons. The LW model helps detect overfitting by adjusting for Correlation Traps via `AlphaHat` in very heavy-tailed regimes. Finally, the IW model offers a parametric fit for both HT and VHT tails, with κ tuning alignment to real-world ESDs (cf. Figure 11). By integrating these models, practitioners can compute precise layer-quality metrics that optimize both performance and interpretability.

5.5 Computational Random Matrix Analysis

The R-transform is the generating function for the *Free Cumulants* of RMT. Formally–and if we assume a model without a branch-cut or troublesome poles– one can define $R(z)$ as a series expansion in z,

$$R(z) := \kappa_1 + \kappa_2 z + \kappa_3 z^2 + \dots \tag{156}$$

where the coefficients κ_k are the free cumulants, which can be expressed in terms of the matrix moments m_k[**?**], defined (here) as

$$m_k := \mathrm{Tr}\left[\tilde{\mathbf{X}}^k\right] = \sum_{i=1}^{\tilde{M}} (\tilde{\lambda})^k \tag{157}$$

where $\tilde{\lambda}_k$ is the k-th eigenvalue of the effective correlation matrix $\tilde{\mathbf{X}}$, which has been mean-centered and normalized by its standard deviation.

The free cumulants are defined recursively as

$$\kappa_k := m_k - \sum_{\text{partitions of } n \text{ blocks } B} \prod m_{|B|} \tag{158}$$

The first 5 *Cumulants* are, explicitly,

$$
\begin{aligned}
k_1 &= m_1 \\
k_2 &= m_2 - m_1^2 \\
k_3 &= m_3 - 3m_2m_1 + 2m_1^3 \\
k_4 &= m_4 - 4m_3m_1 - 2m_2^2 + 10m_2m_1^2 - 5m_1^4 \\
k_5 &= m_5 - 5m_4m_1 + 15m_3m_1^2 + 15m_2^2m_1 - 35m_2m_1^3 - 5m_3m_2 + 14m_1^5
\end{aligned}
\tag{159}
$$

Using these definitions, we can estimate the Layer Quality matrix $\bar{\mathcal{Q}}^2$ for our experimental models (in Section 6 by computing $\mathcal{G}(\lambda)$ for the effective correlation space (i.e. the tail of the layer ESD), however it is defined. That is, we use

$$
\mathcal{G}(\lambda_i) = \kappa_1 \frac{\tilde{\lambda}}{\tilde{M}} + \frac{\kappa_2}{2}\left(\frac{\tilde{\lambda}}{\tilde{M}}\right)^2 + \cdots \tag{160}
$$

Note that we evaluate $\mathcal{G}(\lambda)$ using eigenvalues normalized with $\frac{1}{\tilde{M}}$. This is implemented implemented in the `WeightWatcher` package. In Section 6.4, we examine some examples of this approach to computing Layer Qualities directly.

6 Empirical Studies

In this section, we present empirical results. Our goals are to justify key technical claims, including key assumptions underlying our `SETOL` approach, and to illustrate the behavior of `SETOL` with respect to various parameters and hyperparameters. Importantly, it is *not* our goal to demonstrate that layer PL exponent `Alpha` and `AlphaHat` perform well for diagnostics and predicting model quality for SOTA NN models, as that has been demonstrated previously [26, 24, 27].

Since the `SETOL` theory presented in Section 5 is (effectively) a single layer theory, in order to carefully test (as opposed to simply use) `SETOL`, we need to limit the degree of inter-layer interactions present in the model. To do so, we consider a three-layer Multi-Layer Perceptron (MLP3), trained on MNIST [102]. We refer to the hidden layers as "FC1 and "FC2. Their output sizes and parameter counts are shown in Table 10.

Layer	Units	Weight Parameters	% of total
FC1	300	$768 \times 300 = 230,700$	88.2%
FC2	100	$300 \times 100 = 30,000$	11.4%
out	10	$10 \times 100 = 1000$	0.38%

Table 10: Dimensions of each FC layer in the MLP3 model, along with weight matrix parameter count and fraction of the total.

The following are the main topics we consider.

6.1 **Model Quality: `HTSR` phenomenology.** The `HTSR` phenomenology provides a metric of model quality in the form of the PL exponent α.[36] In particular, smaller values of α (e.g., values of α closer to 2 than 3 or 4) should correspond to better models, e.g., having smaller test errors; and a minimal error should be obtained when $\alpha = 2$. See Section 6.1.

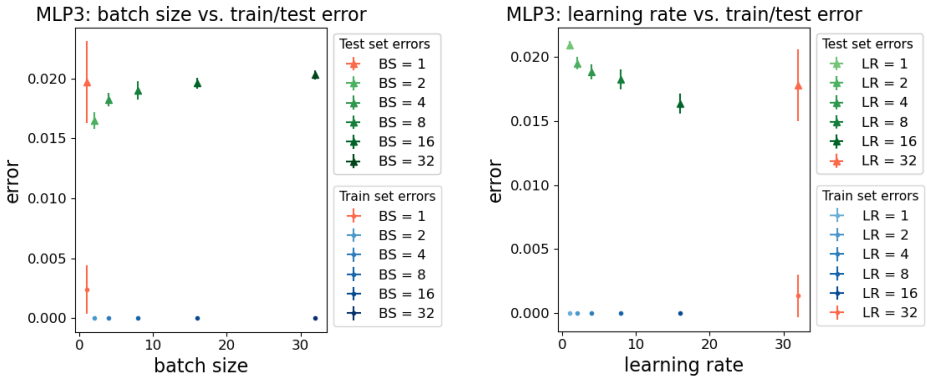
6.2 **Effective Correlation Space.** The `SETOL` theory is based on the notion of an Effective Correlation Space, in which the learning and generalization occurs. This is the low-rank subspace $\mathbf{W}^{ecs}$ of each layer $\mathbf{W}$ that approximates the teacher $\mathbf{T}$. In particular, our measure of model quality should be restricted to $\mathbf{W}^{ecs}$. The Effective Correlation Space can be identified from the tail of the ESD, $\rho_{tail}(\lambda)$, and it can be chosen according to one of several related Model Selection Rules. See Section 6.2.

6.3 **Evaluating the Trace-Log Condition.** In the `HTSR` phenomenology, when a model is very well-trained, the layer PL exponent $\alpha \simeq 2$. In the `SETOL` theory, when a model is very well-trained, the eigenvalues in the tail will satisfy the Empirical Trace-Log Condition, given in Eqn. (123). In Section 6.3, we provide a detailed analysis of this effect.

6.4 **Computational Model Qualities.** In Section 6.4, we empirically compare the HTSR layer-quality exponent α against the computational R-transform-derived metric $\bar{\mathcal{Q}}^2$ on fully trained MLP3 models. We show that for the FC1 layer, $\bar{\mathcal{Q}}^2$ closely tracks α with small error bars and the expected batch-size trend, whereas for FC2, $\bar{\mathcal{Q}}^2$ exhibits much larger variability—demonstrating that α provides a more stable and robust measure of layer quality.

6.5 **Correlation Traps.** Recall, from Section 3.3.1, that if a layer weight matrix $\mathbf{W}$ has a Correlation Trap (and, in particular, those arising from SGD training with very large learning

[36] Prior work has shown that the `AlphaHat` metric ($\hat{\alpha}$) accurately describes variations in model quality as a function of architecture changes [24]. Since we do not vary the depth of the model in our evaluations, the `Alpha` metric (α) is of interest in this setting.



(a) Batch size experiment (b) Learning rate experiment

Figure 14: Train / test errors in the MLP3 model as a function of batch size, and learning rate. Observe the inverse relationship between batch size (a) and learning rate (b). As batch size decreases, test error generally decreases, until batch size reached $bs = 1$. Similarly, as learning rate increases, test error decreases until $lr = 32\times$ the SGD default value of 0.01.

rates) then it is likely that the test (and train) accuracy will be degraded, and α will drop below its optimal value. See Section 6.5 for an empirical demonstration of this.

6.6 **Overloading and Hysteresis Effects.** The experiments described so far validate that `SETOL` makes valid predictions in the $\alpha \gtrsim 2$ range. Beyond that point, `SETOL` only predicts "atypicality, in the sense of spin glass theory [103]. See Section 6.6 for an examination of how the MLP3 behaves when it is pushed further out of that range of validity, e.g., by training only one layer, while keeping the others frozen. In particular, we compare results when a single layer is either over- vs under-parameterized.

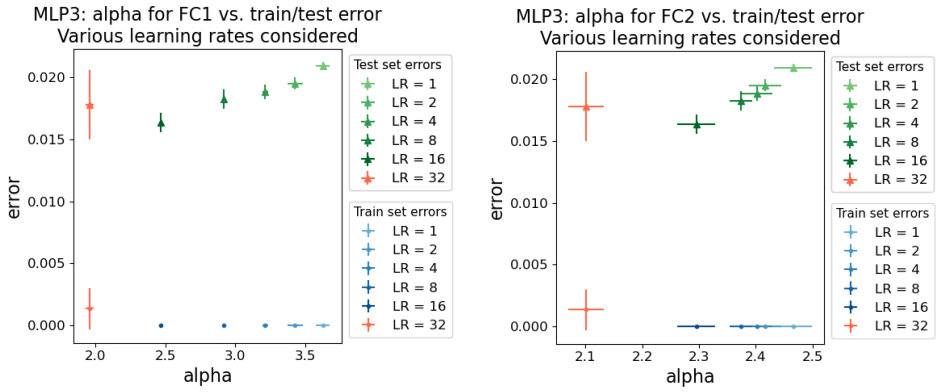
We trained the MLP3 model independently, using both the Tensorflow 2.0 framework (using the Keras api, and on Google Colab) and pytorch, with the goal of consistent, reproducible results. Each setting of batch size, learning rate, and trainable layer was trained with 5 separate starting random seeds, and error bars shown in plots below represent one standard deviation taken over these 5 random seeds. Each training run used the same early stopping criteria on the train loss: training was halted when train loss did not decrease by more than $\Delta E_{train} = 0.0001$, over a period of 3 epochs. In doing so, each model was trained with a different number of epochs; and, at the end, the best weights were chosen for the model. See Appendix A.5 for more details on the MLP3 model and the training setup. We provide a Google Colab notebook with the exact details, allowing the reader to reproduce the results as desired.

The dominant generalizing components of $\mathbf{W}$ reside in $\mathbf{W}^{ecs}$ such that it captures the functional contribution of $\mathbf{W}$ to the NN; and thus

6.1 `HTSR` Phenomenology: Predicting Model Quality via the `Alpha` metric

Here, we want to determine how the quality of our MLP3 model varies with the `Alpha` metric. From previous work [26, 24, 27], we expect that `Alpha` metrics for the FC1 and FC2 layers should be well-correlated with the test accuracy, while varying some suitable training knob, such as learning rate or batch size, that can modulate the test accuracy.[37]

[37] Since we do not change the depth of the model here, we expect the `Alpha` metric to follow the `AlphaHat` metric, also predicting the test accuracies [24].



(a) α_{FC1} (b) α_{FC2}

Figure 15: Train / test errors in the MLP3 model in the ***Learning Rate*** experiment as a function of α_{FC1} (a) and α_{FC2} (b). Observe the regular downward progression of α and error as the learning rate increased in both (a) and (b). When learning rate was 32×, (shown in red), α_{FC1} fell below 2, coinciding with a drastic increase in both train and test error. The results here almost perfectly replicate those of the Batch Size experiment, shown in Figure 16.

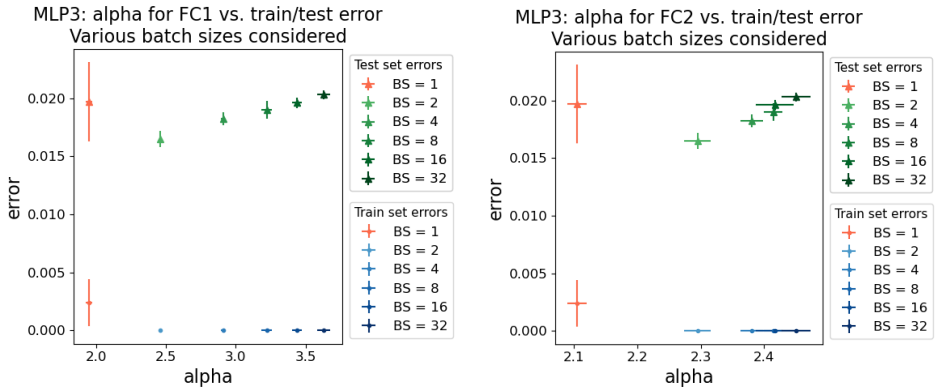
We vary the batch size from small to large, i.e., $bs \in [1, 2, 4, 8, 16, 32]$, following the setup of previous work on the `HTSR` phenomenology [25]. We expect similar effects by varying the learning rate, as it is known that small batch sizes correspond directly to large learning rates [104, 105]. Thus, we conducted a second set of experiments where the learning rate was varied by a factor of $[1\times, 2\times, 4\times, 8\times, 16\times, 32\times]$, relative to the SGD default value of 0.01. Adjusting the learning rate or batch size allows us to systematically vary the layer PL exponent α between roughly 2 and 4, i.e., within the range in which `SETOL` should make the most reliable predictions. As an added benefit, it also allows us to use the very small batch size of 1 to force the model into a state of over-regularization, which we also analyze below.

Consider Figure 14, which plots the final train and test accuracies as a function of the hyperparameter (batch size or learning rate) used during training for the MLP3 model. Figure 14(a) varies batch size, and Figure 14(b) varies learning rate. Recall that error bars represent one standard deviation taken over 5 independent starting random seeds. In Figure 14(a), we see that by decreasing the batch size (bs), and holding other knobs constant, we can systematically improve the train and test accuracy, up to a point. In particular, for $bs \geq 2$, both the test and train accuracies increase with decreasing batch size, consistent with previous work [25]. Further decrease beyond $bs = 2$ leads to *lower* model quality, i.e., larger error and larger error variability. In Figure 14, we see that increasing the learning rate (lr) by a factor x has an exactly analogous effect as decreasing the batch size by $1/x$.[38]

The transition between $lr = 16\times$ (or $bs = 2$), which is locally optimal for the setting of other hyperparameters, and $lr = 32\times$ normal (or $bs = 1$), which is not, provides a demonstration of a distinct change in the behavior of `Alpha`, concordant with the sudden increase in the error and error variability. To explore this in the context of `SETOL`, consider Figure 15 and Figure 16, which plot error as a function of `Alpha`, for different learning rates and batch sizes, respectively.

Figure 15 plots the `Alpha` metrics α_{FC1} and α_{FC2}, as learning rate is varied, demonstrating that both metrics are well-correlated with the test accuracies, for all learning rates less than 16× normal. In particular, as we drive `Alpha` in FC1 down to an Ideal value of $\alpha \simeq 2$, the test error

[38] One could, of course, mitigate this by fiddling with other knobs of the training process, but that is not our goal. Our goal here is not to use a toy model to demonstrate the properties and predictions of `SETOL`.



(a) α_{FC1} (b) α_{FC2}

Figure 16: Train / test errors in the MLP3 model in the **Batch Size** experiment as a function of α_{FC1} (a) and α_{FC2} (b). Observe the regular downward progression of α and error as the batch size decreased in both (a) and (b). When batch size was 1, (shown in red), α_{FC1} fell below 2, coinciding with a drastic increase in both train and test error. The results here almost perfectly replicate those of the Learning Rate experiment, shown in Figure 15.

decreases monotonically (Figure 15(a)). Beyond that point, further decrease of the batch size sees `Alpha` decrease below its Ideal value of 2 in the FC1 layer. This corresponds not only with *higher errors*, but also with *larger error bars*, on both train and test error. The dramatic increase in train error and train error variability is particularly telling, because it suggests that when α_{FC1} passes below 2, the model enters into a "glassy state, and is unable to relax down to 0 train error.

In Figure 15(b), we consider `Alpha` for FC2, and we see that α_{FC2} approaches 2, but does not reach it, even for $lr = 32\times$. This failure to achieve $\alpha_{FC2} \approx 2$, along with the much greater size of FC1, (See Table 10,) suggests that FC1 is the critical layer for the models performance. This also highlights some of the interplay between the layers, (which is not captured by a single layer theory) – as α_{FC1} has narrow error bars throughout, α_{FC2} shows much more variation by way of its wider error bars. Thus, while model accuracy kept improving as learning rate increased up to 16×, this was likely driven by a better α_{FC1}, more than by α_{FC2}.

In Figure 16, we consider batch size, and we see a near identical replication of these results, in terms of the relation of train error, test error and `Alpha` in the two layers. Consequently, the remainder of the experiments will focus on the learning rate experiment, as both produced substantially the same results.

6.2 Testing the Effective Correlation Space

Here, we will address the question:

How shall we *test the assumption* of the Effective Correlation Space?

Recall that the `SETOL` theory estimates model quality by evaluating the ST Generalization Error as an integral over the theoretical training data $\boldsymbol{\xi}$. This integral assumes each layer weight matrix can be replaced with an effectively lower rank form, i.e., $\mathbf{W} \rightarrow \mathbf{W}^{ecs}$, corresponding to the span of the eigencomponents defined by the tail of the ESD, $\rho_{tail}(\lambda)$. In the `HTSR` phenomenology, the tail is defined by the fact that $\rho_{tail}(\lambda)$ follows a PL distribution, above some minimal value λ_{min}. In our `SETOL` theory, the tail is defined by choosing the minimal value λ_{min} to satisfy the Empirical Trace-Log condition. These methods of realizing $\mathbf{W}^{ecs}$ are essentially *Model Selection*

Rules (MSRs) for the Effective Correlation Space. Importantly, in neither approach is λ_{min} just some "rank parameter to be chosen by yet some other MSR on the basis of rank, or magnitude alone, (that, in particular, does not know about `HTSR` or `SETOL`, which consider the *shape* of the ESD).

Thus, to test the assumption of the Effective Correlation Space, we want to show that the models predictions are in fact controlled predominantly by the tail, where the specific choice of the rank parameter depends on `HTSR` or `SETOL` as we expect. We can emulate this theoretical construct and estimate (trends in the) test accuracies by evaluating the train and/or test accuracies of the trained MLP3 model – after replacing the MLP3 layer weight matrices $\mathbf{W}_{FC1}$ and $\mathbf{W}_{FC2}$ with a low-rank approximation consisting of *only* the tail:

$$\mathbf{W}_{FC1}^{ecs} \coloneqq P_{tail}\mathbf{W}_{FC1}$$
$$\mathbf{W}_{FC2}^{ecs} \coloneqq P_{tail}\mathbf{W}_{FC2},$$

where P_{tail} is a projection operator selecting only the tail of the ESD with TrucnatedSVD. (That is, we will use the low-rank TruncatedSVD approximation at the inference step, not at the training step, as is more common.) A Truncated model is one whose weight matrices $\mathbf{W}_*$ have been replaced by truncated matrices $\mathbf{W}_*^{ecs}$. We denote the difference between the original models accuracy and the Truncated models train and test accuracy as ΔE_{train} and ΔE_{test}, respectively:

$$\Delta E_{train} \coloneqq E_{train}(\mathcal{D}) - E_{train}^{ecs}(\mathcal{D})$$
$$\Delta E_{test} \coloneqq E_{test}(\mathcal{D}) - E_{test}^{ecs}(\mathcal{D}),$$

where E_{train}^{ecs} denotes the error of the TruncatedSVD model on the training portion of the dataset $\mathcal{D}$,a and E_{test}^{ecs} denotes corresponding test error for the TruncatedSVD model.

The `PowerLaw` and `TRACE-LOG` Model Selection Rules If we use good MSRs, then we expect that $\Delta E_{train} \to 0$ and $\Delta E_{test} \to 0$ as the models approach Ideal Learning. We consider the following MSRs,[39] which are associated with the `HTSR` and `SETOL` approaches.

- The `PowerLaw` MSR: All eigenvalues lying in the tail of the ESD, $\lambda_i \geq \lambda_{min}^{PL}$, where λ_{min}^{PL} is the start of the PL tail, as determined by the `WeightWatcher` PL fit, which is based on [67].
- The `TRACE-LOG` MSR: All eigenvalues lying in the tail of the ESD, such that they satisfy the Trace-Log Condition, i.e., $\lambda_i \geq \lambda_{min}^{|detX|=1}$, where $\prod \lambda_{i:\lambda_i \geq \lambda_{min}^{|detX|=1}} \simeq 1$.

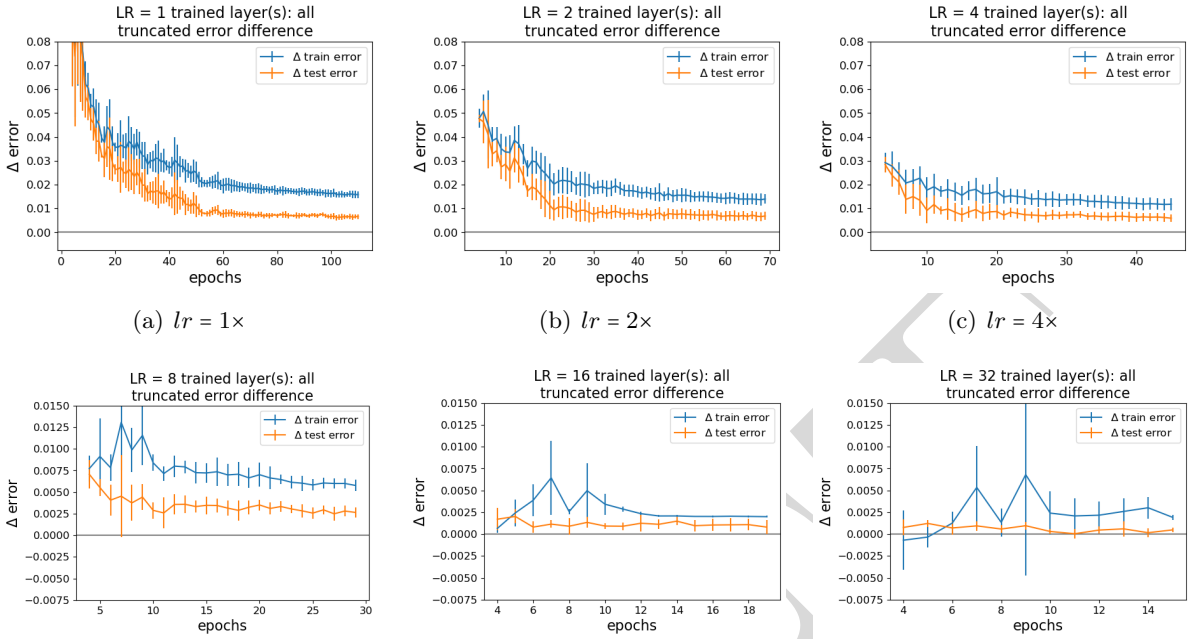
6.2.1 Train and test errors by epochs

To see how the Effective Correlation Space forms, we plot how ΔE_{train} and ΔE_{test} evolve over training, for each of the various learning rates considered.[40]

We start with the effect of the `PowerLaw` MSR. See Figure 17, where we see that ΔE_{train} and ΔE_{test} generally trend downwards as they approach minimum train error. When the learning rate is larger, the models converge more quickly, and ΔE_{train} and ΔE_{test} also converge to lower values. Recall from Figure 14(b) that $lr = 16\times$ had the lowest test error. In Figure 17(e), we see that it also has the lowest ΔE_{train} and ΔE_{test}. A lower ΔE_{train} or ΔE_{test} means that more of the models

[39] We considered other MSRs that do not "know about `HTSR` or `SETOL`, but they (expectedly) perform in trivial or uninteresting ways for testing the assumption of the Effective Correlation Space. Thus, we are not introducing just some arbitrary low-rank approximation, as is common, but instead that the specific `SETOL`-based MSR matters.

[40] When batch size was varied, results did not significantly differ, and so they are omitted.



(a) $lr = 1\times$ (b) $lr = 2\times$ (c) $lr = 4\times$

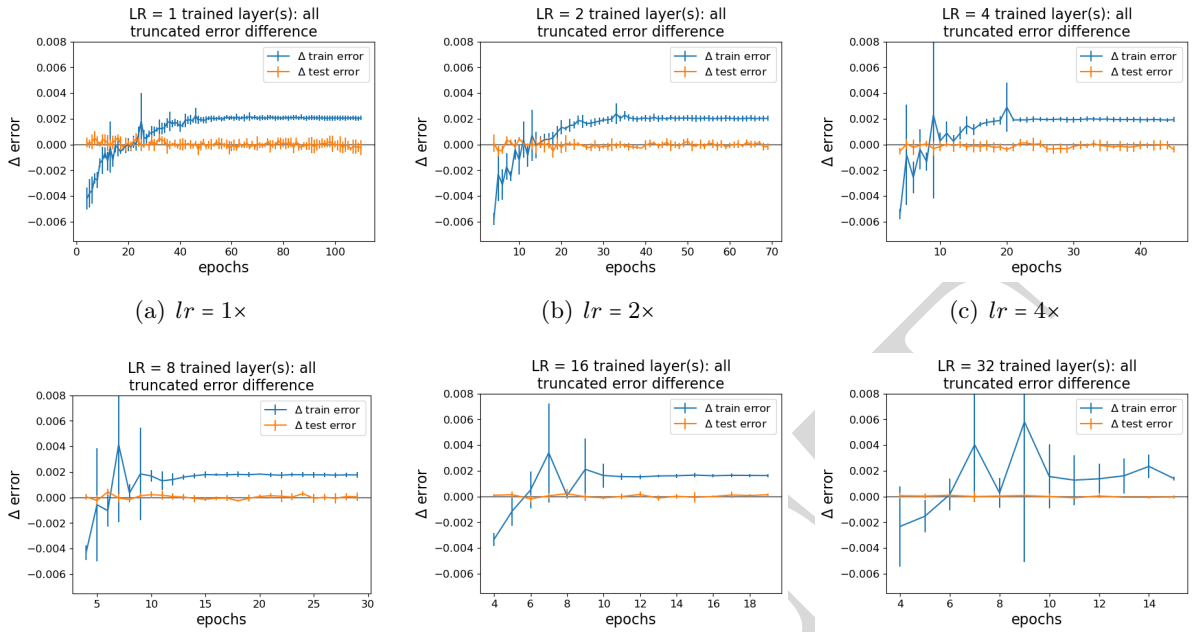
(d) $lr = 8\times$ (e) $lr = 16\times$ (f) $lr = 32\times$

Figure 17: ΔE_{train} (blue) and ΔE_{test} (orange) for various learning rates, using the `PowerLaw` MSR. As learning rate increases we can see that ΔE_{train} and ΔE_{test} both tend towards lower asymptotic minima, which they reach after fewer epochs of training. We can also see that (after the first few epochs,) ΔE_{train} (blue) is always higher than ΔE_{test} (orange). Observe that in the bottom row (d–f) the yaxis is contracted to make the variation more visible. In (f) we can see that as learning rate surpasses its optimal setting, the gap between ΔE_{train} and ΔE_{test} begins to increase again, and has wider error bars, suggesting that the excessively large learning rate is disrupting the MLP3s ability to lear the Effective Correlation Space.

correct predictions are due to the low rank tail, meaning that the tail generalizes better, and we see here that when the tail generalizes best, the model was the most accurate.

In each plot, we also see that the error bars are wide early on, before suddenly becoming much narrower. This transition is more visible in the larger learning rates shown in 17(d)–17(f), but can also be seen in 17(a)–17(c), albeit less clearly. Most interestingly of all, this transition is preceded by a brief period, sometimes a single epoch, in which the error bars are drastically wider, in a way that is reminiscent of a first-order phase transition. Again, this phenomenon can be seen most clearly in 17(d)–17(f).

We next consider the effect of the `TRACE-LOG` MSR. See Figure 18, which also shows the development of ΔE_{train} and ΔE_{test} over epochs, where we see a very different pattern in the train error and test error. The difference in the train error is because, as the model is untrained in the early epochs, the `TRACE-LOG` MSR *over*-estimates the tail by choosing a λ_{min} that is too small. Thus, ΔE_{train} actually increases to its asymptotic value at the final epoch. In the earliest epochs, the truncated train error is even *less* than the full MLP3 models error, suggesting that signal is forming in the large eigenvalues in these early epochs, but is swamped by the randomness of the early initial weights, some of which is then removed by truncation. As epochs progress, this effect disappears. Here again we can see the "phase-transition-like behavior of the train error, as the error bars are wide early on, up to a transition having an abnormally large error bar, after which they stabilize.



(a) $lr = 1\times$ (b) $lr = 2\times$ (c) $lr = 4\times$

(d) $lr = 8\times$ (e) $lr = 16\times$ (f) $lr = 32\times$

Figure 18: ΔE_{train} (blue) and ΔE_{test} (orange) for selected learning rates, using the `TRACE-LOG` MSR. For all learning rates, ΔE_{test} is centered on 0, meaning that the `TRACE-LOG` Effective Correlation Space explains almost all variation in out-of-sample predictions, but it does *not* explain all of the training set predictions (blue). NOTE: The y axis is the same in all plots, and is much narrower than in Figure 17. In (a–e) we can see that ΔE_{train} converges to approximately 0.002. Compare with Figure 17(e), which reaches a minimum of approximately 0.0025. As in Figure 17(f), the learning rate of 32× normal disrputs the MLP3s ability to discover the `TRACE-LOG` Effective Correlation Space.

Perhaps most interestingly of all, we see that under the `TRACE-LOG` MSR, ΔE_{test} is flat *throughout training*, and for all learning rates. Considering that this implies that there is *no* point in training where the `TRACE-LOG` tail generalizes badly, this is a rather striking observation. It also bears noticing that under the `PowerLaw` MSR, (Figure 17,) this is decidedly not the case, meaning that a small, but significant amount of generalization comes from the gap between these tails, but none at all comes from the bulk. [confusing]

There are two final points of comparison between Figures 17 and 18. First, although it appears in Figure 18 that ΔE_{train} converges to a larger value than in Figure 17, this is because the scale of the y-axis is 10× smaller. That is, the `PowerLaw` MSR is biased towards over-estimating λ_{min}, which means it over-truncates, producing a larger ΔE_{train} or ΔE_{test} than the `TRACE-LOG` MSR, which is biased towards under-estimating λ_{min}. Second, in both Figure 17 and 18, we can see that ΔE_{test} is consisitently lower than ΔE_{train}. Clearly, truncating has a larger effect on train predictions, meaning that no matter how long the model is trained, some of the train predictions are still derived from the bulk. Yet, the test predictions are far less affected, meaning that the Effective Correlation Space contributes to the models ability to generalize.

6.2.2 Truncation and Generalization

Given that ΔE_{test} is always lower than ΔE_{train} for the `PowerLaw` MSR, and similarly for `TRACE-LOG` after a certain point in training, it is clear that the Effective Correlation Space

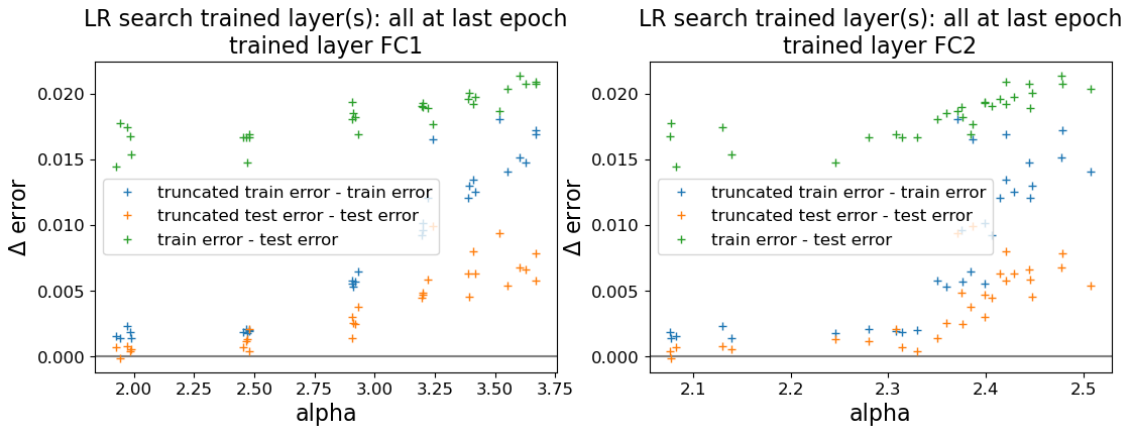


Figure 19: Train and test error gaps using the `PowerLaw` MSR, as a function of alpha in the FC1 and FC2 layers of MLP3 models, at the *final epoch* of training. We can see that as alpha decreases towards 2, (right to left), ΔE_{train} and ΔE_{test} generally decrease as well, meaning that the closer `Alpha` is to 2, the more the Effective Correlation Space explains the train and test predictions. The gap between un-truncated train and test error, (Eqn.81,) decreases as well until $\alpha_{FC1} < 2$.

has something to do with generalization. However, this leaves open the question of what role precisely `Alpha` plays. In Figure 19, we plot ΔE_{train} and ΔE_{test} with the `PowerLaw` MSR as a function of `Alpha` (rather than epochs) for layers FC1 and FC2, as well as the Generalization Gap– that is, $E_{test} - E_{train}$. (Recall Eqn. 81.) Learning rate is not explicitly shown, but its effect can be seen in the clusters of points that each learning rate generates.

In both layers, ΔE_{train} and ΔE_{test} steadily decrease with α_{FC1}, until it passes below 2, after which the relation deteriorates somewhat. This is especially prominent in FC2. Recall from Figure 15, Section 6.1, that when α_{FC1} passed below 2, the train error and test error both increased and exhibited larger variability. From this we interpret `Alpha` as a measure of *regularization* (which is consistent with its introduction as a measure of implicit self-regularization [25]). Regularization has the effect of keeping train and test accuracy closer together, and generally, as `Alpha` in the dominant layer decreases towards 2 from above, the train-test error gap decreases.

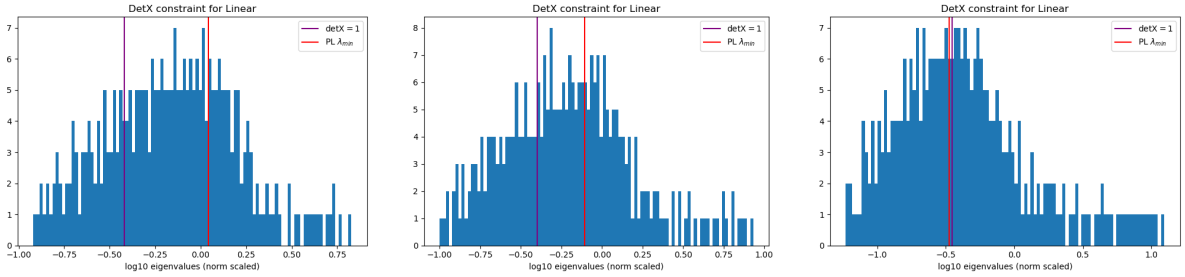
6.3 Evaluating the Trace-Log Condition

Having established that the PL tail of the ESD, defined by eigenvalues above λ_{min}^{PL}, is a major factor in determining model quality in the MLP3 model, we now examine how well the Trace-Log Condition compares with it. In particular, we demonstrate that when the tail of a layer ESD is described well by the `HTSR` phenomenology, i.e., when it is well-fit by a PL with $\rho(\lambda)_{tail} \sim \lambda^{-\alpha}$, with PL exponent $\alpha \simeq 2$, then the eigenvalues in the tail defined by the PL fit, i.e., $\lambda \geq \lambda_{min}^{PL}$, *also* satisfy the Trace-Log Condition of Eqn. 123—*a key assumption of the `SETOL` theory.* This is a rather remarkable empirical result that couples `HTSR` and `SETOL`; it has its basis in our `SETOL` derivation; and it provides the basis for an inductive principle that is based on the product of eigenvalues rather than an eigenvalue gap.

We can denote the eigenvalue that best fits the Trace-Log Condition as $\lambda_{min}^{|detX|=1}$. Then, to measure how well this condition holds, we can compute

$$\Delta\lambda_{min} = \lambda_{min}^{PL} - \lambda_{min}^{|detX|=1}. \tag{161}$$

In Sections 6.3.1 and 6.3.2, we will see the trend that as α approaches 2, λ_{min}^{PL} and $\lambda_{min}^{|detX|=1}$ also



(a) $LR = 8\times$ (b) $LR = 16\times$ (c) $LR = 32\times$

Figure 20: Log-Linear ESDs for three learning rates in the FC1 layer of MLP3. The red line shows λ_{min}^{PL}, and the purple line shows $\lambda_{min}^{|detX|=1}$. Observe that the purple line is to the left of the red line, but as the LR increases they move closer together. However, when LR is 32×, where both train and test accuracy suffered (c), the red line is to the left of the purple line. This is often a signature of an Over-Regularized layer, and indeed the FC1 layer in this model had $\alpha < 2$. (See Figure 15(a) in Section 6.1.)

approach one another, and hence $\Delta\lambda_{min}$ goes to 0, from above, both for our toy MLP3 model as for SOTA models. In our MLP3 model, we will see that a crossing of the equality condition coincides with over-regularization and a degradation in model accuracy. In pre-trained ResNet[106], VGG[107] and ViT[108] models, we will also see, empirically, that in general $\Delta\lambda_{min}$ remains positive, just as α remains above 2.

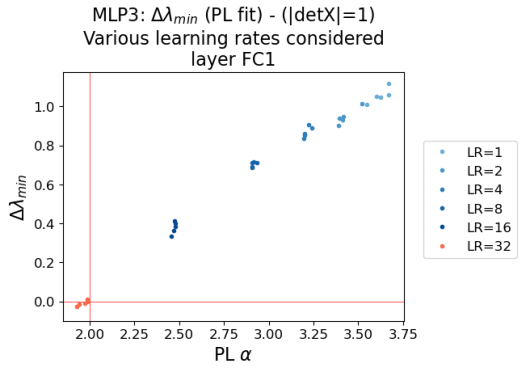
6.3.1 The MLP3 model

Consider Figure 20, which shows λ_{min}^{PL} and $\lambda_{min}^{|detX|=1}$ in the FC1 layer of three MLP3 models, each sharing a common starting random seed, that were trained with the largest learning rates. The λ_{min}^{PL} and $\lambda_{min}^{|detX|=1}$ eigenvalues are marked by red and purple vertical lines, respectively; and thus $\Delta\lambda_{min}$ is the distance between red and purple lines. As learning rate increases, the red and purple lines draw closer, and they are closest for $lr = 32\times$ (Figure 20(c)). (Compare this with Figure 15, Section 6.1, which shows that this corresponds with an increase in test accuracy, up to $lr = 16\times$, but at $lr = 32\times$ `Alpha` fell below 2 and accuracy suffered.) In Figure 20(a)-20(b), the purple line is left of the red line; but in Figure 20(c), the red and purple lines cross, such that $\lambda_{min}^{PL} < \lambda_{min}^{|detX|=1}$. This is analogous to the case where α crosses below 2. This suggests that the absolute Trace-Log is minimized when $\alpha \simeq 2$, which (remarkably) is exactly when the `HTSR` phenomenology predicts the layer is Ideal.

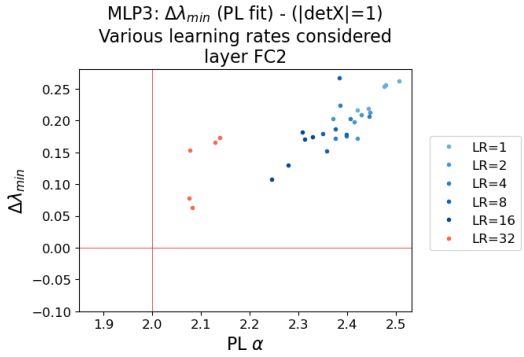
(Observe that Ideal does not necessarily mean optimal under a finite sized training set, but rather that the finite-sized system behaves the most similarly to its infinite limit.)

We can compare α and $\Delta\lambda_{min}$ more broadly by plotting $\Delta\lambda_{min}$ directly as a function of α in a single plot spanning all random seeds and learning rates or batch sizes. This is shown in Figure 21. Critical values of $\alpha = 2$ and $\Delta\lambda_{min} = 0$ are shown as vertical and horizontal red lines, respectively. Values for various learning rates are plotted for layer FC1 (Figure 21(a)) and FC2 (Figure 21(b)), as well as for various batch sizes in layer FC1 (Figure 21(c)) and FC2 (Figure 21(d)).

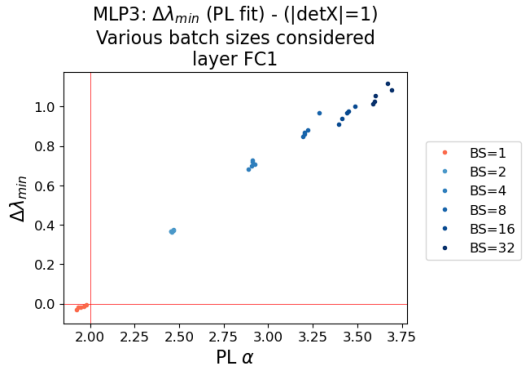
For layer FC1 (Figures 21(a) and 21(c)), in both cases we see near-linear march towards the critical tuple of $(\alpha, \Delta\lambda_{min}) = (2, 0)$. In addition, passing this critical value coincides with diminished train and test accuracy, (recall Figure 14), suggesting that just as $\alpha = 2$ is a threshold of over-regularization, $\lambda_{min}^{PL} < \lambda_{min}^{|detX|=1}$ may be as well. Since FC1 is the dominant layer, comprising roughly 8/9 of the weights of the model, (Table 10,) we expect FC1 to most closely match the



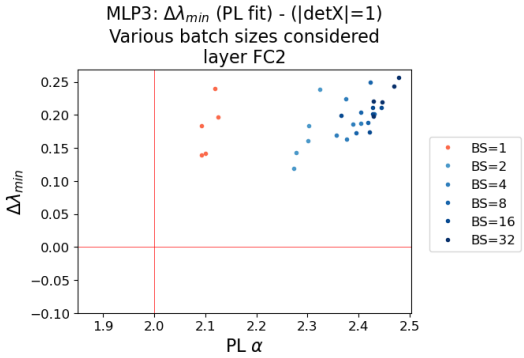
(a) Layer FC1 for various learning rates



(b) Layer FC2 for various learning rates



(c) Layer FC1 for various batch sizes



(d) Layer FC2 for various batch sizes

Figure 21: MLP3 Model: Comparison of the PL `Alpha` (x-axis), with the difference between λ_{min}^{PL} and $\lambda_{min}^{|detX|=1}$ (y-axis). The thin red lines indicate critical values of $\alpha = 2$ and $\Delta\lambda_{min} = 0$. As learning rate increases (a–b) or batch size decreases (c–d), we can see that in layer FC1, which dominates the model, (See Table 10,) α goes to 2, and $\Delta\lambda_{min}$ goes to 0. Observe that both critical values are crossed at the most extreme hyper-parameter selection, (red,) corresponding with over-training. Layer FC2 shows a weaker tendency towards the critical values (b, d), and is disrupted at the most extreme hyper-parameter values (red).

performance of the model as a whole.

For layer FC2, which comprises roughly the other 1/9 of the models weights, there is a similar coevolution, but it is weaker. As learning rate or batch size exceeds their critical values, rather than going to $(2, 0)$ as in FC1, we instead see that the relationship simply breaks down, with the gap growing larger even as α decreases. Given that FC1 *has* passed the critical $\alpha = 2$ threshold, we conjecture that the breakdown of the relationship between α and $\Delta\lambda_{min}$ is due to FC1 becoming atypical.

6.3.2 State-of-the-Art (SOTA) models

Here, we consider SOTA models, in particular VGG pre-trained models [107], the ResNet series [106], the ViT series [108], and the DenseNet series [109]. We show that as α approaches 2, the Log-Trace Condition holds better and better, i.e., $\Delta\lambda_{min}$ approaches 0.

Figure 22 plots α versus the difference $\Delta\lambda_{min}$, (Eqn. 161). Layer matrices from all models

in each series are pooled to generate the plots.[41] Notice that in Figure 22(a) – 22(d), $\Delta\lambda_{min}$ approaches zero as $\alpha \to 2$ from above. Individual points may have a large $\Delta\lambda_{min}$ for an α near to 2, but the overall trend is apparent. The rapid decrease of $\Delta\lambda_{min}$ as α approaches 2 from above, implies that the PL tail rapidly takes on a unit Trace-Log (if it doesnt already have it.)

As we saw in comparison of Figures 17 and 18, (Section 6.2.1,) the `TRACE-LOG` tail is generally larger, and always has highly generalizing components. Thus, it is plausible that as layers reach the limit of the amount of information that can be encoded in them, i.e. as α goes to 2, the `PowerLaw` tail expands to fill the `TRACE-LOG` tail. This effect can be seen clearly in Figure 22 in the condensing of the "funnel shape.

Recall from Figure 21 that layer FC1 dominated the model, (Table 10,) producing a clear progression of α and $\Delta\lambda_{min}$ towards $(2, 0)$, as a function of learning rate or batch size, whereas FC2 showed a slightly less clear relationship. In larger models having dozens of layers, we would not expect any one layer to dominate as thoroughly. Moreover, it is the architecture that varies between models in each series, not (necessarily) the hyperparameters, meaning that there would not be a straight linem, as in Figures 21(a) and 21(c). However, with all of the layers contributing to varying degrees, we nevertheless see a clear trend in both plots of Figure 22. These results show how the single-layer `SETOL` theory extends from the MLP3 model, which is dominated by a single layer, to larger models where many layers interact in complex ways.

The overall pattern of relationship between $\Delta\lambda_{min}$ and α can also be seen in Figure 23, which shows plots for Large Language Models (LLMs) of the Falcon [110] and LLAMA [111] model families, for different numbers of parameters. Observe that each subfigure 23(a)–23(d) shows a single model, rather than a collection of models in a family, as in Figure 22. The y-axis is the same between models in the same family. As in Figure 22, there is a general outline of a "funnel shape pointing towards the critical point $(2, 0)$, with the exception that it is only reached in the case of LLAMA-65b, (Figure 23(d)). This suggests that these LLMs are larger than they necessarily need to be, consistent with prior work [79], but also that they are well guarded against Over-Regularized layers beyond the critical point ($\alpha = 2$ and $\Delta\lambda_{min} = 0$).

6.4 Layer Qualities with Computational R-transforms

In this Section, we look at how the `HTSR` layer quality HT PL metric α compares to computing the Layer Quality-Squared $\bar{\mathcal{Q}}^2$ using the Computational R-transform method proposed in Section 5.5 for fully trained MLP3 model(s). Figure 24 presents results for FC1 and FC2 layers, comparing both the batch size and the mean α metric to the mean $\bar{\mathcal{Q}}^2$, and the results are quite different.

For FC1, as shown in Figure 24(a), the quality metric $\bar{\mathcal{Q}}^2$ increases as the batch size decreases, following the expected trend. Notably, for batch size = 1, the quality metric exceeds 100%, which suggests that the layer may be overfit in this scenario, which is similar to results obtained earlier. Furthermore, in Figure 24(b), the quality metric $\bar{\mathcal{Q}}^2$ shows a strong correlation with the α metric, consistent with the theoretical predictions of the `HTSR` framework. Earlier results suggest that the FC1 layer captures most of the correlation in the data, so it is interesting that this layer also has much smaller error bars, indicating more consistent quality across different batch sizes.

In contrast, for FC2, while the general trends are similar (as shown in Figures 24(c) and 24(d)), the error bars on the quality metric $\bar{\mathcal{Q}}^2$ are much larger. This indicates significantly higher variability in computational quality compared to FC1. The large error bars for FC2 suggest that, while the $\bar{\mathcal{Q}}^2$ metric is theoretically grounded, it is less effective than the existing `HTSR`

[41] For convolutional layers, the `WeightWatcher` tool first computes eigenvalues for all channels-to-channels linear operators separately, and then pools them in order to compute α, λ_{min}^{PL} and $\lambda_{min}^{|detX|=1}$. For instance, in a $64 \times 64 \times 3 \times 3$ weight tensor, there would be 9 separate linear operators of 64×64, giving 576 eigenvalues, which would then be pooled to compute α, λ_{min}^{PL} and $\lambda_{min}^{|detX|=1}$.

layer quality metric α, which exhibits greater stability and reliability in capturing layer behavior. These differences highlight the distinct computational characteristics of the two layers, with FC1 demonstrating more predictable trends that align closely with theoretical expectations.

6.5 Inducing a Correlation Trap

Previous work on the HTSR phenomenology [26, 24] has shown that one can look for quantitative deviations from the necessary pre-conditions of tradtitional RMT (particularly that the weights are 0-mean and finite variance) to detect when a model layer suffers from some other anomaly in the elements, which we call a "Correlation Trap (see Section 3.3.1 and [26, 24]). A Correlation Trap may cause, or be caused by, the over-regularization leading Alpha to fall below 2, (see Section 3.3.2). Here, we explore this in greater detail, in light of our SETOL.

Lets look at the ESDs of the FC1 layer of the MLP3 model, for learning rates lr = 16× and lr = 32× normal. We will be interested in the general shape of the ESD of $\frac{1}{N}\mathbf{W}^T\mathbf{W}$. For the purposes of detecting a Correlation Trap, we will randomize $\mathbf{W}^T\mathbf{W}$ element-wise, and then observe its largest eigenvalue λ_{rand}^{max}.

Figure 25 shows the ESDs of the original matrix (green) and the element-wise randomized rand($\frac{1}{N}\mathbf{W}^T\mathbf{W}$) (red). Observe in particular λ_{rand}^{max} for each learning rate factor. For lr = 16×, Figure 25(a) shows that the ESD of $\frac{1}{N}\mathbf{W}^T\mathbf{W}$ is HT, whereas the ESD of the randomized matrix is essentially a distorted semi-circle—as expected from the well-known MP result; and that λ_{rand}^{max} lies at the edge of the random MP Bulk ESD. (A similar result is seen for smaller learning rates.) In contrast, for lr = 32×, Figure 25(b) shows that while the original ESD is again HT, the ESD of the randomized has one large element, λ_{rand}^{max}, that pulls out from the MP bulk. This is the signature of a Correlation Trap; and it co-occurrs with the exact learning rate setting that degraded the train and test accuracies, pushing Alpha below its optimal value of $\alpha \simeq 2$. When this happens, both the estimation of Alpha and the formation of a PL tail are potentially disrupted.

Correlation Traps have been observed previously [26, 24], using the HTSR phenomenology. However, SETOL provides an explanation for why this would be expected to occur — non-standard element-wise distributions will tend to interfere with the properties of the spectrum which SETOL analyzes. Our derivation in Section 6 suggests that in order to apply the SETOL effectively one must avoid (or remove) such traps. This too has been observed previously [26, 24].

6.6 Overloading and the Hysteresis Effect

Here, we do such-and-such. XXX. HOW PRECISELY TO FRAME.

Obtaining a value of α outside the $\alpha \gtrsim 2$ range is indicative of "overloading [8, ?] that layer. This can be accomplished, e.g., by training only one layer in the MLP3 model. As the two layers have very different sizes, we see markedly different behaviors, that are nevertheless consistent with theory.

The SETOL theory is based on the idea that NNs undergoing training behave like Statistical Mechanic systems relaxing to an equilibrium. So far, we have tested the theory under conditions that are approaching Ideal. However, for the theory to be useful in practice, we must also examine how it performs in non-Ideal situations. Of particular interest, we would like to examine the theory under conditions where the training dynamics slows down, i.e., when it is in a "glassy or meta-stable state. One way we can do this is to train only one layer, and freeze the rest. Doing so overloads the single trainable weight matrix, as a function of the ratio of examples to trainable parameters [8, ?, 89], and we expect this to cause α_{FC1} or α_{FC2} to drop well below 2.

6.6.1 Baseline: Loading onto both FC1 and FC2

To start, Figure 26 shows α_{FC1} and α_{FC2}, binned in units of 0.05, versus train and test error over all epochs of training [42] for each of the different learning rates considered. (Cf. Figures 15 and 16, Section 6.2, which show only the final epoch.) Binning was done so as to facilitate averaging over the 5 starting random seeds; linear fits are shown separately for each learning rate; and error bars represent one standard deviation within each bin. The critical value of $\alpha = 2$ is shown as a vertical red line in all plots. For each learning rate, train error and α decrease together during training, which can be seen for both α_{FC1} (a) and α_{FC2} (c), reading each line from the top right to bottom left. Test errors, (b, d) show a similar trend, but with wider error bars. Observe that the range of the y-axis is narrower for test error to make detail more visible.

We see in Figure 26 (a,b) that the 32× learning rate causes α_{FC1} to decrease faster than any other, putting it on course to fall below 2 before train error reaches $\simeq 0$. (See Figure 14, at the beginning of this Section.) We also see that the slower learning rates cause train error to reach $\simeq 0$ well before α_{FC1} can reach 2, and this offers an explanation as to why their test error was higher.

6.6.2 Overloading FC1

In contrast, when only FC1 is trained, Figure 27 shows that, as expected, α_{FC1} decreases well below 2. Training error (a) generally trends downward as α_{FC1} decreases, but no matter the learning rate, the relation is the same, or nearly so, because only one layer is being trained. Consequently, all learning rates were pooled for one linear fit.~~, for visibility.~~

In Figure 27 we can see demonstration of a crucial claim of `SETOL` theory: that for $\alpha_{FC1} > 2$, (vertical red line,) the test error declines linearly (b) is almost perfectly linear with decline in α_{FC1}, however, when $\alpha_{FC1} < 2$, the curve bends upward. Furthermore, the precision with which the trajectory changes as α_{FC1} passes the threshold provides ample validation that the estimator of [67] is indeed accurate. We observe that test error may continue to decrease after $\alpha_{FC1} < 2$, however the rate of decrease is significantly less. We can also see that in some sense, the model is "doomed to always have train error reach $\simeq 0$ when $\alpha_{FC1} \simeq 1.7$, i.e., after $\alpha = 2$, because of the number of trainable parameters, and perhaps because of the lack of a modulating influence of FC2 seen in Figure 26.

6.6.3 Overloading FC2

The MNIST dataset has 60,000 training examples, which means that an FC1-only model is over-parameterized, but FC2 is substantially *under*-parameterized [112]. (Table 10) This drastically changes the meaning of an experiment where only FC2 is trained. Figure 26 (Section 6.6.1) shows α_{FC2} vs. train error (c) and test error (d) when *all* layers are trained. There we see a different relationship between α_{FC2} and train and test error for each learning rate. None of the training runs reached an α_{FC2} of 2, as the load was split between both layers FC1 and FC2.

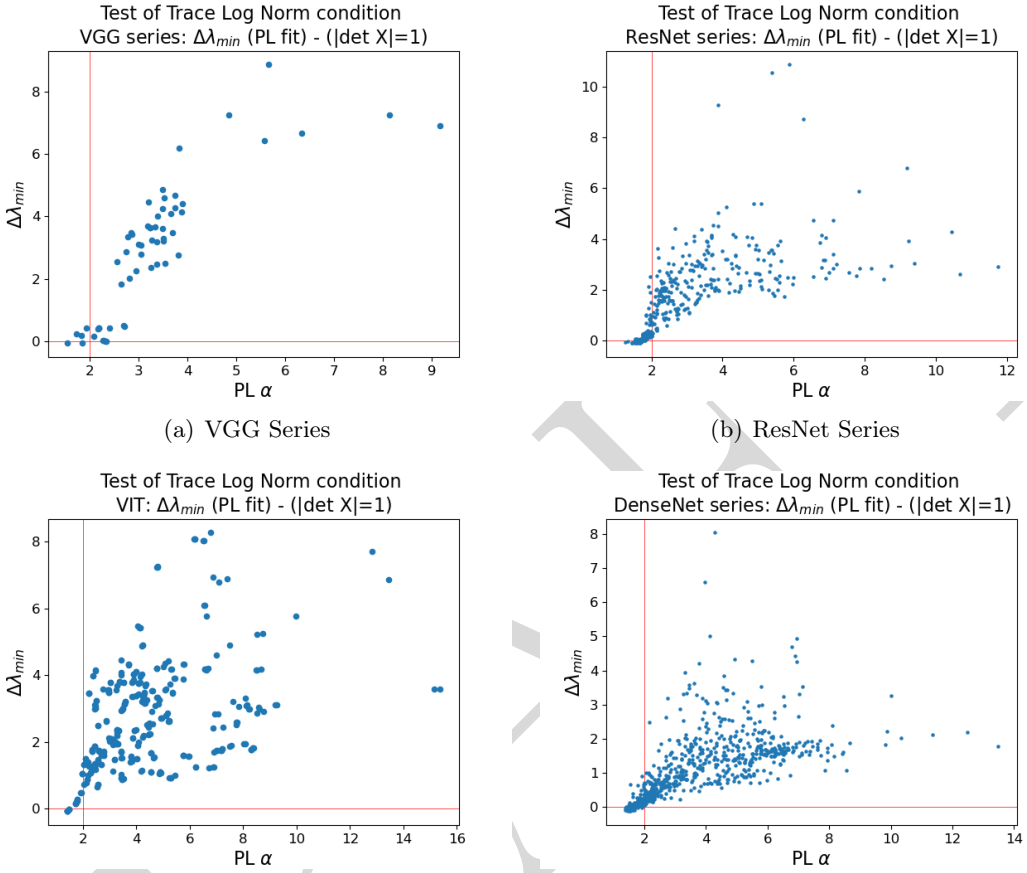
Figure 28, however, shows a starkly different relationship. When only FC2 is trained, each random seed produces a different trajectory, meaning that they cannot be plotted as a single curve, even with error bars. Train (a, c) and test (b, d) error rates are shown as a function of α_{FC2} for the two highest learning rate factors, 16× (a, b) and 32× (c, d). Lower learning rate factors, (not shown,) showed the same trend, except that they did not progress as far as those shown in Figure 28.

First we see, for all seeds, α_{FC2} decreases all the way down to 1.5, after which it begins to rebound to the right. As it does, train error continues to decrease. We can also see that test

[42] Excluding the first four epochs when the matrix is still essentially random.

error continues to decrease as well, down to its minimum value of slightly more than 0.03, as α_{FC2} continues to increase for a short time. However, at the exact point where test error reaches a minimum, the PL tail itself begins to fracture, leading to different estimates of α_{FC2} for each seed. As each of the five starting random seeds are shown separately, we can see that each of them terminates at a different point, some of which are closer to 2 than others.

Such a reversion to a more typical value of α_{FC2}, prior to the fracturing of the PL tail, is reminiscent of a spin glass system relaxing towards its minimum energy configuration, in a way that retains some memory of the path taken on the way to its current state. We conjecture that if the model were trained for sufficiently many epochs, (perhaps many thousands,) then the tail would re-form, and α_{FC2} would reform, and revert all the way back to its stable value of 2.



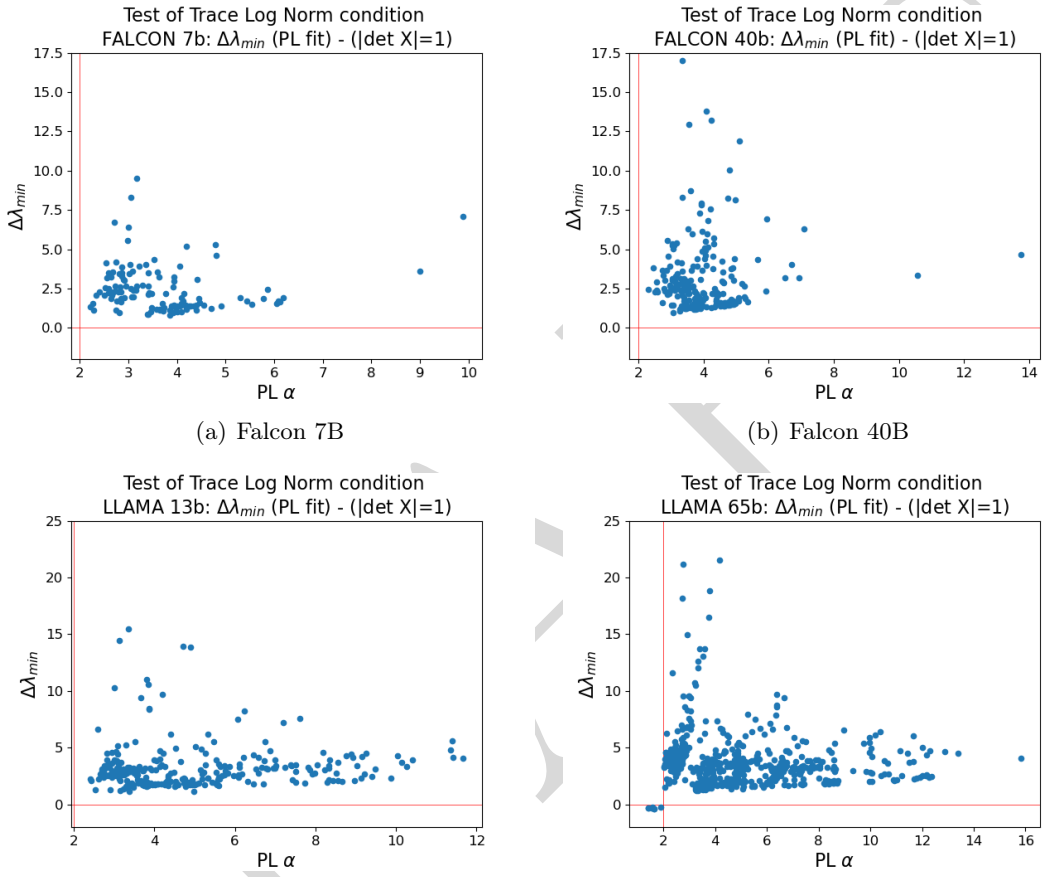
(a) VGG Series

(b) ResNet Series

(c) ViT Series

(d) DenseNet Series

Figure 22: Difference between the two λ_{min} estimates, $\Delta\lambda_{min}$, (Eqn. 161), as a function or α, for linear and convolutional layers in series of VGG [107], ResNet [106], ViT models [108] and DenseNet models [109]. Layer matrices for all models in the series were pooled to create each plot. In (a), (VGG,) we see three clusters of points – those with $\Delta\lambda_{min}$ close to 0 and α close to 2, those with $\Delta\lambda_{min}$ above 2 and $\alpha > 2.5$, and those with $\Delta\lambda_{min}$ above 6 and $\alpha > 3.5$. In (b), (ResNet,) we see that in general, as α shrinks towards 2, $\Delta\lambda_{min}$ tends towards 0, overshooting slightly. We also see that the difference $\Delta\lambda_{min}$ is almost always positive, with few exceptions, and even the layers that do not overshoot form a kind of "funnel shape pointing towards the critical point $(2, 0)$. In (c), (ViT,) we also see the same general relationship between α and $\Delta\lambda_{min}$ across layers of several ViT models. Observe that ViT models do not have convolutional layers, and in spite of this, the overall pattern is similar. In (d), (DenseNet,) we see a similar overall trend as in (b), except that $\Delta\lambda_{min}$ tends to decrease sooner, but there are also more layers with $\alpha < 2$ and $\Delta\lambda_{min}$ above 0.



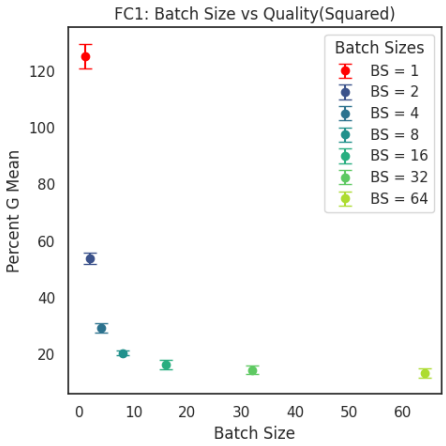
(a) Falcon 7B

(b) Falcon 40B

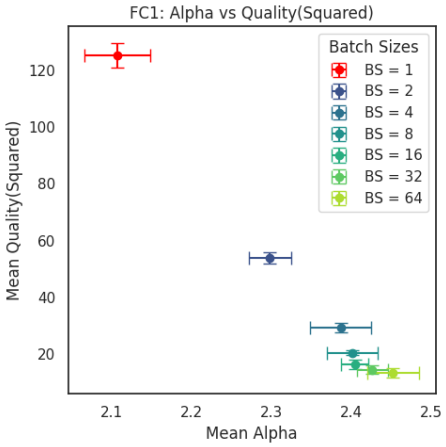
(c) Llama 13B

(d) Llama 65B

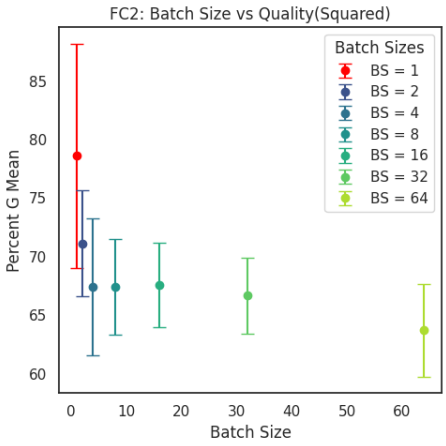
Figure 23: Difference between the two λ_{min} estimates, $\Delta\lambda_{min} = \lambda_{min}^{PL} - \lambda_{min}^{|detX|=1}$, as a function of α, for all linear layers in the FALCON[110](a-b) and LLAMA [111](c-d) language models for varying numbers of parameters. As in Figure 22, we see that in recent Large Language Models, $\Delta\lambda_{min}$ remains positive, except where $\alpha < 2$ (d). Otherwise, a "funnel shape can still be seen leading towards the critical point $(2, 0)$ as in Figures 22(b) and 22(c) Observe that the x- and y-axes are different between sub-figures due to the differences in scale of each model.



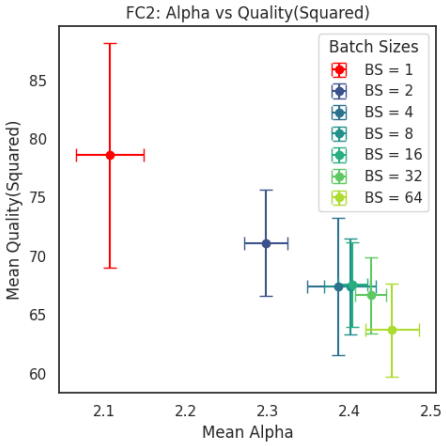
(a) FC1 batch size vs $\bar{\mathcal{Q}}^2$



(b) FC1 Layer α vs $\bar{\mathcal{Q}}^2$

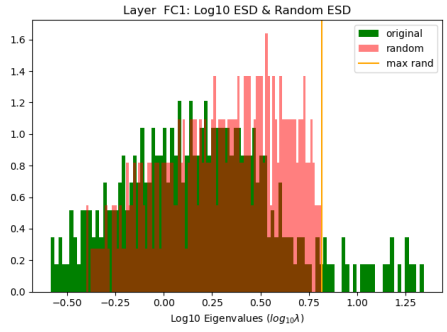


(c) FC2 batch size vs $\bar{\mathcal{Q}}^2$

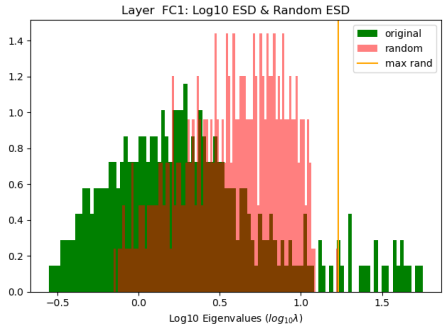


(d) FC2 Layer α vs $\bar{\mathcal{Q}}^2$

Figure 24: Evaluation of the computational R-transform Layer Quality-Squared metric $\bar{\mathcal{Q}}^2$ on the fully trained MLP3 model(s) for different batch sizes.



(a) MLP3 Learning Rate 16×



(b) MLP3 Learning Rate 32×

Figure 25: ESD plots for learning rate $lr = 16\times$ and $lr = 32\times$ normal, shown on Log-Lin scale, as computed using the `WeightWatcher` tool, for the FC1 weight matrix $\mathbf{W}$ (green) and an element-wise randomized rand($\mathbf{W}$) (red). This provides an example of inducing a Correlation Trap in the MLP3 model, simply by increasing the learning rate used during model training. See Section 3.3.1.

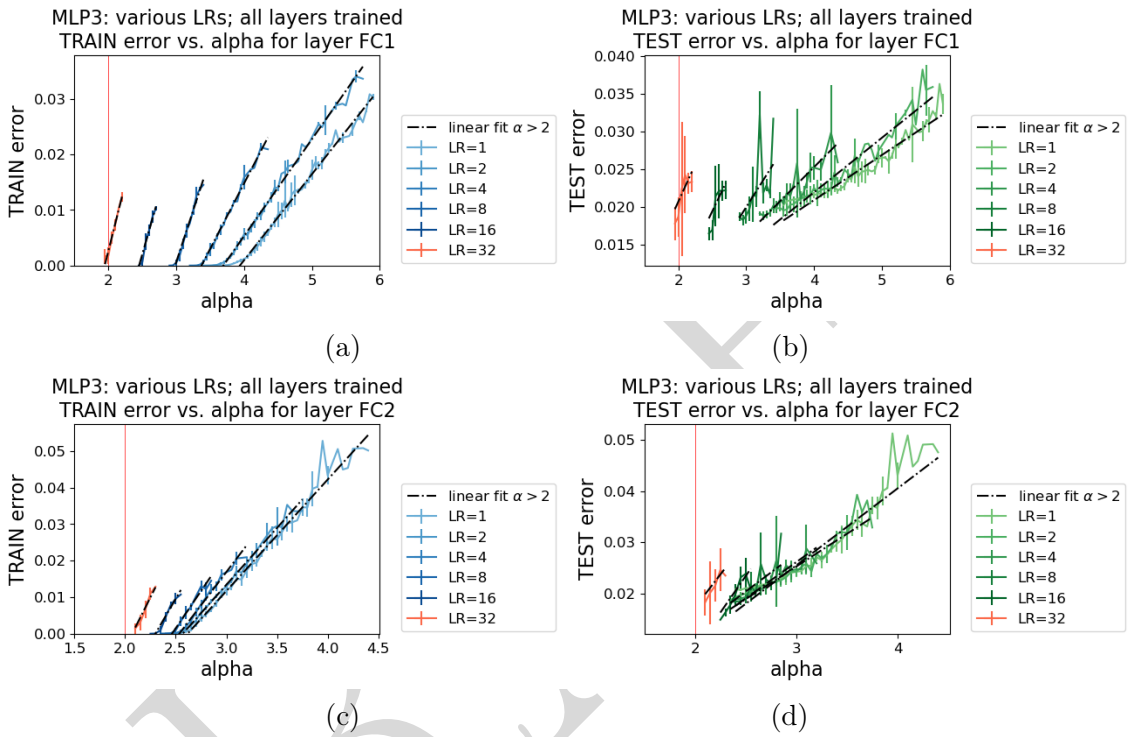
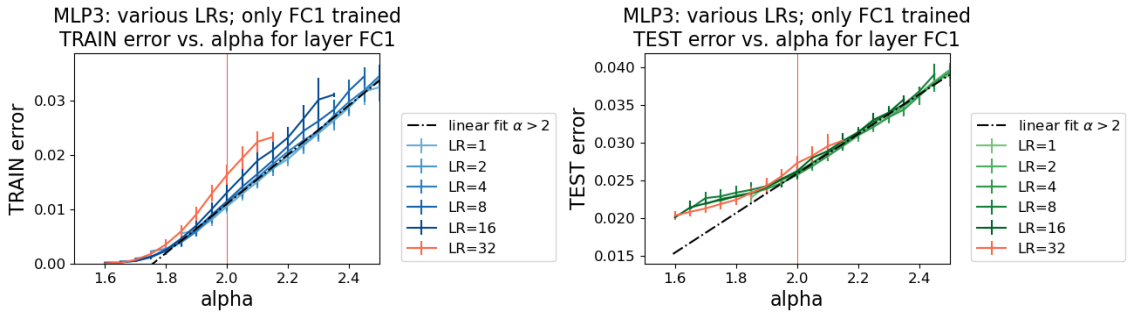


Figure 26: Train (a, c) and test (b, d) accuracy as a function of α_{FC1} (a, b) and α_{FC2} (c, d) when **all layers are trained**. Red vertical lines show the critical value of $\alpha = 2$, and dashed black lines show linear fits of error, using only points where $\alpha > 2$ and train error > 0.001. For FC1 (a, b), we can see that each learning rate produces a different trajectory of train error (a) and test error (b) as a function of α_{FC1}, showing that even though FC1 dominates overall, (Table 10,) FC2 still plays a modulating role. (Cf. Figure 27 where there is only one trajectory.) In (c, d) we can see that α_{FC2} never goes below 2. As in (a, b), each learning rate produces a different trajectory, though there is greater overlap for lower learning rates. See discussion in Section 6.6.1.



(a) (b)

Figure 27: Train and test accuracy as a function of α_{FC1} when only FC1 is trained (a, b). Red vertical lines show the critical value of $\alpha_{FC1} = 2$, and dashed black lines show linear fits of error, using only points where $\alpha_{FC1} > 2$. In contrast with Figure 26, when only FC1 is trained, we can see that no matter the learning rate, there is ony one trajectory, for both train error (a) and test error (b). Hence, for visibility, only one linear fit, using all learning rates pooled, is shown. Crucially, we can see in (b) that as α_{FC1} passes below 2, the *test error* trajectory changes, for all learning rates, even as the *train error* trajectory does not, until it reaches ~ 0. This suggests that even though test accuracy can still decrease when $\alpha_{FC1} < 2$, it does so at a decreased rate relative to α_{FC1}. See discussion in Section 6.6.2.

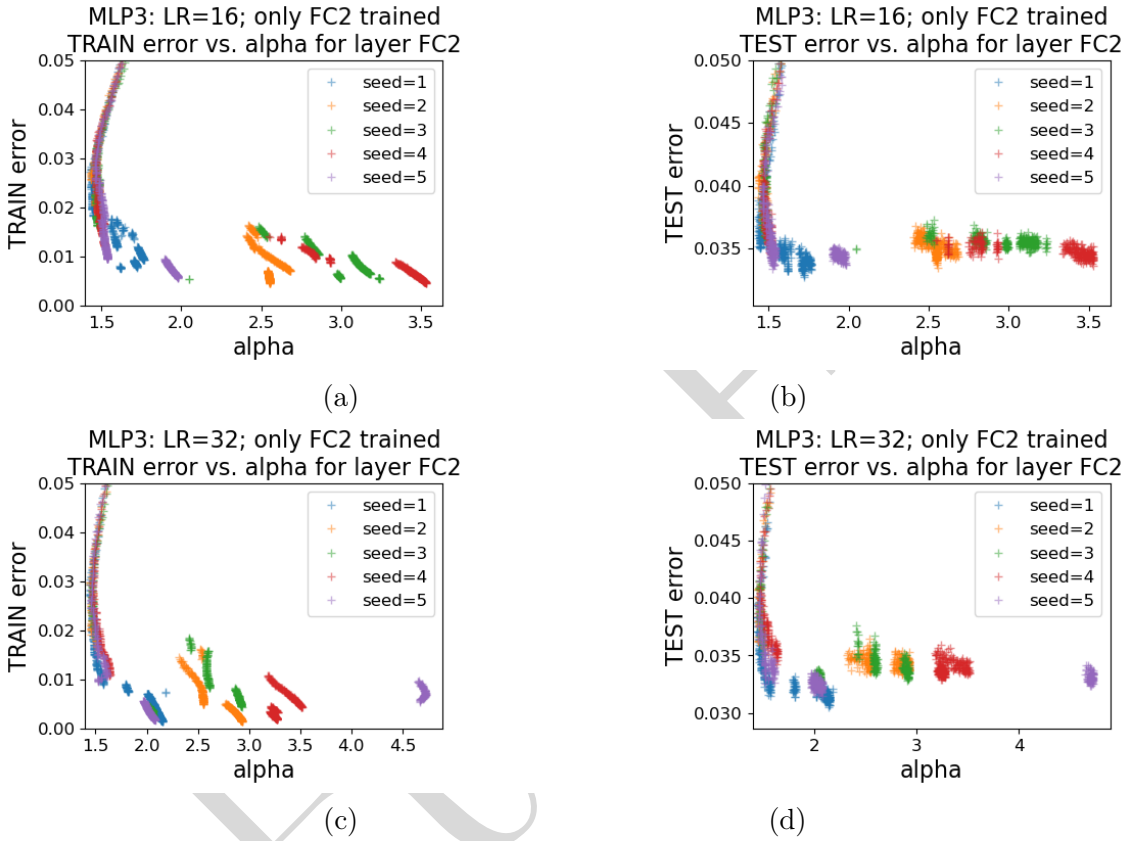


Figure 28: Train error (a, c) and test error (b, d) as a function of α_{FC2} when all other layers are frozen, for the two largest learning rates, $lr = 16\times$, (a, b) and $lr = 32\times$(c, d). Cf. Figure 26, wherein all layers were trained. Due to the markedly different behavior of each random seed, they cannot be plotted as means and error bars, and are instead shown separately. The path taken by all seeds up to $\alpha_{FC2} = 1.5$ has a slight curvature characteristic of a hysteresis-like behavior. Observe also that the fragmenting into separate paths, due to the breakdown of the PL tail, coincides roughly with each seeds path reaching its minimum test error. The y-axis is scaled differently for train and test error to make variation more visible. See discussion in Section 6.6.3.

7 Conclusion and Future Directions

In this work, we have introduced SETOL, a *Semi-Empirical Theory of (Deep) Learning* that unifies concepts from Statistical Mechanics (StatMech), Heavy-Tailed (HT) Random Matrix Theory (RMT), and quantum-chemistry-inspired approaches to strongly correlated systems [?, ?, ?]. SETOL aims to provide a solid theoretical foundation for the Heavy-Tailed Self-Regularization (HTSR) phenomenology, including the widely used Alpha and AlphaHat HT Power Law (PL) Layer Quality metrics, which are implemented in the open-source WeightWatcher toolkit. Specifically, SETOL reformulates the Neural Network (NN) learning problem for a single layer as a matrix generalization of the classic Student-Teacher (ST) model for Perceptron learning and is analyzed within the Annealed Approximation (AA) at high temperatures (high-T). This reformulation results in a model expressed as an integral over random Student correlation matrices, commonly referred to as the Harish-Chandra–Itzykson–Zuber (HCIZ) integral. To evaluate this integral, we recast the solution using a technique derived from first principles, analogous to a single step of the Exact Wilson Renormalization Group (RG) theory [?]. Leveraging recent results [82, ?], we express the Layer Quality as a sum of integrated matrix-cumulants from RMT (i.e., R-transforms). Finally, we conduct both direct and observational experiments to validate key assumptions of the SETOL framework and empirically connect it to the HTSR theory.

Key Contributions and Observations.

A. *Rigor for the HTSR Phenomenology.* SETOL explains *why* power-law (PL) exponents in the layer spectral densities (e.g. Alpha and AlphaHat) act as robust diagnostics of generalization, even in large, complex architectures without access to training or test data. Our analysis ties these Heavy-Tailed ESDs to a Volume Preserving Transformation associated with an *Effective* Free Energy landscape, and suggesting (in analogy with traditional StatMech phases in learning theory) that the HTSR condition $\alpha \approx 2$ marks a phase boundary between optimal generalization and overfitting.

B. *Matrix-Generalized Student-Teacher (ST) Model.* SETOL is formulated as matrix generalization of the classical (vector-based) ST perceptron learning, incorporating $N \times M$ layer weight matrices, $\mathbf{w} \rightarrow \mathbf{W}$. Key to this generalization is isolating the top eigenvalue/eigenvector directions—called the *Effective Correlation Space* (ECS)—before evaluating the resulting partition function (or HCIZ integral). The ECS contains the HTSR PL tail, validating that the tail captures the dominant layer generalizing components.

C. *Trace-Log Condition & Alpha = 2.* A remarkable empirical observation, predicted by SETOL, is that layers near *ideal* training also satisfy $\ln\left(\prod \lambda_i\right) \approx 0$ in their tail eigenvalues; equivalently, $\sum \ln \lambda_i \approx 0$. We call this the TRACE-LOG condition (or DetX in WeightWatcher). Empirically, this condition appears when the HTSR Alpha ≈ 2. SETOL thereby *unifies* two previously separate heuristics for "optimal" or so-called Ideal behavior.

D. *Empirical Validation of SETOL* To validate the ECS and the TRACE-LOG condition, we trained small (3-layer) Multi-Layer Perceptron (MLP3) on MNIST and under varying batch sizes and learning rates. Using this, we verified that when the HTSR $\alpha \approx 2$ the SETOL TRACE-LOG condition also (usually) holds, and that one can reproduce the training accuracy by retaining only the ECS. This is further confirmed by using the WeightWatcher tool to examine the Alpha and DetX metrics common, open-source CV and NLP models, including modern LLMs (ResNets, DenseNets, ViTs, and LLMs like LLaMA and Falcon).

E. *Correlation Traps & OverFitting.* We observe that when layer ESDs with $\alpha < 2$, they exhibit behavior that can be interpreted as *over-regularization* and/or *overfitting*. For example, we

observe what we call *Correlation Traps*, large rank-one perturbations in the (randomized) layer weight matrix $\mathbf{W}$ that can be induced by training with excessively small batch sizes (bs=1) and are associated with degraded test accuracy, and which cause the `HTSR` to drop to $\alpha < 2$.

Additionally, we can induce a overfitting by freezing all but one layer and then training, which causes the layer $\alpha < 2$. By training in the underparameterized regime, we observe path-dependent, "glassy" behavior.

F. *Connection to Semi-Empirical Methods.* Conceptually, `SETOL` parallels well-known *Semi-Empirical* methods in quantum chemistry [**?**, **?**, **?**], wherein complicated many-body Hamiltonians are approximated by effective theories, but fitted or validated using empirical data. By retaining only the largest spectral modes (`ECS`) and imposing the `TRACE-LOG` condition, we can describe crucial low-rank correlations while discarding less relevant interactions, much like in Freed–Martin Effective Hamiltonian theories and/or Wilson's Renormalization Group (RG) approach.[56]

RG Analogy: A One-Step View. From a Renormalization Group perspective, restricting to the measure on the partition function to the `ECS` is akin to performing a *single step* of the Wilson Exact Renormalization Group (ERG). In doing this, we are discarding bulk "uninteresting" degrees of freedom in favor of the strongly correlated HT *long-ranged* modes. This leads to an effective model with fewer degrees of freedom but *renormalized* interactions–interactions that are dominated by the largest eigenvalues. This analogy with RG theory suggests that the `HTSR` phenomenology, where $\alpha \in [2, 6]$ in the Fat-Tailed Universality Class, is essentially describing a near-critical phase when $\alpha \approx 2$ and satisfies $\ln(\prod \lambda_i) \approx 0$ in its ECS. Departing from this point ($\alpha < 2$) leads to suboptimal results, consistent with the multi-phase pictures in `StatMech` spin glass theories of learning [8, 6, 4, 5].

Toward Understanding "Why Deep Learning Works." A key question in deep learning theory is why large neural networks achieve strong generalization despite operating in highly non-convex optimization landscapes. From the perspective of RG exact theory, this phenomenon can be partially understood through the concentration of generalization-relevant components. Specifically, models trained in regimes exhibiting *Lévy-like* or *Power-Law (PL)* couplings lead to the emergence of effective low-dimensional descriptions, where irrelevant modes are suppressed. The `WeightWatcher Alpha` metric quantitatively captures this concentration by characterizing how the optimization landscape stabilizes near criticality, with $\alpha \approx 2$ signaling optimal generalization in a near-critical regime.

Relation to Levy Spin Glasses and Heavy-Tailed Random Matrix Models. One long-standing puzzle is why large NNs avoid the worst of highly non-convex optimization, despite nominal exponential degeneracies. `SETOL` offers a partial explanation: if the trained model has *Levy-like* or *PL* couplings, then typical spin-glass degeneracies can be lifted, leaving the layer in a finite number of near-critical minima. In analogy with older work on *Levy Spin Glasses*[40], it is proposed `WeightWatcher Alpha` metric effectively measures how *rugged yet stable* the effective energy landscape is for each layer, with $\alpha \approx 2$ signifying a sweet spot of Ideal generalization.

7.1 Future Directions

1. Multi-Layer RG and Layer Interactions. While `SETOL` is formulated per-layer, modern DNNs stack many layers, each potentially with different α. An improved approach would treat

layer-layer interactions, potentially with a multi-step RG or other approach, that could track how these exponents evolve or couple across depth. It is possible that certain layers (e.g. final fully-connected heads in LLMs) exhibit α far from 2, while others converge near 2—raising questions about how best to address or combine them.

2. Explicit R-transform Expansions for Heavy-Tailed ESDs. In practice, large pre-trained models often show truncated power-law (TPL) ESDs, with $\alpha \geq 2$, and with sharp tail cutoff (e.g. $\lambda_{\max}$) [27, 28]. A possible extension of `SETOL` could treat these cases by deriving the explicit form of the R-transform for PL/TPL ESDs with $\alpha \in [2, 3, 4]$, and applying it even in cases with the `TRACE-LOG` condition does not strictly hold.

3. Practical Diagnostics and Fine-Tuning. The open-source `WeightWatcher` tool has already seen success diagnosing layer quality. Integrating `SETOL` 's `TRACE-LOG` condition may refine this further, helping users identify correlation traps or "under-exploited" layers. There is also strong potential for using `Alpha` or `TRACE-LOG`-based signals during training or fine-tuning: e.g. automatically adapting learning rates to push each layer closer to $\alpha = 2$, for fine-tuning models with significantly less memory[113], for compressing large LLMs[30], and other practical applications.

4. Correlation Traps and Meta-Stable States. Although our experiments show how small batch sizes or large learning rates can induce correlation traps, a quantitative theory of *where and why* traps occur remains open. Clarifying these states could enable *trap-avoidance* strategies, e.g. partial re-initialization or specialized regularizers that favor lower-rank updates in the ECS. In large-language-model (LLM) contexts, correlation traps might manifest as *hallucinations* or *mode collapse*, motivating deeper analysis.

4. Analyzing the Layer Null Space. One critical but often overlooked factor is the potential null space within model layers, which can emerge during overfitting. This null space represents parameter directions that fail to contribute meaningfully to generalization but instead encode redundant or overly specific patterns tied to the training data which might be ignored or forgotten. Future work should examine if and when NN layers have components in their null space, which contribute significantly to the performance of the model.

5. Layer-Layer Cross-Terms. `SETOL` is a single layer theory, however, as noted in Section 5.1, it would be desirable to extend the theory to including layer-layer cross terms. While we don't have an exact expression for this, we can propose a phenomenological guess that the leading order term would be the integrated R-transform, defined for the overlap between nearest-neighbor weight matrices (i.e., $\mathbf{W}_1, \mathbf{W}_2$) that can be algined along a common axis. This term would take the form $\mathbb{G}_{\mathbf{R}_{1,2}}(\lambda_{1,2})$ where $\mathbf{R}_{1,2} \sim \mathbf{W}_1^\top \mathbf{W}_2$ and $\lambda_{1,2}$ is an eigenvalue of $\mathbf{R}_{1,2}$. Note that the open-source `WeightWatcher` tool can identify and compute the intra-layer interactions.[66]

Concluding Remarks. `SETOL` as a Semi-Empirical theory merges first-principles methods from `StatMech` and `RMT` with empirical insights from `HTSR` and the open-source `WeightWatcher` tool. It clarifies *how* Heavy-Tailed layer weight matrices can emerge from training on realistic data, and *why* their spectral exponents so reliably predict generalization quality without peeking at training/test sets. In so doing, `SETOL` not only offers new insights into the "*why does it work*" question of deep learning but also suggests a roadmap for improving DNN models by focusing attention on that near-critical subspace of their largest eigenvalues. We are optimistic that future developments along these lines—extending the single-step Renormalization Group analogy, refining

TPL expansions, and systematically diagnosing correlation traps—will yield more robust, data-free metrics for training, fine-tuning, and compressing next-generation neural networks.

Acknowledgements. We would like to thank Matt Lee of Triaxiom Capital and Carl Page of the Anthropocene Institute. We also thank Mirco Milletarì and Michael Mahoney for helpful conversations.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[2] Nobel Prize Organization. The nobel prize in physics 2024, 2024.

[3] Nobel Prize Organization. The nobel prize in chemistry 2024, 2024.

[4] A. Engel. Complexity of learning in artificial neural networks. *Theoretical Computer Science*, 265(1–2):285–306, 2001.

[5] A. Engel and C. P. L. Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, New York, NY, USA, 2001.

[6] E Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257, jan 1988.

[7] H. Sompolinsky, N. Tishby, and H. S. Seung. Learning from examples in large neural networks. *Phys. Rev. Lett.*, 65:1683–1686, Sep 1990.

[8] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992.

[9] E. Levin, N. Tishby, and S. A. Solla. A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE*, 78(10):1568–1574, 1990.

[10] Erin Grant, Sandra Nestler, Berfin Şimşek, and Sara Solla. Statistical physics, Bayesian inference and neural information processing. *arXiv e-prints*, page arXiv:2309.17006, September 2023.

[11] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[12] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79(8):2554–2558, 1982.

[13] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[14] G. E. Hinton and T. J. Sejnowski. Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, and CORPORATE PDP Research Group, editors, *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, pages 282–317. MIT Press, 1986.

[15] W. A. Little. The existence of persistent states in the brain. *Math. Biosci.*, 19:101–120, 1974.

[16] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. USA*, 116:15849–15854, 2019.

[17] Marco Loog, Tom Viering, Alexander Mey, and David MJ Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625, 2020.

[18] M. Opper. Learning to generalize. In D. Baltimore, editor, *Frontiers of Life: Intelligent Systems*, pages 763–775. Academic Press, Cambridge, 2001.

[19] David A. Roberts, Sho Yaida, and Boris Hanin. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, 2022.

[20] V. Vapnik, E. Levin, and Y. Le Cun. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6(5):851–876, 1994.

[21] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65(2):499–556, 1993.

[22] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2):195–236, 1996.

[23] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. Technical Report Preprint: arXiv:1703.11008, 2017.

[24] C. H. Martin and M. W. Mahoney. Post-mortem on a deep learning contest: a Simpson's paradox and the complementary roles of scale metrics versus shape metrics. Technical Report Preprint: arXiv:2106.00734, 2021.

[25] C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

[26] C. H. Martin, T. S. Peng, and M. W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(4122):1–13, 2021.

[27] Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E. Gonzalez, Kannan Ramchandran, Charles H. Martin, and Michael W. Mahoney. Evaluating natural language processing models with generalization metrics that do not need access to any training or testing data. Technical Report Preprint: arXiv:2202.02842, 2022.

[28] Y. Yang, R. Theisen, L. Hodgkinson, J. E. Gonzalez, K. Ramchandran, C. H. Martin, and M. W. Mahoney. Test accuracy vs. generalization gap: Model selection in NLP without accessing training or testing data. In *Proceedings of the 29th Annual ACM SIGKDD Conference*, pages 3011–3021, 2023.

[29] Yefan Zhou, TIANYU PANG, Keqin Liu, Charles Martin, Michael W Mahoney, and Yaoqing Yang. Temperature balancing, layer-wise weight analysis, and neural network training. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 63542–63572. Curran Associates, Inc., 2023.

[30] Haiquan Lu, Yefan Zhou, Shiwei Liu, Zhangyang Wang, Michael W. Mahoney, and Yaoqing Yang. Alphapruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 0000–0000, 2024.

[31] John Negele. Hans Bethe and the theory of nuclear matter. *Physics Today*, 58(10):58, 2005.

[32] D. Ivanenko. The proton-neutron hypothesis of atomic nuclei. *Nature*, 129:798, 1932.

[33] Maria Goeppert-Mayer. On closed shells in nuclei. ii. *Physical Review*, 75(10):1969–1970, 1949.

[34] J. Hans D. Jensen, Otto Haxel, and Hans Suess. On the "magic numbers" in nuclear structure. *Physical Review*, 75:1766, 1949.

[35] Eugene Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955.

[36] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

[37] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller. Random matrix theories in quantum physics: Common concepts. *Physics Reports*, 299:190, 1998.

[38] A. Zee. Law of addition in random matrix theory. *Nuclear Physics B*, 474(3):726–744, September 1996.

[39] Melih K. Sener and Klaus Schulten. General random matrix approach to account for the effect of static disorder on the spectral properties of light harvesting systems. *Physical Review E*, 65(3):031916, 2002. Received 6 June 2001; revised manuscript received 29 August 2001; published 6 March 2002.

[40] S. Gallucio, J.-P. Bouchaud, and M. Potters. Rational decisions, random matrices and spin glasses. *Physica A*, 259:449–456, 1998.

[41] R. Cherrier, D. S. Dean, and A. Lefèvre. Role of the interaction matrix in mean-field spin glass models. *Physical Review E*, 67(4), April 2003.

[42] Rudolph Pariser and Robert G. Parr. A semi-empirical theory of the electronic spectra and electronic structure of complex unsaturated molecules. i. *The Journal of Chemical Physics*, 21(3):466–471, 1953.

[43] J. Hubbard. Electron correlations in narrow energy bands. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 276(1365):238–257, 1963.

[44] Michael J. S. Dewar and Walter Thiel. Ground states of molecules. 38. the mindo/3 method. approximations and parameters. *Journal of the American Chemical Society*, 97(16):4899–4907, 1975.

[45] John Ridley and Michael C. Zerner. Intermediate neglect of differential overlap spectroscopy: a reexamination using a modified neglect of differential overlap approach. *Theoretica Chimica Acta*, 32:111–134, 1973.

[46] James J. P. Stewart. Mopac: A semiempirical molecular orbital program. *Journal of Computer-Aided Molecular Design*, 4:1–103, 1990.

[47] Arieh Warshel and Michael Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, 103(2):227–249, 1976.

[48] J. Hubbard. Calculation of partition functions. *Physical Review Letters*, 3(2):77–78, 1959.

[49] K. F. Freed. Theoretical basis for semiempirical theories. In G.A. Segal, editor, *Semiempirical Methods of Electronic Structure Calculation*, volume 7 of *Modern Theoretical Chemistry*. Springer, 1977.

[50] Karl F. Freed. Is there a bridge between ab initio and semiempirical theories of valence? *Accounts of Chemical Research*, 16:137–144, Mar 1983.

[51] Charles H. Martin and Karl F. Freed. Ab initio computation of semiempirical π-electron methods. v. geometry dependence of hν π-electron effective integrals. *The Journal of Chemical Physics*, 105(4):1437–1450, 1996.

[52] Charles H. Martin. Highly accurate ab initio π-electron hamiltonians for small protonated schiff bases. *The Journal of Physical Chemistry*, 100:14310–14315, 1996.

[53] Charles H Martin. Redesigning semiempirical-like *pi*-electron theory with second order effective valence shell hamiltonian (hv) theory: application to large protonated schiff bases. *Chemical Physics Letters*, 257(3-4):229–237, 1996.

[54] Charles H. Martin and Robert R. Birge. Reparametrizing mndo for excited-state calculations by using ab initio effective hamiltonian theory: Application to the 2,4-pentadien-1-iminium cation. *The Journal of Physical Chemistry A*, 102(5):852–860, 1998.

[55] Nobel Prize Committee. The nobel prize in physics 1982: Kenneth g. wilson. https://www.nobelprize.org/prizes/physics/1982/wilson/, 1982. Accessed: 2024-12-09.

[56] Wolfgang Wenzel and Kenneth G. Wilson. Basis set reduction in hilbert space. *Phys. Rev. Lett.*, 69:800–803, Aug 1992.

[57] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI*, 2018.

[58] J. M. Jumper, K. F. Freed, and T. R. Sosnick. Maximum-likelihood, self-consistent side chain free energies with applications to protein molecular dynamics. Technical Report Preprint: arXiv:1610.07277, 2016.

[59] D. A. Roberts, S. Yaida, and B. Hanin. *The Principles of Deep Learning Theory*. Cambridge University Press, 2021.

[60] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. Technical Report Preprint: arXiv:1711.00165, 2017.

[61] G. Yang. Tensor programs III: Neural matrix laws. Technical Report Preprint: arXiv:2009.10685, 2021.

[62] C. H. Martin and M. W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4284–4293, 2019.

[63] C. H. Martin and M. W. Mahoney. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 20th SIAM International Conference on Data Mining*, 2020.

[64] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24:2613–2626, Sep 1981.

[65] Joseph D. Bryngelson and Peter G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 84:7524–7528, Nov 1987.

[66] Charles H. Martin. Weightwatcher, 2023. Accessed: 2024-06-03.

[67] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[68] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777, 2014.

[69] Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Physical Review E*, 106(5):054124, 2022.

[70] J.-P. Bouchaud and M. Potters. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, 2003.

[71] A. Edelman and Y. Wang. Random matrix theory and its innovative applications. In R. Melnik and I. Kotsireas, editors, *Advances in Applied Mathematics, Modeling, and Computational Science*. Springer, 2013.

[72] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: for Physicists, Engineers and Data Scientists*. Cambridge University Press, 2020.

[73] Francis R Bach and Michael I Jordan. Learning spectral clustering, with application to speech separation. *The Journal of Machine Learning Research*, 7:1963–2001, 2006.

[74] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[75] Per Bak. *How nature works: the science of self-organized criticality*. Oxford University Press, Oxford, UK, 1997.

[76] D. Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer-Verlag, Berlin, 2006.

[77] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83(7):1467–1470, 1999.

[78] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud. Random matrix theory and financial correlations. *Mathematical Models and Methods in Applied Sciences*, pages 109–11, 2005.

[79] Y. Yang, L. Hodgkinson, R. Theisen, J. Zou, J. E. Gonzalez, K. Ramchandran, and M. W. Mahoney. Taxonomizing local versus global structure in neural network loss landscapes. Technical Report Preprint: arXiv:2107.11228, 2021.

[80] C. H. Martin and M. W. Mahoney. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. Technical Report Preprint: arXiv:1901.08278, 2019.

[81] H. Sompolinsky, N. Tishby, and H. S. Seung. Learning from examples in large neural networks. *Phys. Rev. Lett.*, 65:1683–1686, 1990.

[82] T. Tanaka. On dualistic structure involving shannon transform and integrated r-transform. pages 1651–1654, 2007.

[83] T. Tanaka. Asymptotics of Harish-Chandra-Itzykson-Zuber integrals and free probability theory. *J. Phys.: Conf. Ser.*, 95(1):012002, 2008.

[84] M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The heavy-tail phenomenon in SGD. Technical Report Preprint: arXiv:2006.04740, 2020.

[85] Chaim Baskin, Evgenii Zheltonozhkii, Tal Rozen, Natan Liss, Yoav Chai, Eli Schwartz, Raja Giryes, Alexander M Bronstein, and Avi Mendelson. Nice: Noise injection and clamping estimation for neural network quantization. *Mathematics*, 9(17):2144, 2021.

[86] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

[87] A. Engel, C. Van den Broeck, and C. Broeck. *Statistical Mechanics of Learning*. Statistical Mechanics of Learning. Cambridge University Press, 2001.

[88] Hanoch Gutfreund, Haim Sompolinsky, and Daniel Stein. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985.

[89] C. H. Martin and M. W. Mahoney. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. Technical Report Preprint: arXiv:1710.09553, 2017.

[90] B. H. Brandow. Foundations of the nuclear shell model. *Physics Letters*, 4(8):294–296, 1963.

[91] C. H. Martin and M. W. Mahoney. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. Technical Report Preprint: arXiv:1710.09553v1, 2017.

[92] G. Parisi and M. Potters. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A: Mathematical and General*, 28(18):5267–5286, 1995.

[93] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. Technical Report Preprint: arXiv:1611.03530, 2016.

[94] Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[95] F. Rosenblatt. *Principles of Neurodynamics*. Spartan, New York, NY, USA, 1962.

[96] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[97] Joël Bun. *Application of Random Matrix Theory to High Dimensional Statistics*. Phd thesis, Université Paris Saclay (COmUE), 2016. NNT: 2016SACLS245, tel-01400544.

[98] Z. Burda, J. Jurkiewicz, M. A. Nowak, G. Papp, and I. Zahed. Lévy matrices and financial covariances. Technical Report Preprint: arXiv:cond-mat/0103108, 2001.

[99] Z. Burda, J. Jurkiewicz, M. A. Nowak, G. Papp, and I. Zahed. Random Lévy matrices revisited. Technical Report Preprint: arXiv:cond-mat/0602087, 2006.

[100] Z. Burda and J. Jurkiewicz. Heavy-tailed random matrices. Technical Report Preprint: arXiv:0909.5228, 2009.

[101] Jipeng Li, Xueqiong Yuan, and Ercan Engin Kuruoglu. Exploring weight distributions and dependence in neural networks with α-stable distributions. *IEEE Transactions on Artificial Intelligence*, 5(11):5519–5535, November 2024.

[102] LeCun Yann. The mnist database of handwritten digits. *R*, 1998.

[103] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, Oxford, 2001.

[104] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le. Don't decay the learning rate, increase the batch size. Technical Report Preprint: arXiv:1711.00489, 2017.

[105] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

[106] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. Technical Report Preprint: arXiv:1512.03385, 2015.

[107] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. Technical Report Preprint: arXiv:1409.1556, 2014.

[108] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[109] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[110] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[111] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[112] M. Dereziński, F. Liang, and M. W. Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. Technical Report Preprint: arXiv:1912.04533, 2019.

[113] Peijun Qing, Chongyang Gao, Yefan Zhou, Xingjian Diao, Yaoqing Yang, and Soroush Vosoughi. Alphalora: Assigning lora experts based on layer training quality, 2024.

[114] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, Cambridge, UK, 2001.

[115] François Chollet. keras. https://github.com/fchollet/keras, 2015.

[116] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Workshop on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[117] 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.

[118] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[119] R. Cherrier, D. S. Dean, and A. Lefèvre. Role of the interaction matrix in mean-field spin glass models. *Physical Review E*, 67:046112, 2003.

A Appendix

A.1 Data Vectors, Weight Matrices, and Other Symbols

See Table 11 for a summary of various vectors and matrices, including their dimensions; see Table 12 for summary of various various symbols used throughout the text; and see Table 13 for summary of types of "Energies" used throughout the text.

Number of NN Layers	index L	N_L
Number of Data Examples	index μ	$N = N_D$
Number of (input) Features	index i, j	N_f
Actual Data (Matrix)	D	$N_f \times N_D$
Model Data (Matrix)	$\mathcal{D}$	$N_f \times N$
Teacher Perceptron Weight Vector	$\mathbf{t}$	m
Student Perceptron Weight Vector	$\mathbf{s}$	m
Actual Input Data Vector	$\mathbf{x}_\mu$	$N_f \times 1$
Gaussian model of Input Data Vector	$\boldsymbol{\xi}_\mu$	$N_f \times 1$
Actual Input Data Label	y_μ	$+1 \vert -1$
Model Input Data Label	y_μ	$+1 \vert -1$
General Weight Matrix	$\mathbf{W}$	$N \times M$
General Correlation Matrix	$\mathbf{X} = \frac{1}{N}\mathbf{W}^\intercal\mathbf{W}$	$M \times M$
Input Layer Weight Matrix	$\mathbf{W}_1$	$N \times M$
Hidden Layer Weight Matrix	$\mathbf{W}_2$	$N \times M$
Output Layer Weight Matrix	$\mathbf{W}_3$	$M \times 2$
Teacher Weight Matrix	$\mathbf{T}$	$N \times M$
Student Weight Matrix	$\mathbf{S}$	$N \times M$
Student-Teacher Overlap Matrix	$\mathbf{R} = \frac{1}{N}\mathbf{S}^T\mathbf{T}$	$M \times M$
Student Correlation Matrix 1	$\mathbf{A}_1 = \frac{1}{N}\mathbf{S}^\intercal\mathbf{S}$	$M \times M$
Student Correlation Matrix 2	$\mathbf{A}_2 = \frac{1}{N}\mathbf{S}\mathbf{S}^\intercal$	$N \times N$
ECS Student Correlation Matrix 1	$\tilde{\mathbf{A}}_1$	$\tilde{M} \times \tilde{M}$
ECS Student Correlation Matrix 2	$\tilde{\mathbf{A}}_2$	$N \times N$
ECS Teacher Correlation Matrix	$\tilde{\mathbf{X}}$	$\tilde{M} \times \tilde{M}$

Table 11: Summary of of various vectors and matrices, including their dimensions.

Perceptron Student-Teacher (ST) Overlap	$R = \frac{1}{N}\mathbf{s}^\top\mathbf{t} = \frac{1}{N}\sum_i s_i t_i$
Student-Teacher (ST) Overlap Operator	$\mathbf{R} = \frac{1}{N}\mathbf{S}^\top\mathbf{T}$
Matrix Generalized ST Overlap	$\frac{1}{N^2}\,\mathrm{Tr}\,[\mathbf{R}^\top\mathbf{R}]$
Student-Teacher Self-Overlap	$\eta(\boldsymbol{\xi}) = \mathbf{y}_T^\top\mathbf{y}_S$
ℓ_2-Energy Difference or ℓ_2-Error	$\Delta E_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi})$
ℓ_2-Energy Difference Operator Form	$\boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n) := \sum_{\boldsymbol{\xi}}\Delta E_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi})$
ℓ_2-Energy Difference Matrix Operator Form	$\boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{S},\mathbf{T}) := \mathbf{I}_M - \frac{1}{N}\mathbf{S}^\top\mathbf{T}$
Annealed Error Potential	$\epsilon(R) = \epsilon(S,T) = \langle\boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n)\rangle_{\overline{\boldsymbol{\xi}^N}}$
Linear Perceptron $\epsilon(R)$ at high-T, large-N	$\epsilon(R) = 1 - R$
Annealed Approximation (AA)	$\langle\ln Z\rangle_{\overline{\boldsymbol{\xi}^N}} = \ln\langle Z\rangle_{\overline{\boldsymbol{\xi}^N}}$
Annealed Hamiltonian	$H^{an}(\mathbf{w})$
Annealed Hamiltonian at high-T	$H^{an}_{hT}(\mathbf{w}) = \epsilon(\mathbf{w})$
Average Student-Teacher Generalization Error	$\bar{\mathcal{E}}^{ST}_{gen} = \langle\epsilon(R)\rangle^\beta_{\mathbf{s}}$
Average Student-Teacher Generalization Accuracy	$1 - \bar{\mathcal{E}}^{ST}_{gen} = \langle\eta(R)\rangle^\beta_{\mathbf{s}}$
Matrix Layer Quality-Squared	$\bar{\mathcal{Q}}^2 = \langle\mathbf{R}^\top\mathbf{R}\rangle^\beta_{\mathbf{S}}$
Equivalent Notation for Averages	$\langle\cdots\rangle_A = \int\cdots d\mu(\mathbf{A}) = \mathbb{E}_{\mathbf{A}}[\cdots]$
Projection Operator onto `ECS`	$\mathbf{P}^{ecs} := \sum\|\tilde{\lambda}_i\rangle\langle\tilde{\lambda}_i\|,\ i = 1\cdots\tilde{M}$
Average over `ECS` Student Correlation Matrices	$\langle\cdots\rangle^\beta_{\tilde{\mathbf{A}}} = \int\cdots d\mu(\tilde{\mathbf{A}}) = \mathbb{E}_{\tilde{\mathbf{A}}}[\cdots]$
`TRACE-LOG` Determinant Relation	$\mathrm{Tr}\,[\ln\mathbf{A}] = \ln\det\mathbf{A}$
`TRACE-LOG` (Volume Preserving) Condition	$\mathrm{Tr}\,[\ln\mathbf{A}] = 0$ or $\det\mathbf{A} = 1$
Effective Correlation Measure Transform	$d\mu(\mathbf{S}) \to d\mu(\tilde{\mathbf{A}})$
HCIZ Integral (Tanaka's Notation)	$\mathbb{E}_{\tilde{\mathbf{A}}}[\exp(\beta\,\mathrm{Tr}\,[\mathbf{T}^\top\mathbf{A}\mathbf{T})]$
HCIZ Integral	$\left\langle\exp\left(N\,\beta\,\mathrm{Tr}\left[\frac{1}{N}\,\mathbf{T}^\top\,\tilde{\mathbf{A}}_2\,\mathbf{T}\right]\right)\right\rangle_{\tilde{\mathbf{A}}}$
Layer Quality Generating Function	$\beta\boldsymbol{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg1} := N\beta\sum_{\mu=1}^{\tilde{M}}\int_{\tilde{\lambda}_{\min}}^{\tilde{\lambda}_\mu} dz R(z)$
Norm Generating Function	$G_A(\gamma) = \int_{\lambda^{ECS}_{min}}^{\tilde{\lambda}} R_A(z)dz$
Eigenvalue for $\mathbf{X} = \frac{1}{N}\mathbf{W}^\top\mathbf{W}$	λ or λ_i for $i = 1\cdots M$
Power Law ESD Tail for $\mathbf{X}$	$\rho_{tail}(\lambda) \sim \lambda^{-\alpha}$
Effective Correlation Space ESD Tail for $\mathbf{X}$	$\rho^{ECS}_{tail}(\tilde{\lambda}),\ \mathrm{Tr}\left[\ln\prod_{j=1}^{\tilde{M}}\tilde{\lambda}_j\right] = 0$
Schatten Norm	$\|\mathbf{X}\|^\alpha_\alpha = \sum_j\lambda_j^\alpha$
ECS Tail (or Trace) Norm	$\frac{1}{\tilde{M}}\sum_i^{\tilde{M}}\tilde{\lambda}_i,\ \ \tilde{\lambda}_i \in \rho^{ECS}_{tail}(\tilde{\lambda})$
Spectral Norm	$\|\mathbf{X}\|_\infty = \lambda_{max}$
`WeightWatcher` Start of PL Tail	$\lambda^{PL}_{min} = \lambda_{min}$
Start of `ECS` Tail	$\lambda^{ECS}_{min} = \lambda^{\|detX\|=1}_{min}$
`ECS`-PL Gap between start of tails	$\Delta\lambda_{min} := \lambda^{ECS}_{min} - \lambda^{\|detX\|=1}_{min}$
`WeightWatcher Alpha` (layer) quality metric	α
`WeightWatcher AlphaHat` (layer) quality metric	$\hat{\alpha} = \alpha\log_{10}(\lambda_{max})$

Table 12: Summary of of various various symbols used throughout the text.

Explanation	Examples	Refs
Energy Landscape or Output function	E	Sec. 2.1
The output of the NN given a single input data point.		1,74,??,97
Energy Difference or Student-Teacher (ST) Error	ΔE, $\mathbf{\Delta E}$	Sec. 4, A.2
The difference between the output of a Student NN and its prescribed label y for a single data point	$\Delta E \coloneqq y - E_S$	19
And as the total error for a sample of N data points	$\mathbf{\Delta E} = \sum \Delta E$	22
Or between the outputs of the Student and the Teacher NNs.	$\mathbf{\Delta E} \coloneqq \sum E_S - E_T$	83,98,180
Annealed Hamiltonian (and Potentials)	$H^{an}\ (\mathcal{E}(R), \epsilon(R))$	Sec. 4.2.4,A.2
The Annealed Error Potential (for the Error) is defined as:	$\epsilon(R) = \frac{1}{N}\mathcal{E}(R) = \langle \mathbf{\Delta E} \rangle_{\overline{\boldsymbol{\xi}^N}}$	94
The Annealed Hamiltonian (for the Error):	$\beta H^{an} \coloneqq \frac{1}{N} \ln \langle e^{-\beta \mathbf{\Delta E}} \rangle_{\boldsymbol{\xi}^N}$	44, 165
At high-T, the relation between H^{an}_{hT} and $\epsilon(R)$ is	$H^{an}_{hT}(R) \coloneqq \epsilon(R)$	49
The full ST model Hamiltonian:	$\beta H^{an}(\beta, R) \coloneqq \frac{1}{2}\ln[1 + 2\beta(1-R)]$	172
At high-T, the ST model Hamiltonian is:	$\beta H^{an}_{hT}(R) \coloneqq \beta(1-R)$	94, 173
The matrix-generalized ST model Hamiltonian:	$\beta H^{an}(\mathbf{R}) \coloneqq \frac{1}{2}\ln[1 + 2\beta(\mathbf{I}_M - \mathbf{R})]$	200
At high-T, the matrix-generalized ST model Hamiltonian is:	$H^{an}_{hT}(\mathbf{R}) \coloneqq \mathbf{I}_M - \mathbf{R}$	202
The Layer Quality-Squared Hamiltonian (for the Accuracy):	$\mathbf{H}_{\bar{\mathcal{Q}}^2} \coloneqq \mathbf{R}^\intercal \mathbf{R}$	16, 214
Different Average Model Errors	$\bar{\mathcal{E}}$	Sec. 4.2.5
Empirical Training, Teacher, and Test Errors	$\bar{\mathcal{E}}^{emp}_{train}$, $\bar{\mathcal{E}}^{T} \approx \bar{\mathcal{E}}^{emp}_{gen}$	75,76,80
Empirical Generalization Gap	$\bar{\mathcal{E}}^{emp}_{gap} \coloneqq \bar{\mathcal{E}}^{emp}_{train} - \bar{\mathcal{E}}^{emp}_{gen}$	81
Student-Teacher Training and Generalization Errors	$\bar{\mathcal{E}}^{ST}_{train}$, $\bar{\mathcal{E}}^{ST}_{gen}$	168,171
Neural Network (MLP) Training and Generalization Errors, the (abstract) matrix generalization of ST error	$\bar{\mathcal{E}}^{NN}_{train}$, $\bar{\mathcal{E}}^{NN}_{gen}$	
Average Training and/or Generalization Error	$\bar{\mathcal{E}}_{train}, \bar{\mathcal{E}}_{gen}$	Sec. 4.2.5
In the AA and at High-T, these are the same, and are just the Thermal Average of $\epsilon(R)$	$\bar{\mathcal{E}}^{an,hT}_{train} = \bar{\mathcal{E}}^{an,hT}_{gen} = \langle \epsilon(R) \rangle^{\beta}_{\mathbf{s}}$	??,??
For the ST model, we always assume AA and High-T	$\bar{\mathcal{E}}^{ST}_{gen} = \bar{\mathcal{E}}^{an,hT}_{gen}$	
Likewise, when generalizing $\bar{\mathcal{E}}^{ST}_{gen}$ to matrices,	$\bar{\mathcal{E}}^{ST}_{gen} \to \bar{\mathcal{E}}^{NN}_{gen} = \bar{\mathcal{E}}^{an,hT}_{gen}$	
Layer Qualities	$\bar{\mathcal{Q}}$, $\bar{\mathcal{Q}}^2$	Sec. 5.2.1, A.3
For the ST Perceptron, $\bar{\mathcal{Q}}^{ST}$ is the generalization accuracy	$\bar{\mathcal{Q}}^{ST} \coloneqq 1 - \bar{\mathcal{E}}^{ST}_{gen} = 1 - \langle \epsilon(R) \rangle^{\beta}_{\mathbf{s}}$	
in terms of the Self-Overlap $\eta(R)$	$\bar{\mathcal{Q}}^{ST} \coloneqq \langle \eta(R) \rangle^{\beta}_{\mathbf{s}}$	
In the AA, and at high-T, $\langle \epsilon(R) \rangle^{\beta}_{\mathbf{s}} = 1 - R$	$\bar{\mathcal{Q}}^{ST} \coloneqq 1 - \bar{\mathcal{E}}^{an,ht}_{gen} = \langle R \rangle^{\beta}_{\mathbf{s}}$	96
For an MLP / NN, we approximate the total accuracy as a product of layer qualities $\bar{\mathcal{Q}}$ (in the AA, at high-T)	$\bar{\mathcal{Q}}^{NN} \coloneqq \prod \bar{\mathcal{Q}}^{NN}_{L}$	7
For a matrix, the Layer Quality-Squared $\bar{\mathcal{Q}}^2$ is an HCIZ integral	$\bar{\mathcal{Q}}^2 \coloneqq \langle \mathbf{R}^\intercal \mathbf{R} \rangle^{\beta}_{\mathbf{S}}$	107
We approximate $\bar{\mathcal{Q}}$ using the quality squared	$\bar{\mathcal{Q}} \coloneqq \sqrt{\bar{\mathcal{Q}}^2} \approx Q^{NN}_{L}$	9,108

Table 13: Summary of types of "Energies," with simplified examples of the notation, and references to definitions. This is a guide to understanding how the various Energies are defined and used. See the text for exact definitions, dependent variables, etc.

A.2 Summary of the Statistical Mechanics of Generalization (SMOG)

In this section, we derive the Annealed Hamiltonian for two variants of the ST model, in the high-T limit: in Appendix A.2.1, we derive an expression for $H^{an}(R)$ for the ST Perceptron model, when the students and teachers are modeled as N-vectors $\mathbf{w}$ (as in [8]); and in Appendix A.2.2, we derive an expression for $H^{an}(\mathbf{R})$ for Matrix-Generalized case, i.e., when the students and teachers are modeled as $N \times M$ matrices $\mathbf{W}$ (as our SETOL requires). From these, we will obtain expressions for the Average ST Model Generalization Accuracy $\bar{\mathcal{E}}^{ST}_{gen}$ and the Average NN Model Generalization Accuracy $\bar{\mathcal{E}}^{NN}_{gen}$, as well as for the corresponding data-averaged errors. Although the functional form for these quantities will be the same for the vector case and the matrix case, there are several important differences in the derivation of $H^{an}(\mathbf{R})$, most notably having to do with a normalization for the weight matrix.

A.2.1 Annealed Hamiltonian $H^{an}(R)$ when Student and Teachers are Vectors

In this section, we derive an expression for the Annealed Hamiltonian $H^{an}(R)$, in the AA and the high-T approximation, when student and teachers are are modeled as N-vectors. From this, we obtain an expression for the data-averaged ST error $\epsilon(R)$, which is the same as the expression given in Eqn. 94.

The procedure starts by computing the associated quenched average of the Free Energy, defined for the model error as

$$\begin{aligned}\langle -\beta F \rangle_{\overline{\boldsymbol{\xi}^N}} &:= \langle \ln Z \rangle_{\overline{\boldsymbol{\xi}^N}} \\ &= \left\langle \ln \int d\mu(\mathbf{s}) e^{-\beta \boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n)} \right\rangle_{\overline{\boldsymbol{\xi}^N}} \\ &= \frac{1}{N} \int d\mu(\boldsymbol{\xi}^N) \ln \int d\mu(\mathbf{s}) e^{-\beta \boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n)}, \end{aligned} \tag{162}$$

where $d\mu(\boldsymbol{\xi}^N) := \prod_{i=1}^{N} d\boldsymbol{\xi}_i P(\boldsymbol{\xi}^N)$ (see Eqn. ??) and where the data-dependent ST error, $\boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n)$, is defined in Eqn. 83, with an $\mathcal{L} = \ell_2$ loss. [43] If we apply the AA (see Eqn. 41) to Eqn. 162, then we obtain

$$\langle -\beta F \rangle_{\overline{\boldsymbol{\xi}^N}} \simeq \ln \frac{1}{N} \int d\mu(\boldsymbol{\xi}^N) \int d\mu(\mathbf{s}) e^{-\beta \boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n)}. \tag{163}$$

Notice that we have interchanged the logarithm (ln) and and the data average (the "disorder average") $\langle \cdots \rangle_{\overline{\boldsymbol{\xi}^N}}$ over the data; this is the essence of the AA, as it lets the disorder fluctuate rather than forcing the system to be quenched to the data. We will now switch the order of integration in Eqn. 163, giving

$$\langle -\beta F \rangle_{\overline{\boldsymbol{\xi}^N}} \simeq \ln \int d\mu(\mathbf{s}) \frac{1}{N} \int d\mu(\boldsymbol{\xi}^N) e^{-\beta \boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n))}. \tag{164}$$

We now recall the definition of the Annealed Hamiltonian, $H^{an}(R)$ (see Eqn. 44 in Section 4.2, which is analogous to Eqn. (2.31) of [8]):

$$\beta H^{an}(R) = \beta H^{an}(\beta, \mathbf{s}, \mathbf{t}) := -\frac{1}{N} \ln \int d\mu(\boldsymbol{\xi}^N) e^{-\beta \boldsymbol{\Delta}\mathbf{E}_{\ell_2}(\mathbf{s},\mathbf{t},\boldsymbol{\xi}^n)}. \tag{165}$$

[43] Also, recall that the Teacher T is fixed and is not learned, so we do not integrate over $d\mu(\mathbf{t})$. In fact, for the (vector) Perceptron model, the Teacher T weights are simply subsumed into the overlap R, and even in the more general cases, such as non-Linear/Boolean Perceptron, in the full Replica calculations, etc. See the original literature for more details. [18, 8, 114] as well as [89].

where we have denoted the Hamiltonian as $H^{an}(\beta, \mathbf{s}, \mathbf{t})$ to indicate the explicit dependence on β, and we have added β to the R.H.S. to because the L.H.S. is unitless. Using this definition, we can express the Annealed Partition Function, Z_N^{an}, in the AA in terms of the Annealed Hamiltonian $H^{an}(R)$ (as in Eqn. 46 in Section 4.2, and as in Eqn. (2.31) of [8]:

$$Z_N^{an} := N\langle Z \rangle_{\overline{\boldsymbol{\xi}^N}} = \int d\mu(\mathbf{s}) e^{-N\beta H^{an}(\beta, \mathbf{s}, \mathbf{t})}. \tag{166}$$

Following Section 4.2.5, we can write the *Average Model Training Error* $\bar{\mathcal{E}}_{train}^{ST}$, in the AA, in terms of Annealed Partition Function, Z_N^{an}:

$$\bar{\mathcal{E}}_{train}^{ST} := -\frac{1}{N}\frac{\partial}{\partial \beta} \ln Z_N^{an} \tag{167}$$

This now lets us write the Average Model Training Error $\bar{\mathcal{E}}_{train}^{ST}$ in the AA in terms of the Annealed Hamiltonian $H^{an}(R)$:

$$\begin{aligned} \bar{\mathcal{E}}_{train}^{ST} &= \frac{1}{Z_N^{an}} \int d\mu(\mathbf{s}) \frac{\partial \beta H^{an}}{\partial \beta} e^{-N\beta H^{an}(\beta, \mathbf{s}, \mathbf{t})} \\ &= \frac{1}{Z_N^{an}} \int d\mu(\mathbf{s}) \langle \boldsymbol{\Delta \mathbf{E}}_{\ell_2}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \rangle_{\overline{\boldsymbol{\xi}^N}} e^{-N\beta H^{an}(\beta, \mathbf{s}, \mathbf{t})}, \end{aligned} \tag{168}$$

where $\langle \mathbf{E}_{\ell_2}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \rangle_{\overline{\boldsymbol{\xi}^N}}^{\beta}$ is a Thermal Average but defined over the specific ST error in the AA for the chosen set of training data $\boldsymbol{\xi}^{train} = \boldsymbol{\xi}^N$, and is denoted by

$$\langle \boldsymbol{\Delta \mathbf{E}}_{\ell_2}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \rangle_{\overline{\boldsymbol{\xi}^N}}^{\beta} := \frac{\partial \beta H^{an}}{\partial \beta}. \tag{169}$$

This is analogous to defining the average error $\boldsymbol{\Delta \mathbf{E}}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)$ as a Thermal Average, but as one over the data $\boldsymbol{\xi}^n$ instead of the weights. This can be seen by expanding Eqn. 44, setting $\mathbf{w} = \mathbf{s}$, fixing $\mathbf{t}$ (implicitly), and taking the partial derivative:

$$\begin{aligned} \frac{\partial \beta H^{an}}{\partial \beta} &= \frac{\partial}{\partial \beta}\left(-\frac{1}{N} \ln \int d\mu(\boldsymbol{\xi}^n) e^{-\beta \boldsymbol{\Delta \mathbf{E}}_{\ell_2}(\mathbf{w}, \boldsymbol{\xi}^n)}\right) \\ &= -\frac{1}{N}\frac{\partial}{\partial \beta}\frac{1}{N} \ln \int d\mu(\boldsymbol{\xi}^n) e^{-\beta \boldsymbol{\Delta \mathbf{E}}_{\ell_2}(\mathbf{w}, \boldsymbol{\xi}^n)} \\ &= -\frac{1}{N}\left(\int d\mu(\boldsymbol{\xi}^n) e^{-\beta \boldsymbol{\Delta \mathbf{E}}_{\ell_2}(\mathbf{w}, \boldsymbol{\xi}^n)}\right)^{-1} \frac{\partial}{\partial \beta} \int d\mu(\boldsymbol{\xi}^n) e^{-\beta \boldsymbol{\Delta \mathbf{E}}_{\ell_2}(\mathbf{w}, \boldsymbol{\xi}^n)} \\ &= -\frac{1}{N}\left(\int d\mu(\boldsymbol{\xi}^n) e^{-\beta \boldsymbol{\Delta \mathbf{E}}_{\ell_2}(\mathbf{w}, \boldsymbol{\xi}^n)}\right)^{-1} \int d\mu(\boldsymbol{\xi}^n)(-\boldsymbol{\Delta \mathbf{E}}_{\mathcal{L}}(\mathbf{w}, \boldsymbol{\xi}^n)) e^{-\beta \boldsymbol{\Delta \mathbf{E}}_{\ell_2}(\mathbf{w}, \boldsymbol{\xi}^n)} \end{aligned} \tag{170}$$

We can also write the Model Generalization Accuracy $\bar{\mathcal{E}}_{gen}^{ST}$ as Boltzmann-weighted average of $\epsilon(R)$, weighted by $H^{an}(\beta, S; T)$ (as in (2.32) in [8]), as:

$$\bar{\mathcal{E}}_{gen}^{ST} = \frac{1}{Z_N^{an}} \int d\mu(\mathbf{s}) \epsilon(R) e^{-N\beta H^{an}(\beta, \mathbf{s}, \mathbf{t})}, \tag{171}$$

where $\epsilon(R) = \epsilon(\mathbf{s})$ is the average ST error, for a fixed Teacher T, averaged over *all* possible data inputs, i.e., not just over the specific training data. (Note that we have dropped the subscript *train* on $\boldsymbol{\xi}$ since it is clear from the context.)

In the high-T (small β) limit, the two model errors become formally equivalent (i.e., $\bar{\mathcal{E}}_{train}^{ST} = \bar{\mathcal{E}}_{gen}^{ST}$ as $T \to \infty$). To show this, consider the Annealed Hamiltonian $H^{an}(R)$, for the Linear Perceptron

with the ℓ_2 loss. As shown in Eqn. (C6) of [8], this takes a simple analytic form–in the large-N limit–in terms of the ST overlap R:

$$H^{an}(R) = \frac{1}{2}\ln\left[1 + 2\beta(1-R)\right]. \tag{172}$$

Eqn. 172 holds in the AA, but not in the High-T limit. If we evaluate $\dfrac{\partial H^{an}}{\partial \beta}$ in the High-T (small β) limit, then we can use the approximation $(\ln[1+x] \simeq x + \cdots)$ to obtain the High-T approximation:

$$\beta H^{an}(R) \simeq \beta H^{an}_{hT}(R) := \beta(1-R), \quad \beta \text{ small}. \tag{173}$$

where we now see that $H^{an}_{hT}(R)$ no longer explicitly depends on β. By Eqn. 50, this gives

$$\epsilon(R) = \langle \mathbf{E}_{\ell_2}(\mathbf{s}, \mathbf{t}, \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}^N} \simeq 1 - R \text{ as } N \to \infty, \tag{174}$$

which we recognize as the same as the data-averaged ST error $\epsilon(R)$ in Eqn. 94.

A.2.2 Annealed Hamiltonian $H^{an}(\mathbf{R})$ for the Matrix-Generalized ST Error

In this section, we derive an expression for our matrix generalization of the Annealed Hamiltonian of the Linear Perceptron, in the AA and the high-T approximation, when student and teachers are are modeled as $N \times M$ matrices $\mathbf{W}$, i.e., $H^{an}(R) \to H^{an}(\mathbf{R})$. See Eqn. **??**, which has the same form as Eqn. 165 for the vector case. From this, we obtain an expression for the data-averaged ST error $\epsilon(R)$, again when the student and teachers are are modeled as matrices. There is a subtle normalization issue here, about which we need to be careful. However, when we normalize appropriately, we will obtain an expression for "data-averaged ST error" (i.e., Annealed Error Potential) $\epsilon(R)$ that is of the same form as we obtained in the vector case (as given in Eqn. 174 and Eqn. 94). The difference will be that in the vector case we take $R = \frac{1}{N}\mathbf{s}^\top\mathbf{t}$, while in the matrix case we take $R = \frac{1}{N}\mathbf{S}^\top\mathbf{T}$.

We will need to evaluate an average over the N random M-dimensional training data vectors $\boldsymbol{\xi}$, which are i.i.d Gaussian with 0 mean and σ^2 variance:

$$\|\boldsymbol{\xi}\|^2 := \frac{1}{N}\sum_{\mu=1}^{N} \boldsymbol{\xi}_\mu \boldsymbol{\xi}_\mu^\top = \sigma^2 \mathbf{I}_M, \tag{175}$$

where $\boldsymbol{\xi}_\mu$ is a vector of length M, and $\mathbf{I}_M$ is an $M \times M$ identity matrix. The expected value of the squared norm is:

$$\mathbb{E}[\|\boldsymbol{\xi}_\mu\|^2] = M\sigma^2. \tag{176}$$

If we let $\sigma^2 \sim \frac{1}{M}$, then $\mathbb{E}[\|\boldsymbol{\xi}_\mu\|^2] = 1$, i.e., the data vectors can be normalized to 1. Let the probability distribution over the N data vectors be

$$\begin{aligned} P(\boldsymbol{\xi}^N) &= \prod_{\mu=1}^{N}\left(\frac{1}{\sqrt{(2\pi\sigma^2)^M}}\right) e^{-\frac{\|\boldsymbol{\xi}_\mu\|^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sqrt{(2\pi\sigma^2)^M}}\right)^N \exp\left[-\frac{M}{2}\sum_{\mu=1}^{N}\|\boldsymbol{\xi}_\mu\|^2\right] \\ &= \mathcal{N}\exp\left[-\frac{M}{2}\sum_{\mu=1}^{N}\|\boldsymbol{\xi}_\mu\|^2\right], \end{aligned} \tag{177}$$

where $M = N_f$ is the number of features in the data, where the normalization $\mathcal{N}$ is

$$\mathcal{N} := \left(\frac{1}{\sqrt{(2\pi\sigma^2)^M}}\right)^N = \left(\frac{M}{2\pi}\right)^{MN/2}. \tag{178}$$

The Total Data Sample Error ($\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T})$) and the Matrix Normalization First, let us express the matrix-generalized Total Data Sample Error, $\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T})$, for a single layer, in operator form (for each training example)

$$\frac{1}{N}\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T}) \coloneqq \mathrm{Tr}\left[\mathbf{I}_M - \frac{1}{N}\mathbf{S}^\top\mathbf{T}\right] = M - \mathrm{Tr}\left[\mathbf{R}\right] \tag{179}$$

where $\mathbf{I}_M$ is a diagonal matrix of dimension M. Notice that $\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T})$ scales as $N \times M$, the total number of parameters in the system. Also, importantly, when all the overlaps are perfect, then the error is zero, i.e. if $\mathrm{Tr}\left[\mathbf{R}\right] = M$ then $\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T}) = 0$.

We can define the data-dependent form (i.e., in the basis of the data $\boldsymbol{\xi}$) as

$$\begin{aligned}\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T},\boldsymbol{\xi}) &\coloneqq \sum_{\mu=1}^{N}(\boldsymbol{\xi}^\mu)^\top\left(\mathbf{I}_M - \frac{1}{N}\mathbf{S}^\top\mathbf{T}\right)\boldsymbol{\xi}^\mu \\ &= \sum_{\mu=1}^{N}\sum_{i,j=1}^{M}\boldsymbol{\xi}_i^\mu\left(\delta_{ij} - \frac{1}{N}[\mathbf{S}^\top\mathbf{T}]_{ij}\right)\boldsymbol{\xi}_j^\mu.\end{aligned} \tag{180}$$

The Annealed Hamiltonian (per-parameter, $h^{an}(\mathbf{R})$) The definition of the Annealed Hamiltonian, $H^{an}(\mathbf{R})$, for the matrix-generalized case must be extended to account for the M parameters per training example. We then have that the total energy is then the sum of the entries of M independent parts, as expected by Size-Consistency:

$$\mathrm{Tr}\left[H^{an}(\mathbf{R})\right] = M\,\mathrm{Tr}\left[h^{an}(\mathbf{R})\right] \tag{181}$$

where the Annealed Hamiltonian per-parameter, $h^{an}(\mathbf{R})$, is obtained from Eqn. 165 as

$$\begin{aligned}\beta h^{an}(\mathbf{R}) &\coloneqq -\frac{1}{N}\ln\int \mathcal{D}\boldsymbol{\xi}^N\, e^{-\beta\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T})}P(\boldsymbol{\xi}^N) \\ &= -\frac{1}{N}\ln\mathbb{I}_H,\end{aligned} \tag{182}$$

where

$$\mathbb{I}_H \coloneqq \int \mathcal{D}\boldsymbol{\xi}^N\, e^{-\beta\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T})}P(\boldsymbol{\xi}^N). \tag{183}$$

That is, $h^{an}(\mathbf{R})$ represents the Energy or Error each of the M parameters contributes (integrated over the N training examples $\boldsymbol{\xi}^N$).

The goal will be to derive the high-Temperature Annealed Hamiltonian, $H^{an}_{hT}(\mathbf{R})$, which is now defined such that:

$$\mathrm{Tr}\left[H^{an}_{hT}(\mathbf{R})\right] \coloneqq M\,\mathrm{Tr}\left[h^{an}_{hT}(\mathbf{R})\right] \tag{184}$$

If we were to evaluate the trace of $h^{an}_{hT}(\mathbf{R})$, then we could also define a matrx-generalized Annealed Error Potential,

$$\epsilon(\mathbf{R}) \coloneqq \mathrm{Tr}\left[h^{an}_{hT}(\mathbf{R})\right], \tag{185}$$

which would be like a mean-field potential, but we need something different for the matrix case.

To evaluate the integral, notice that $\mathbb{I}_H$ is really just an average over i.i.d. data, and so it is just a product over N independent terms

$$\mathbb{I}_H \coloneqq \int \mathcal{D}\boldsymbol{\xi}^N e^{-\beta\mathbf{\Delta E}_{\ell_2}(\mathbf{S},\mathbf{T})}P(\boldsymbol{\xi}^N) \to \left[\int \mathcal{D}\boldsymbol{\xi}\,[\cdots]\right]^N, \tag{186}$$

as in Eqn. 190 below. Moreover, when taking $\ln \mathbb{I}_H$, the N term pulls down and become a prefactor

$$-\ln \mathbb{I}_H = -\ln \Big[\int \mathcal{D}\boldsymbol{\xi}\, [\cdots] \Big]^N = -N \ln \Big[\int \mathcal{D}\boldsymbol{\xi}\, [\cdots] \Big]. \tag{187}$$

Thus, as with the vector case, $H^{an}(\mathbf{R})$ is like a mean-field average over the data $\boldsymbol{\xi}$, independent of the sample size N. Also, since the final result must scale as $N \times M$, the integral should scale as M, i.e., $[\int \mathcal{D}\boldsymbol{\xi}\, [\cdots]\,] \sim M$.

If we substitute $\mathbf{\Delta E}_{\ell_2}(\mathbf{S}, \mathbf{T}, \boldsymbol{\xi})$, Eqn. 180, into the integral $\mathbb{I}_H$, Eqn. 183, then we obtain

$$\begin{aligned}
\mathbb{I}_H &= \int \mathcal{D}\boldsymbol{\xi}^N \exp\left(-\beta \sum_{\mu=1}^{N} (\boldsymbol{\xi}^\mu)^\top \left(\mathbf{I}_M - \frac{1}{N} \mathbf{S}^\top \mathbf{T} \right) \boldsymbol{\xi}^\mu \right) P(\boldsymbol{\xi}^N) \qquad (188) \\
&= \int \mathcal{D}\boldsymbol{\xi}^N \exp\left(-\beta \sum_{\mu=1}^{N} (\boldsymbol{\xi}^\mu)^\top \left(\mathbf{I}_M - \frac{1}{N} \mathbf{S}^\top \mathbf{T} \right) \boldsymbol{\xi}^\mu \right) \mathcal{N} \exp\left(-\sum_{\mu=1}^{N} \frac{\|\boldsymbol{\xi}^\mu\|^2}{2\sigma^2} \right) \\
&= \mathcal{N} \int \mathcal{D}\boldsymbol{\xi}^N \exp\left(-\beta \sum_{\mu=1}^{N} (\boldsymbol{\xi}^\mu)^\top (\mathbf{I}_M - \tfrac{1}{N} \mathbf{S}^\top \mathbf{T}) (\boldsymbol{\xi}^\mu) - \sum_{\mu=1}^{N} \frac{\|\boldsymbol{\xi}^\mu\|^2}{2\sigma^2} \right) \\
&= \mathcal{N} \int \mathcal{D}\boldsymbol{\xi}^N \exp\left(-\frac{1}{2\sigma^2} \sum_{\mu=1}^{N} 2\beta\sigma^2 (\boldsymbol{\xi}^\mu)^\top (\mathbf{I}_M - \tfrac{1}{N} \mathbf{S}^\top \mathbf{T}) (\boldsymbol{\xi}^\mu) + \|\boldsymbol{\xi}^\mu\|^2 \right) \\
&= \mathcal{N} \int \mathcal{D}\boldsymbol{\xi}^N \exp\left(-\frac{1}{2\sigma^2} \sum_{\mu=1}^{N} (\boldsymbol{\xi}^\mu)^\top [2\beta\sigma^2 (\mathbf{I}_M - \tfrac{1}{N} \mathbf{S}^\top \mathbf{T}) + \mathbf{I}_M] (\boldsymbol{\xi}^\mu) \right). \qquad (189)
\end{aligned}$$

By combining the exponents, we obtain

$$\begin{aligned}
\mathbb{I}_H &= \mathcal{N} \int \mathcal{D}\boldsymbol{\xi}^N \exp\left[-\frac{1}{2\sigma^2} \sum_{\mu=1}^{N} (\boldsymbol{\xi}^\mu)^\top (\mathbf{M})\, \boldsymbol{\xi}^\mu \right] \\
&= \mathcal{N} \int \mathcal{D}\boldsymbol{\xi} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{\xi})^\top (\mathbf{M})\, \boldsymbol{\xi} \right]^N, \qquad (190)
\end{aligned}$$

where $\mathbf{M} = 2\beta\sigma^2 (\mathbf{I}_M - \frac{1}{N} \mathbf{S}^\top \mathbf{T}) + \mathbf{I}_M$ is an $M \times M$ matrix.

We now use the familiar property of multi-variant Gaussian integrals,

$$\int d\mathbf{x} e^{-\frac{1}{2\sigma^2} (\mathbf{x})^\top \mathbf{M}(\mathbf{x})} = (2\pi\sigma^2)^{m/2} \frac{1}{\sqrt{\det(\mathbf{M})}} \tag{191}$$

where $\mathbf{x}$ is an m-dim vector (with zero mean), and $\mathbf{M}$ is a square postive-definite matrix, and $\det(\mathbf{M})$ is the determinant of $\mathbf{M}$. Using Eqn. 191, we can rewrite $\mathbb{I}_H$ in Eqn. 190 as

$$\begin{aligned}
\mathbb{I}_H &= \mathcal{N} \left[\frac{(2\pi\sigma^2)^{M/2}}{\sqrt{\det(\mathbf{M})}} \right]^N \qquad (192) \\
&= \mathcal{N} (2\pi\sigma^2)^{NM/2} \left[\sqrt{\det\left(2\beta\sigma^2 (\mathbf{I}_M - \tfrac{1}{N} \mathbf{S}^\top \mathbf{T}) + \mathbf{I}_M \right)} \right]^{-N} \\
&= \left(\frac{1}{2\pi\sigma^2} \right)^{MN/2} (2\pi\sigma^2)^{M/2} \left[\sqrt{\det\left(\mathbf{I}_M + 2\beta\sigma^2 (\mathbf{I}_M - \tfrac{1}{N} \mathbf{S}^\top \mathbf{T}) \right)} \right]^{-N}, \qquad (193)
\end{aligned}$$

where Eqn. 193 follows by inserting $\mathcal{N}$ from Eqn. 178. We can now identify $\sigma^2 = \frac{1}{M}$ to obtain

$$\begin{aligned}\mathbb{I}_H &= \left[\sqrt{\det\left(\mathbf{I}_M + 2\beta\sigma^2(\mathbf{I}_M - \tfrac{1}{N}\mathbf{S}^\top\mathbf{T})\right)}\right]^{-N} \\ &= \left[\sqrt{\det\left(\mathbf{I}_M + \tfrac{2\beta}{MN}(\mathbf{I}_M - \tfrac{1}{N}\mathbf{S}^\top\mathbf{T})\right)}\right]^{-N} \\ &= \left[\det\left(\mathbf{I}_M + \tfrac{2\beta}{M}(\mathbf{I}_M - \tfrac{1}{N}\mathbf{S}^\top\mathbf{T})\right)\right]^{-N/2}. \end{aligned} \tag{194}$$

Large-N Approximation. Using the expression $\det(\mathbf{I}_M + \epsilon\mathbf{\Omega}) \approx 1 + \epsilon\,\mathrm{Tr}\left[\mathbf{\Omega}\right]$, which holds for an arbitrary matrix Ω for small ϵ, we can evaluate the determinant in Eqn. 194 in the large-N approximation, which gives

$$\mathbb{I}_H \approx \left[1 + \tfrac{2\beta}{M}(\,\mathrm{Tr}\left[\mathbf{I}_M - \tfrac{1}{N}\mathbf{S}^\top\mathbf{T}\right])\right]^{-N/2}. \tag{195}$$

Inserting this into Eqn. 182, we obtain

$$\beta h^{an}(\mathbf{R}) = -\frac{1}{N}\ln\left[1 + \tfrac{2\beta}{M}(\,\mathrm{Tr}\left[\mathbf{I}_M - \tfrac{1}{N}\mathbf{S}^\top\mathbf{T}\right])\right]^{-N/2}. \tag{196}$$

Matrix-Generalized ST Error $H^{an}_{hT}(\mathbf{R})$ when $M = 1$. To start, observe that when $M = 1$, Eqn. 196 becomes

$$\begin{aligned}\beta H^{an}(\mathbf{R})|_{M=1} &= \beta h^{an}(\mathbf{R})\|_{M=1} \\ &= -\frac{1}{N}\ln\left[1 + 2\beta(1 - \tfrac{1}{N}\mathbf{S}^\top\mathbf{T})\right]^{-N/2} \qquad (197) \\ &= \frac{1}{2}\ln\left[1 + 2\beta(1 - \tfrac{1}{N}\mathbf{S}^\top\mathbf{T})\right] \\ &= \frac{1}{2}\ln\left[1 + 2\beta(1 - \mathbf{R})\right], \qquad (198)\end{aligned}$$

meaning that Eqn. 196 reduces to Eqn. 172, as desired.

This ensures the Hamiltonian scales as M so the Free Energy scales as $N \times M$, the number of free paramaters in the system. Notice that for the final Layer Quality-Squared Hamiltonian $\mathbf{H}_{\bar{\mathcal{Q}}^2}$, this will change.

$$\mathbb{I}_H \approx \left[1 + \frac{2\beta}{M}\,\mathrm{Tr}[(\mathbf{I}_M - \tfrac{1}{N}\mathbf{S}^\top\mathbf{T})\right]^{-N/2}, \tag{199}$$

for any $M > 1$. Given this, it follows from Eqn. 196 and Eqn. 182 that we can define

$$\beta H^{an}(\mathbf{R}) := \frac{M}{2}\ln\det\left[1 + \frac{2\beta}{M}(\mathbf{I}_M - \mathbf{R})\right], \tag{200}$$

To obtain the high-Temperature form, we use

$$\ln\det\left[\mathbf{I}_M + \epsilon\mathbb{M}\right] \approx \mathrm{Tr}\left(\epsilon\mathbb{M}\right) \quad (\epsilon \ll 1), \tag{201}$$

with $\epsilon = \frac{2\beta}{M}$ and $\mathbb{M} = \mathbf{I}_M - \mathbf{R}$. We mow obtain the folloiwing result for the matrx-generalized high-T of $H^{an}_{hT}(\mathbf{R})$ using

$$\mathrm{Tr}\left[H^{an}_{hT}(\mathbf{R})\right] = \mathrm{Tr}\left[\mathbf{I}_M - \mathbf{R}\right] = M - \mathrm{Tr}\left[\mathbf{R}\right] \tag{202}$$

which is of the same functional form as Eqn. 172, as desired:

This form of the Hamiltonian, $H^{an}(\mathbf{R})$, however, is not symmetric, and we will eventually want a symmetric operator or matrix. Fortunately, the high-T form, $H^{an}_{hT}(\mathbf{R})$, can be made symmetric, as shown below.

A.3 Expressing the Layer Quality

In this section, we obtain an approximation expression for the Layer Quality-Squared from the IZ Free Energy for the Generalization Error, given in Eqn. 110 in Section 5.2.1.

For the required Free Energy $\beta\mathbf{F}^{IZ}$, we will use the matrix-generalized Hamiltonian from Eqn. 202 for the Layer Quality, $H_{hT}^{an}(\mathbf{R}) = \mathbb{I}_M - \mathbf{R}$, giving a Boltzmann distribution and the corresponding Thermal Average. Expanding this out, we have

$$-\beta\mathbf{F}^{IZ} = -\ln \int d\mu(\mathbf{S}) \exp\left[-N\beta \operatorname{Tr}[H_{hT}^{an}(\mathbf{R})]\right] \tag{203}$$

$$\tag{204}$$

We could also express $\beta\mathbf{F}^{IZ}$ In terms of the matrix-generalized Annealed Error Potential $\epsilon(\mathbf{R})$ (Eqn. 185), giving

$$-\beta\mathbf{F}^{IZ} = -\ln \int d\mu(\mathbf{S}) \exp\left[-N\beta\epsilon(\mathbf{R})\right] \tag{205}$$

In analogy with Eqn. 49, as $H_{hT}^{an}(\mathbf{R}) = M - \mathbf{R}$, write

$$-\beta\mathbf{F}^{IZ} = -\ln \int d\mu(\mathbf{S}) \exp\left[-N\beta \operatorname{Tr}[M - \mathbf{R}]\right] \tag{206}$$

Using the approximation $\operatorname{Tr}[\mathbf{R}] \approx \sqrt{\operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]}$, we have

$$-\beta\mathbf{F}^{IZ} \approx -\ln \int d\mu(\mathbf{S}) \exp\left[-N\beta(M - \sqrt{\operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]})\right] \tag{207}$$

$$= -\ln \int d\mu(\mathbf{S}) \exp[-NM\beta] \exp\left[N\beta\sqrt{\operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]}\right], \tag{208}$$

$$= -\ln e^{-NM\beta} \int d\mu(\mathbf{S}) \exp\left[N\beta\sqrt{\operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]}\right], \tag{209}$$

$$= -\ln e^{-NM\beta} - \ln \int d\mu(\mathbf{S}) \exp\left[N\beta\sqrt{\operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]}\right], \tag{210}$$

$$= NM\beta - \ln \int d\mu(\mathbf{S}) \exp\left[\beta N\sqrt{\operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]}\right], \tag{211}$$

Notice that, as expected, the Free Energy scales $\beta\mathbf{F}^{IZ}$ as $N \times M$. Since Eqn. 203 equals Eqn. 208, we can write the Free Energy in terms of $\operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]$. From Eqn. 211, we can identify a generating function ($\Gamma_{\bar{\mathcal{Q}}}$) for the layer accuracy, or Quality. For example, to compute the average Quality $\bar{\mathcal{Q}}$, we would use

$$\beta\Gamma_{\bar{\mathcal{Q}}}^{IZ} := \ln \int d\mu(\mathbf{S}) \exp\left[N\beta\sqrt{\operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]}\right], \tag{212}$$

and to compute the average Quality (squared) $\bar{\mathcal{Q}}^2$, we would use

$$\beta\Gamma_{\bar{\mathcal{Q}}^2}^{IZ} := \ln \int d\mu(\mathbf{S}) \exp\left[N\beta \operatorname{Tr}\left[\mathbf{R}^\top\mathbf{R}\right]\right]. \tag{213}$$

We have recovered Eqn. 110. We can now also define the Layer Quality-Squared Hamiltonian as

$$\mathbf{H}_{\bar{\mathcal{Q}}^2} := \mathbf{R}^\top\mathbf{R} \tag{214}$$

which is a symmetric operator, as desired. Consequently, we may also write

$$\beta\Gamma_{\bar{\mathcal{Q}}^2}^{IZ} := \ln \int d\mu(\mathbf{S}) \exp\left[N\beta \operatorname{Tr}[\mathbf{H}_{\bar{\mathcal{Q}}^2}]\right]. \tag{215}$$

A.4 Derivation of the TRACE-LOG Condition

A.4.1 Setting up the Saddle Point Approximation (SPA)

As in Eqn. 115, we can write Eqn. 11 in terms of the $\mathbf{A}_2$ form of the Student Correlation matrix, giving

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} = \ln \int_{\mathbf{S}} d\mu(\mathbf{S}) \exp\left(N\beta \,\mathrm{Tr}\left[\tfrac{1}{N}\mathbf{T}^\top \mathbf{A}_2 \mathbf{T}\right]\right) \tag{216}$$

where $d\mu(\mathbf{S})$ is the measure over all $N \times M$ real-valued random matrices, although we really want to limit this to all $N \times M$ real matrices that resemble the Teacher $\mathbf{T}$, which we clarify below.

To transform $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$ into a form we can evaluate using Tanakas result, we need to change the measure from an integral over all random $N \times M$ student weight matrices $d\mu(\mathbf{S})$ to an integral over all $N \times N$ student correlation matrices $d\mu(\mathbf{A})$, i.e., $d\mu(\mathbf{S}) \to d\mu(\mathbf{A})$. To accomplish this, we can insert an integral over the Dirac Delta function

$$\mathbf{I} \coloneqq \int d\mu(\mathbf{A}_1)\delta(N\mathbf{A}_1 - \mathbf{S}^\top\mathbf{S}) = \int d\mu(\mathbf{A})\delta(N\mathbf{A}_1 - \mathbf{S}^\top\mathbf{S}). \tag{217}$$

(This is simply a resolution of the Identity.) This gives

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} = \ln \int_{\mathbf{S}} d\mu(\mathbf{S}) \int_{\mathbf{A}} d\mu(\mathbf{A})\delta\left(N\mathbf{A}_1 - \mathbf{S}^\top\mathbf{S}\right) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\top\mathbf{A}_2\mathbf{T}]}, \tag{218}$$

where $d\mu(\mathbf{A}) = \mathbf{Pr}[\mathbf{A}]\, d\mathbf{A}$ and $\mathbf{Pr}[\mathbf{A}]$ is the (still unspecified) probability density over the new random matrix $\mathbf{A}$. Let us express Eqn. 218 at large-N as

$$\lim_{N\gg 1} \beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} = \lim_{N\gg 1} \ln \int d\mu(\mathbf{A}) \int d\mu(\mathbf{S})\delta(N\mathbf{A}_1 - \mathbf{S}^\top\mathbf{S}) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\top\mathbf{A}_2\mathbf{T}])}. \tag{219}$$

Now we assume we can first evaluate the term

$$\lim_{N\gg 1} \int d\mu(\mathbf{S})\delta(N\mathbf{A}_1 - \mathbf{S}^\top\mathbf{S}) \tag{220}$$

at large-N using a Saddle Point Approximation (SPA).

Using the relation,

$$\delta(N\mathbf{A}_1 - \mathbf{S}^\top\mathbf{S}) = \mathcal{N} \int_{\hat{\mathbf{A}}} d\mu(\hat{\mathbf{A}}) e^{iNTr[\hat{\mathbf{A}}\mathbf{A}_1]} e^{-iTr[\hat{\mathbf{A}}\mathbf{S}^\top\mathbf{S}]}, \tag{221}$$

where $\hat{\mathbf{A}}$ is a $M \times M$ auxiliary matrix, and the domain of integration $d\mu(\hat{\mathbf{A}})$ is all $M \times M$ real-valued matrices, and where the normalization $\mathcal{N}_1$ is

$$\mathcal{N}_1 \coloneqq \frac{1}{(2\pi)^{M(M+1)/4}}, \tag{222}$$

because $\mathbf{A}$ is a symmetric matrix with $M(M+1)/2$ constraints.

This is simply the matrix generalization of $\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\hat{x}x} d\hat{x}$, so we can express the delta function as an exponential, giving

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} = \mathcal{N}_1 \ln \int_{\mathbf{S}} d\mu(\mathbf{S}) \int_{\mathbf{A}} d\mu(\mathbf{A}_1) \int_{\hat{\mathbf{A}}} d\mu(\hat{\mathbf{A}}) e^{iNTr[\hat{\mathbf{A}}\mathbf{A}_1]} e^{-iTr[\hat{\mathbf{A}}\mathbf{S}^\top\mathbf{S}]} e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\top\mathbf{A}_2\mathbf{T}]}. \tag{223}$$

Rearranging terms, we obtain

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} = \ln \int_{\mathbf{A}} d\mu(\mathbf{A}) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\top\mathbf{A}_2\mathbf{T}]} \times \Gamma_1, \tag{224}$$

where we define Γ_1 as

$$\Gamma_1 := \Gamma_1(\mathbf{A}_1) = \mathcal{N}_1 \int_{\mathbf{S}} d\mu(\mathbf{S}) \int_{\hat{\mathbf{A}}} d\mu(\hat{\mathbf{A}}) e^{iNTr[\hat{\mathbf{A}}\mathbf{A}_1]} e^{-iTr[\hat{\mathbf{A}}\mathbf{S}^\top\mathbf{S}]}. \tag{225}$$

We can simplify the complex integral in Γ_1 with the Wick Rotation $i\hat{\mathbf{A}} \to \hat{\mathbf{A}}$. The Wick rotation ensures that the Gaussian integral converges (although this has not been rigorously checked) We may expect $d\mu(\hat{\mathbf{A}})$ to be invariant to rotations in the complex plane so the Wick rotation does not introduce any complex prefactors. This gives

$$\Gamma_1 = \mathcal{N}_1 \int_{\mathbf{S}} d\mu(\mathbf{S}) \int_{\hat{\mathbf{A}}} d\mu(\hat{\mathbf{A}}) e^{NTr[\hat{\mathbf{A}}\mathbf{A}_1]} e^{-Tr[\hat{\mathbf{A}}\mathbf{S}^\top\mathbf{S}]} \tag{226}$$

$$= \mathcal{N}_1 \int_{\mathbf{S}} d\mu(\mathbf{S}) \int_{\hat{\mathbf{A}}} d\mu(\hat{\mathbf{A}}) e^{NTr[\hat{\mathbf{A}}\mathbf{A}_1]} e^{-Tr[\mathbf{S}\hat{\mathbf{A}}\mathbf{S}^\top]}, \tag{227}$$

where the second line follows since the trace is invariant under cyclic permutations (i.e., $\mathrm{Tr}\left[ABC\right] = \mathrm{Tr}\left[BCA\right] = \mathrm{Tr}\left[CAB\right]$). Swapping the order of the integrals yields

$$\Gamma_1 = \Gamma_1(\hat{\mathbf{A}}) = \mathcal{N}_1 \int_{\hat{\mathbf{A}}} d\mu(\hat{\mathbf{A}}) e^{NTr[\hat{\mathbf{A}}\mathbf{A}_1]} \times \Gamma_2, \tag{228}$$

where we define Γ_2 as

$$\Gamma_2 := \Gamma_2(\hat{\mathbf{A}}) = \int_{\mathbf{S}} d\mu(\mathbf{S}) e^{-Tr[\mathbf{S}\hat{\mathbf{A}}\mathbf{S}^\top]}.$$

To evaluate Γ_2, we will make several mathematically convenient approximations. (These will yield an approximate expression which can be verified empirically.) We first assume for the purpose of changing measure that the (data) columns of $\mathbf{S}$ are statistically independent, so that the measure $d\mu(\mathbf{S})$ factors into N gaussian distributions

$$d\mu(\mathbf{S}) = \prod_{\mu=1}^{N} d\mu(\mathbf{s}_\mu) = \prod_{\mu=1}^{N} d\mathbf{s}_\mu, \tag{229}$$

where $\mathbf{s}_\mu$ is an M-dimensional vector. The singular values of $\mathbf{S}$ are invariant to randomly permuting the columns or rows, so the resulting ESD does not change. This is very different from permuting $\mathbf{S}$ element-wise, which will make the resulting ESD Marchenko Pastur (MP).

Using Eqn. 229, Γ_2 reduces to a simple Gaussian integral, which can be evaluated as a product of N Gaussian integrals (over the $M \times M$ matrix $\hat{\mathbf{A}}$)

$$\Gamma_2 = \left[\int_{\mathbf{s}} d\mathbf{s} e^{-\frac{1}{\sigma^2}\mathbf{s}\hat{\mathbf{A}}\mathbf{s}^\top} \right]^N \tag{230}$$

$$= \left[\mathcal{N}_2 \det\left(\hat{\mathbf{A}}\right)^{-1/2} \right]^N, \tag{231}$$

where the normalization $\mathcal{N}_2$

$$\mathcal{N}_2 := \left(\pi\sigma^2\right)^{M/2}, \tag{232}$$

where $\sigma^2 = \mathbf{s}^\top\mathbf{s} = 2?1/M$

For any square, non-singular matrix $\hat{\mathbf{A}}$, $\mathrm{Tr}\left[\ln \hat{\mathbf{A}}\right] = \ln\det\left(\hat{\mathbf{A}}\right)$, so it follows from Eqn. 230 that

$$\ln\Gamma_2 = N \ln\mathcal{N}_2 \left[(det\hat{\mathbf{A}})^{-1/2} \right]$$

$$= N \ln\mathcal{N}_2 - \tfrac{N}{2} \mathrm{Tr}\left[\ln \hat{\mathbf{A}}\right], \tag{233}$$

so that

$$\Gamma_2 = \mathcal{N}_2^N e^{-\frac{N}{2}Tr[\ln \hat{\mathbf{A}}]} \tag{234}$$

Substituting Eqn. 234 into Eqn. 228, we can write Γ_1 as

$$\Gamma_1(\hat{\mathbf{A}}) = \quad C_{\Gamma_1} \int_{\hat{\mathbf{A}}} d\mu(\hat{\mathbf{A}}) e^{NTr[\hat{\mathbf{A}}\mathbf{A}_1]} e^{-\frac{N}{2}Tr[\ln \hat{\mathbf{A}}]}, \tag{235}$$

where

$$C_{\Gamma_1} := \mathcal{N}_1 e^{\mathcal{N}_2}. \tag{236}$$

We now can evaluate the integral in Eqn. 228 over the Lagrange Multiplier $\hat{\mathbf{A}}$ (i.e., $\int_{\hat{\mathbf{A}}}$). If we call this $\Gamma_1(\hat{\mathbf{A}})$, then (following Tanaka [83]) we can define the *Rate Function* $I(\hat{\mathbf{A}}, \mathbf{A}_1)$ such that

$$\Gamma_1(\hat{\mathbf{A}}) = \int_{\hat{\mathbf{A}}} d\mu(\hat{\mathbf{A}}) e^{-NI(\hat{\mathbf{A}}, \mathbf{A}_1)}, \tag{237}$$

where

$$I(\hat{\mathbf{A}}, \mathbf{A}_1) = -\operatorname{Tr}\left[\hat{\mathbf{A}}\mathbf{A}_1\right] + \frac{1}{2}\operatorname{Tr}\left[\ln \hat{\mathbf{A}}\right]. \tag{238}$$

We can formally evaluate the integral in Eqn. 237 in the large-N limit using a Saddle Point Approximation (SPA) (see Section 4.2, Eqn. 67), as

$$\Gamma_1(\hat{\mathbf{A}}) \to \sqrt{\frac{(2\pi)^{N/2}}{N\|I\|}} e^{-NI^*(\hat{\mathbf{A}}, \mathbf{A}_1)}, \tag{239}$$

where $I^*(\hat{\mathbf{A}}, \mathbf{A})$ is the maximum value, obtained using

$$I^*(\hat{\mathbf{A}}, \mathbf{A}_1) := \lim_{N \gg 1} I(\hat{\mathbf{A}}, \mathbf{A}_1) = \sup_{\hat{\mathbf{A}}} \left[-\operatorname{Tr}\left[\hat{\mathbf{A}}\mathbf{A}_1\right] + \frac{1}{2}\operatorname{Tr}\left[\ln \hat{\mathbf{A}}\right] \right], \tag{240}$$

where I at the SPA is defined as

$$I := \frac{\partial}{\partial \hat{\mathbf{A}}} I(\hat{\mathbf{A}}, \mathbf{A}_1) = -\mathbf{A}_1 + \frac{1}{2\hat{\mathbf{A}}} = 0 \tag{241}$$

and I is defined as

$$I = \frac{\partial^2}{\partial \hat{A}^2} I(\hat{\mathbf{A}}, \mathbf{A}_1) = -\frac{1}{2}\left(\frac{1}{2}\hat{A}^{-1}\right) \otimes \left(\frac{1}{2}\hat{A}^{-1}\right) = -\frac{1}{8}\mathbf{A}_1 \otimes \mathbf{A}_1 \tag{242}$$

where $\otimes$ is the Kronecker product, and $\mathbf{A}_1 \otimes \mathbf{A}_1$ is the Hessian of $\mathbf{A}_1$.

Solving the SPA equation, we find that the auxiliary matrix is $\hat{\mathbf{A}} = \frac{1}{2}\mathbf{A}_1^{-1}$ and the prefactor (Hessian) is given as $det\left(-\frac{1}{8}\mathbf{A}_1 \otimes \mathbf{A}_1\right) = \left(-\frac{1}{8}\right)^{M^2} \left(det(\mathbf{A}_1)\right)^M$.

[double check prefactors.]

Substituting for $\hat{\mathbf{A}}$ into Rate Function (Eqn. 238), I becomes

$$\begin{aligned} I^*(\hat{\mathbf{A}}, \mathbf{A}_1) &= -\operatorname{Tr}\left[\mathbb{I}_M\right] + \frac{1}{2}\operatorname{Tr}\left[\ln \mathbf{A}_1\right] \\ &= -M + \frac{1}{2}\operatorname{Tr}\left[\ln \mathbf{A}_1\right]. \end{aligned} \tag{243}$$

In order for this result to be physically meaningful, we need that if $I^*(\hat{\mathbf{A}}, \mathbf{A}_1)$ grows, then it must grow slower than N, and, more importantly, that $\det(\mathbf{A})$ be non-zero. Importantly, When $\det(\mathbf{A}) = 1$ exactly, however, then Γ_1 becomes a constant, and this simplifies things considerably!

A.4.2 Casting the Generating Function ($\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$) as an HCIZ Integral

In this section, we express the Generating Function $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$, given in Eqn. 115 (equivalently, in Eqn. 110), as an HCIZ Integral, as given in Eqn. 119.

Inserting $I^*(\hat{\mathbf{A}}, \mathbf{A})$ from Eqn. 243 into $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$, we obtain

$$\begin{aligned}\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} &= \ln\left[C_{\Gamma_1} e^{-NM} \int d\mu(\mathbf{A}) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal\mathbf{A}_2\mathbf{T}]} e^{\frac{N}{2}\ln(\det(\mathbf{A}_1))}\right] \\ &= \ln C_{\Gamma_1} - NM + \ln\left[\int d\mu(\mathbf{A}) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal\mathbf{A}_2\mathbf{T}]} e^{\frac{N}{2}\ln(\det(\mathbf{A}_1))}\right].\end{aligned} \tag{244}$$

So long as the second term $\mathrm{Tr}\left[\mathbb{I}_M\right]$ does not depend on N, it will vanish when we take the partial derivative of $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$ to obtain the $\bar{\mathcal{E}}^{NN}_{gen}$, in which case it is not important. We can then simply write the Generating Function $\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}$ as in Eqn. 116 as:

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} = \ln\left[\int d\mu(\mathbf{A}_1) e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal\mathbf{A}_2\mathbf{T}]} e^{\frac{N}{2}\ln(\det(\mathbf{A}_1))}\right], \tag{245}$$

or, in Bra-Ket notation, as

$$\beta\mathbf{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2} = \ln\left\langle e^{N\beta Tr[\frac{1}{N}\mathbf{T}^\intercal\mathbf{A}_2\mathbf{T}]} e^{\frac{N}{2}\ln(\det(\mathbf{A}_1))}\right\rangle_{\mathbf{A}}. \tag{246}$$

A.5 MLP3 Model Details

The empirical MLP3 Model implements the assumptions described in Section 5 used the following procedures:

A three-layer Multi-Layer Perceptron was trained for classification on the MNIST dataset[102]. The first Fully Connected (FC) hidden layer has 300 units, the second FC hidden layer has 100 units, and the third FC layer has ten units for classification, matching the ten digit classes of MNIST. Input images are grayscale, and were rescaled to the $[0, 1]$ range. Following the keras[115] defaults, the weights were initialized using the Glorot Normal[116] method, and the biases were initialized to 0. Each model was trained using Categorical Cross Entropy as the loss function. The loss function was *summed* over each mini-batch, which is the default behavior for Keras, rather than being *averaged*, which is the default for pytorch[117].

Optimization was carried out by either Stochastic Gradient Descent (SGD) without momentum, or the Adam algorithm [118]. The Learning Rate (LR) was set to 0.01 for SGD, and 0.001 for Adam. The LR was held constant, i.e., there was no decay schedule. Each algorithm proceeded epoch by epoch until the value of the loss function did not decrease by more than 0.0001 for three consecutive epochs. At each epoch, the `WeightWatcher` tool was used to compute metrics for each layer. Loss values reported are the average loss per labeled example, and not the summed loss over each minibatch. Training loss is averaged over all batches in the epoch, whereas test loss is evaluated once at the end of the epoch.

In some experiments, only one layer was trained, while the others were left frozen. In other experiments all layers were trained. Models were trained using a series of mini-batch sizes ranging from 1 to 32. For each separate training run, the models were re-initialized to the same starting random weights, all random seeds were reset, and deterministic computations were used to train the models.

Separate notebooks are provided for keras and pytorch implementations of the experiments.

A.6 Tanaka's Result

In this section, we will rederive the result by Tanaka [82, 83] that we use in our main derivation, and, importantly, explain how to address the missing Temperature term. For completeness, we restate it here using the notation of the main text:

$$\lim_{N\gg 1}\frac{1}{N}\ln\underbrace{\left\langle\exp\left(\frac{\beta}{2}\operatorname{Tr}\left[\mathbf{W}^{\mathsf{T}}\mathbf{A}_2\mathbf{W}\right]\right)\right\rangle_{\mathbf{A}}}_{\text{HCIZ Integral}}=\frac{\beta}{2}\sum_{i=1}^{M}\mathbb{G}_{\mathbf{A}}(\lambda_i) \tag{247}$$

where $\mathbf{W}$ is the $N\times M$ Teacher weight matrix, $\mathbf{A}=\mathbf{A}_2$ is the $N\times N$ Student (correlation) matrix, but β is now the inverse-Temperature(because we are working with real matrices), and we have added a $\frac{1}{2}$ (which will be clear later). $\mathbb{G}_{\mathbf{A}}(\lambda)$ is a complex analytic function of the eigenvalues λ of (the Teacher Correlation matrix) $\mathbf{X}$, whose functional form will depend on the structure of the limiting form of (the Student) ESD $\rho^{\infty}_{\mathbf{A}}(\lambda)$. We may also write it as $\mathbb{G}_{\mathbf{A}}(\mathbf{X})$ below. We call it perhaps somewhat imprecisely a Norm Generating Function because the final results for the Layer Quality $\bar{\mathcal{Q}}$ will take the form of a *Tail norm* in many cases.

To apply this result, we note that while the term β is just a constant in [83] (1 or 2, depending on whether the random matrix is real or complex), it is not actually inverse Temperature $\beta=\frac{1}{T}$ in the original derivation. Still, we seek a final result that is linear in $\beta=\frac{1}{T}$, so that we can easily evaluate $\bar{\mathcal{Q}}^2$ in the high-T limit, i.e. $\bar{\mathcal{Q}}^2=\frac{\partial}{\partial N}\frac{1}{\beta}\beta\boldsymbol{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1}=\frac{\partial}{\partial\beta}\frac{1}{N}\beta\boldsymbol{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1}$ (see 11). We can introduce $\beta=\frac{1}{T}$ by simply changing the scale of $\mathbf{A}_2$ since the final result is a sum of R-transforms, which by definition are linear, i.e. $\mathbb{G}_{\mathbf{A}}(\beta\lambda)=\beta\mathbb{G}_{\mathbf{A}}(\lambda)$, however, it is instructive to rederive the final result, with β explicitly included.

Notation. We start by re-writing the Tanaka result, Eqn. (247), in our notation for the expected value $\langle\cdots\rangle_{\mathbf{A}}$ operator, as follows:

$$\tfrac{1}{2}\beta\boldsymbol{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1}=\lim_{N\gg 1}\ln\underbrace{\int d\mu(\mathbf{A})\left[\exp\left(\frac{\beta}{2}\operatorname{Tr}\left[\mathbf{W}^{\mathsf{T}}\mathbf{A}_2\mathbf{W}\right]\right)\right]}_{\text{HCIZ Integral}}=N\beta\tfrac{1}{2}\sum_{i=1}^{M}\mathbb{G}_{\mathbf{A}}(\lambda_i). \tag{248}$$

where we have added a $\frac{1}{2}$ for technical convenience (to make the connection with the LDP, below). If we denote the internal HCIZ integral as

$$\mathbb{Z}^{IZ}:=\int d\mu(\mathbf{A})\left[\exp\left(\frac{\beta}{2}\operatorname{Tr}\left[\mathbf{W}^{\mathsf{T}}\mathbf{A}_2\mathbf{W}\right]\right)\right], \tag{249}$$

then it holds that

$$\beta\boldsymbol{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2}:=\ln\mathbb{Z}^{IZ}, \tag{250}$$

from which it follows that

$$\beta\boldsymbol{\Gamma}^{IZ}_{\bar{\mathcal{Q}}^2,N\gg 1}:=\lim_{N\gg 1}\ln\mathbb{Z}^{IZ}. \tag{251}$$

The SPA approximates the Partition Function $\mathbb{Z}^{IZ}$, which is now an HCIZ integral, by its peak value. For this, $\mathbb{G}_{\mathbf{A}}(\lambda)$ itself must either not explicitly depend on N and/or at least not grow faster than N.

The trick here is we can choose an R-transform of $\mathbf{A}$ that is a simple analytic expression based on the observed the empirical spectral density (ESD) of the $\mathbf{X}$. And this can readily be done for the ESDs for a wide range of layer weight matrices observed in modern DNNs because the their

ESDs are Heavy-Tailed Power Law [62]. We can then readily express the Quality $\bar{\mathcal{Q}}$ of the Teacher layer in a simple functional form, (i.e an approximate Shatten Norm),

Importantly, the matrices $\mathbf{X}$ and $\mathbf{A}$ must be well approximated by low rank matrices since the derivation in Tanaka requires this. Fortunately, this appears to be generally true for the layers in very well trained DNNs, which is what allows us to apply this withing the `ECS`. In fact, technically we need to integrate over $d\mu(\tilde{\mathbf{A}})$; this is straightforward as this simply changes the lower bound on the integral from $0 \to \lambda_{min}^{ECS}$, both above and in the subsequent derivation. For now, we omit this detail; I may add it in later

Finally, we note that $\mathbb{G}_{\mathbf{A}}(\mathbf{X})$ is kind of *Generalized Norm* because it can be evaluated as a sum over a function of the M eigenvalues λ_μ of the Teacher correlation matrix $\mathbf{X} = \frac{1}{N}\mathbf{W}^\top\mathbf{W}$. $\mathbb{G}_{\mathbf{A}}(\mathbf{X})$ will turn out to be a simple expression similar to the Frobenius Norm or the Shatten Norm of $\mathbf{X}$, depending on the functional form we choose to model the lmiting form of the Student ESD, $\rho_{\mathbf{A}}^{\infty}(\lambda)$.

A.6.1 Setup and Outline

To evaluate 248, we want to integrate over all Student Correlation matrices $\mathbf{A}$ that "resemble the Teacher Correlation matrix $\mathbf{X}$. To formalize this idea, we need to define the measure over "all desired $\mathbf{A}$, $d\mu(\mathbf{A})$, in terms of the actual M eigenvalues, $\{\lambda_i\}_{i=1}^{M}$, of the Teacher.

Randomness assumption. For real weights $\mathbf{W}$, we assume an *orthogonally invariant* ensemble, $d\mu(\mathbf{W}) = d\mu(\mathbf{U}\mathbf{W}\mathbf{U}^\top)$ for all $\mathbf{U} \in O(M)$, mirroring the isotropic Gaussian initialisation widely used in neural networks. Crucially, Tanaka's large-N analysis shows the resulting HCIZ exponent depends only on the eigenvalue spectrum, so the final integrated–R expression should remain applicable even when full rotational invariance is later broken in training.

Using a source matrix D to represent $d\mu(\mathbf{A})$ with $d\mu(\mathbf{W})$. We consider all matrices $\mathbf{A}$ with the same limiting spectral density, $\rho_{\mathbf{A}}^{\infty}(\lambda)$, as the limiting (*empirical*) ESD of the Teacher. That is, we want $\rho_{\mathbf{A}}^{\infty}(\lambda) = \rho_{\mathbf{W}}^{\infty}(\lambda)$, where $\mathbf{T} = \mathbf{W}$. Of course, there are infinitely many weight matrices $\mathbf{W}$ with the same M eigenvalues, $\{\lambda_i\}_{i=1}^{M}$, as the Teacher. Let us specify these matrices with the measure $d\mu(\mathbf{W})$. Doing this lets us then write the measure $d\mu(\mathbf{A})$ in terms of $d\mu(\mathbf{W})$ as:

$$d\mu(\mathbf{A}) := e^{-\frac{\beta}{2}Tr[\mathbf{W}\mathbf{D}\mathbf{W}^\top]}d\mu(\mathbf{W}), \tag{252}$$

where $\mathbf{D}$ is some $M \times M$ matrix, called the Source Matrix, to be specified below, and the $\frac{1}{2}$ here as well. Indeed, the key idea here will be to define $\mathbf{D}$ in such a way as to obtain the desired final result. Notice also that we have added a β term; this will be factored out later.

We can now represent the partition function $\mathbb{Z}^{IZ}$, by inserting Eqn. 252 into Eqn. 249. $\mathbb{Z}^{IZ}$ is now defined as an integral over all possible (Teacher) weight matrices $\mathbf{W}$

$$\mathbb{Z}^{IZ} = \int d\mu(\mathbf{W}) \exp[\frac{\beta}{2}\left(\mathrm{Tr}\left[\mathbf{W}^\top\mathbf{A}_2\mathbf{W}\right] - \mathrm{Tr}\left[\mathbf{W}\mathbf{D}\mathbf{W}^\top\right]\right)], \tag{253}$$

Observe that this integral only converges when all the eigenvalues of $\mathbf{D}$, $\{\vartheta_\mu\}_{\mu=1}^{M}$, are larger than the maximum eigenvalue of $\mathbf{A}$, i.e., when $\vartheta_\mu > \lambda_{max}$, for $\mu \in [1, M]$ (although below this will become $\beta\vartheta_\mu > \lambda_{max}$). Later, we will place $\mathbf{D}$ in diagonal form, and we will obtain an explicit expression for its M eigenvalues in terms of the M non-zero eigenvalues of $\mathbf{X}$. The eigenvalues of $\mathbf{D}$ will turn out to Lagrange Multipliers, needed later.

The Saddle Point Approximation (SPA) and the Large Deviation Principle (LDP). To evaluate the large-N case of $\beta\mathbf{\Gamma}^{IZ}_{\bar{Q}^2}$ (see 250, 251), we assume that the distribution of possible Teacher correlation matrices, $\mu(\mathbf{X})$, satisfies a *Large Deviation Principle (LDP)*. A LDP applies to probability distributions that take an exponential form, such that $\mu(\mathbf{X}) = e^{-NI(\mathbf{X})}d\mu(\mathbf{X})$, where $I(\mathbf{X})$ is Entropy or Rate function $I(\mathbf{X})$. In applying a LDP, we effectively restrict measure of student correlation matrices $\mathbf{A}$ to those most similar to the empirically observed Teacher correlation matrix $\mathbf{X}$. We expect the measure over all Teacher correlation matrices follows an LDP because the ESD is far from Gaussian, the dominant generalizing components reside in the tail of the ESD, and at finite-size the tail decays at worst as an exponentially Truncated Power Law (TPL).

Two steps to evaluate $\left\langle \mathbb{Z}^{IZ} \right\rangle_{\mathbf{A}}$ in the large-N approximation. The goal is to start with Eqn. 253 and obtain two separate, equivalent relations, Eqns. 254 and 256:

1. **Obtaining an integral transform of $\rho^{\infty}_{\mathbf{A}}(\lambda)$.** First, we expand and reduce Eqn. 253 and evaluate the expected value of $\mathbb{E}_{\mathbf{A}}[\mathbb{Z}^{IZ}] = \mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}]$ in the large-N limit by expressing the $\rho_{\mathbf{A}}(\lambda)$ for the $N \times N$ matrix $\mathbf{A} = \mathbf{A}_2 = \frac{1}{N}\mathbf{S}\mathbf{S}^{\intercal}$ in the continuum representation, i.e., as $\rho^{emp}_{\mathbf{A}}(\lambda) \rightarrow \rho^{\infty}_{\mathbf{A}}(\lambda)$, to obtain:

$$\lim_{N\gg 1}\frac{1}{N}\ln \mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}] = M\ln(\frac{2\pi}{\beta}) - \sum_{\mu=1}^{M}\int \ln(\delta_{\mu}-\lambda)\rho^{\infty}_{\mathbf{A}}(\lambda)d\lambda. \tag{254}$$

   This gives us an $\mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}]$ in terms of an integral transform $\rho^{\infty}_{\mathbf{A}}(\lambda)$, which we can model.[44]

2. **Forming the Saddle Point Approximation (SPA).** We evaluate Eqn. 253 as the expected value of $\mathbb{E}_{\mathbf{A}}[\mathbb{Z}^{IZ}] = \mathbb{E}_{\mathbf{A}_1}[\mathbb{Z}^{IZ}]$ for the $M \times M$ matrix $\mathbf{A} = \mathbf{A}_1 = \frac{1}{N}\mathbf{S}^{\intercal}\mathbf{S}$ (but explicitly in terms of $d\mu(\mathbf{X})$). Then, taking in the large-N approximation using the SPA, (and which can be done implicitly using the LDP), we obtain

$$\lim_{N\gg 1}\mathbb{E}_{\mathbf{A}_1}[\mathbb{Z}^{IZ}] \simeq \int \exp\left(\beta N \operatorname{Tr}\left[\mathcal{G}(\mathbf{X})\right]\right) d\mu(\mathbf{X}) \approx \exp(\beta N \mathcal{G}^{max}) \tag{255}$$

   where $\mathcal{G}(\mathbf{X})$ depends on $\mathbb{G}_{\mathbf{A}}(\mathbf{X})$, and $\mathcal{G}^{max} = \sup_{\mathbf{X}}\mathcal{G}(\mathbf{X})$. We can then write

$$\lim_{N\gg 1}\frac{1}{N}\ln \mathbb{E}_{\mathbf{A}_1}[\mathbb{Z}^{IZ}] \approx \beta\mathcal{G}^{max}, \tag{256}$$

3. **Finding the Inverse Legendre Transform.** To do this, we now equate

$$\lim_{N\gg 1}\frac{1}{N}\ln \mathbb{E}_{\mathbf{A}_1}[\mathbb{Z}^{IZ}] = \lim_{N\gg 1}\frac{1}{N}\ln \mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}] \tag{257}$$

   Then, we can form the inverse Legendre transform which we will let us relate $\mathbb{G}_{\mathbf{A}}(\lambda)$ in Eqn. 247 to the integrated R-transform of $\rho^{\infty}_{\mathbf{A}}(\lambda)$.

(See A.6.5.)

A.6.2 Step 1. Forming the Integral Transformation of ESD ($\rho^{\infty}_{\mathbf{A}}(\lambda)$)

We first establish Eqn. 254, in Steps 1.1 – 1.4. This is done by changing variables under a Unitary transformation, $\mathbf{W} \rightarrow \check{\mathbf{W}}$, evaluating the resulting functional determinant, and then taking the continuum limit of the ESD $\tilde{\rho}_{\mathbf{A}}(\lambda) \rightarrow \rho^{\infty}_{\mathbf{A}}(\lambda)$.

[44] This integral of $\rho^{\infty}_{\mathbf{A}}(\lambda)$ is related to the *Shannon Transform*, an integral transform from information theory that is useful when analyzing the mutual information or the capacity of a communication channel [82].

Step 1.1 To do so, let us first assume that Teacher correlation matrix $\mathbf{X}$ and the source matrix $\mathbf{D}$ are simultaneously diagonalizable (i.e., their commutator is zero: $[\mathbf{X}, \mathbf{D}] = 0$). In this case, we may write the generating function $\mathbb{Z}^{IZ}$ in Eqn. 253 as

$$\mathbb{Z}^{IZ} = \int d\mu(\mathbf{W}) \exp \frac{\beta}{2} \Big(\operatorname{Tr}\big[\mathbf{W}^\top \mathbf{U}^\top \mathbf{\Lambda} \mathbf{U} \mathbf{W}\big] - \operatorname{Tr}\big[\mathbf{W} \mathbf{V}^\top \mathbf{\Delta} \mathbf{V} \mathbf{W}^\top\big] \Big), \tag{258}$$

where we have defined

$$\mathbf{A}_2 = \mathbf{U}^\top \mathbf{\Lambda} \mathbf{U}, \;\; \mathbf{D} = \mathbf{V}^\top \mathbf{\Delta} \mathbf{V}, \tag{259}$$

where $\mathbf{U}$ $(N \times N)$ and $\mathbf{V}$ $(M \times M)$ are Unitary matrices. Since $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$, we can insert these identities into $\mathbb{Z}^{IZ}$ in 258, giving

$$\begin{aligned} \mathbb{Z}^{IZ} = \int d\mu(\mathbf{W}) \exp \frac{\beta}{2} \times \\ \Big(\operatorname{Tr}\big[(\mathbf{V}^\top \mathbf{V}) \mathbf{W}^\top \mathbf{U}^\top \mathbf{\Lambda} \mathbf{U} \mathbf{W} (\mathbf{V}^\top \mathbf{V})\big] - \operatorname{Tr}\big[(\mathbf{U}^\top \mathbf{U}) \mathbf{W} \mathbf{V}^\top \mathbf{\Delta} \mathbf{V} \mathbf{W}^\top (\mathbf{U}^\top \mathbf{U})\big] \Big). \end{aligned} \tag{260}$$

We can identify the reduced weight matrix $\check{\mathbf{W}}$ as

$$\check{\mathbf{W}} = \mathbf{U} \mathbf{W} \mathbf{V}^\top, \;\; \check{\mathbf{W}}^\top = \mathbf{V} \mathbf{W}^\top \mathbf{U}^\top, \tag{261}$$

Rearranging parentheses, this gives

$$\begin{aligned} \mathbb{Z}^{IZ} = \int d\mu(\mathbf{W}) \exp \frac{\beta}{2} \times \\ \Big(\operatorname{Tr}\big[\mathbf{V}^T (\mathbf{V} \mathbf{W}^T \mathbf{U}^T) \mathbf{\Lambda} (\mathbf{U} \mathbf{W} \mathbf{V}^T) \mathbf{V}\big] - \operatorname{Tr}\big[\mathbf{U}^\top (\mathbf{U} \mathbf{W} \mathbf{V}^\top) \mathbf{\Delta} (\mathbf{V} \mathbf{W}^\top \mathbf{U}^\top) \mathbf{U}\big] \Big). \end{aligned} \tag{262}$$

We can now express $\mathbb{Z}^{IZ}$ in terms of $\check{\mathbf{W}}$ as

$$\mathbb{Z}^{IZ} = \int d\mu(\mathbf{W}) \exp \frac{\beta}{2} \Big(\operatorname{Tr}\big[\mathbf{V}^\top \check{\mathbf{W}}^\top \mathbf{\Lambda} \check{\mathbf{W}} \mathbf{V}\big] - \operatorname{Tr}\big[\mathbf{U}^\top \check{\mathbf{W}} \mathbf{\Delta} \check{\mathbf{W}}^\top \mathbf{U}\big] \Big). \tag{263}$$

Since the Trace operator $\operatorname{Tr}[\cdot]$ is invariant to Unitary (Orthogonal) transformations, we can now remove the $\mathbf{U}$ and $\mathbf{V}$ terms, giving the simplified expression for our generating function $\mathbb{Z}^{IZ}$ in terms of the two diagonal matrices $\mathbf{\Lambda}, \mathbf{\Delta}$, the reduced weight matrix $\check{\mathbf{W}}$, and the Jacobian $J(\check{\mathbf{W}})$ transformation for $d\mu(\mathbf{W}) \to d\mu(\check{\mathbf{W}})$, as:

$$\mathbb{Z}^{IZ} = \int d\mu(\check{\mathbf{W}}) J(\check{\mathbf{W}}) \exp \frac{\beta}{2} \Big(\operatorname{Tr}\big[\check{\mathbf{W}}^\top \mathbf{\Lambda} \check{\mathbf{W}}\big] - \operatorname{Tr}\big[\check{\mathbf{W}} \mathbf{\Delta} \check{\mathbf{W}}^\top\big] \Big). \tag{264}$$

Step 1.2 We can now evaluate the integral using the standard relation for the functional determinant for infinite-dimensional Gaussian integrals [114]

$$\mathbb{Z}^{IZ} = \left(\frac{2\pi}{\beta} \right)^{NM/2} \det\left(\mathbf{\Delta} - \mathbf{\Lambda} \right)^{-1/2} \tag{265}$$

where the Jacobian is unity for the Unitary transformation.

$$J(\check{\mathbf{W}}) = 1. \tag{266}$$

since $\mathbf{W} \mapsto \check{\mathbf{W}}$ is an orthogonal transfomation. We now use the standard Trace-Log-Determinant relation [114]

$$\mathrm{Tr}\left[\ln \mathbf{M}\right] = \ln \det \mathbf{M}. \tag{267}$$

Let us insert $(\exp \ln)$ on the R.H.S. of 265, to obtain

$$\begin{aligned}
\mathbb{Z}^{IZ} &= \exp \ln \left[\left(\frac{2\pi}{\beta}\right)^{NM/2} \det\left(\mathbf{\Delta} - \mathbf{\Lambda}\right)^{-1/2} \right] \\
&= \exp\left[\left(\frac{NM}{2}\right) \ln \frac{2\pi}{\beta} - \frac{1}{2} \mathrm{Tr}\left[\ln\left(\mathbf{\Delta} - \mathbf{\Lambda}\right)\right] \right] \\
&= \exp\left[\frac{NM}{2} \ln \frac{2\pi}{\beta} - \frac{1}{2} \ln \det\left(\mathbf{\Delta} - \mathbf{\Lambda}\right) \right].
\end{aligned} \tag{268}$$

Step 1.3 We now want to express the generating function $\mathbb{Z}^{IZ}$ in 268 in terms of an integral over the continuous, limiting spectral density $\rho_{\mathbf{A}}(\lambda)$ of the correlation matrix $\mathbf{A}_2$.

First, we express the Determinant of the matrix $\mathbf{\Delta} - \mathbf{\Lambda}$ in terms of discrete eigenvalues:

$$\det\left(\mathbf{\Delta} - \mathbf{\Lambda}\right)^{-1/2} = \prod_{\mu=1}^{M} \prod_{i=1}^{N} \left(\vartheta_\mu - \lambda_i\right)^{-1/2}. \tag{269}$$

This gives the Log-Determinant in terms of the M (non-zero) eigenvalues of $\mathbf{D}$ and $\mathbf{A}_2$, as

$$\ln \det\left(\mathbf{\Delta} - \mathbf{\Lambda}\right)^{-1/2} = -\frac{1}{2} \sum_{\mu=1}^{M} \sum_{i=1}^{N} \ln\left(\vartheta_\mu - \lambda_i\right). \tag{270}$$

We can express the ESD, $\tilde{\rho}_{\mathbf{A}}(\lambda)$, of the Student Correlation matrix $\mathbf{A}_2$ in terms of the Dirac delta-function, $\delta(x)$, as

$$\tilde{\rho}_{\mathbf{A}}(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_i). \tag{271}$$

Using this, the Expected Value of the Log-Determinant in 270 can be expressed in terms of the ESD of $\mathbf{A}_2$ as

$$\begin{aligned}
\left\langle \ln \det\left(\mathbf{\Delta} - \mathbf{\Lambda}\right)^{-1/2} \right\rangle_{\mathbf{A}_2} &= -\frac{1}{2} \sum_{\mu=1}^{M} \sum_{i=1}^{N} \int d\lambda \ln(\vartheta_\mu - \lambda) \delta(\lambda - \lambda_i) \\
&= -\frac{1}{2} \sum_{\mu=1}^{M} \int d\lambda \ln(\vartheta_\mu - \lambda) \sum_{i=1}^{N} \delta(\lambda - \lambda_i) \\
&= -\frac{1}{2} \sum_{\mu=1}^{M} \int d\lambda \ln(\vartheta_\mu - \lambda) N \tilde{\rho}_{\mathbf{A}}(\lambda).
\end{aligned} \tag{272}$$

Let us insert this back into our expression for the generating function, 268, giving $\mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}]$ in terms of the ESD $\tilde{\rho}_{\mathbf{A}}$ as

$$\mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}] = \exp\left\{ \frac{N}{2} \Big[M \ln \frac{2\pi}{\beta} - \sum_{\mu=1}^{M} \int d\lambda \ln(\vartheta_\mu - \lambda) \tilde{\rho}_{\mathbf{A}}(\lambda) \Big] \right\}. \tag{273}$$

We can now replace the sum over the N eigenvalues λ_i with an integral over the limiting ESD, $\rho(\lambda)$, to obtain

$$\rho_{\mathbf{A}}^{\infty}(\lambda) = \lim_{N \to \infty} \tilde{\rho}_{\mathbf{A}}(\lambda). \tag{274}$$

Observe that this effectively means that we are taking a large-N limit, $N \gg 1$. This lets us write the Expected Value of the generating function $\mathbb{Z}^{IZ}$ in 273 as

$$\lim_{N\gg 1} \mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}] = \exp\left\{\frac{N}{2}\Big[M \ln\frac{2\pi}{\beta} - \sum_{\mu=1}^{M}\int d\lambda \ln(\vartheta_\mu - \lambda)\rho_{\mathbf{A}}^{\infty}(\lambda)\Big]\right\} \tag{275}$$

Step 1.4 Using the Self-Averaging Property,

$$\ln \mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}] \simeq \left\langle \ln \mathbb{Z}^{IZ}\right\rangle_{\mathbf{A}_1}, \tag{276}$$

It follows from Eqn. 275 that

$$\lim_{N\gg 1} \ln \mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}] \simeq \frac{NM}{2}\ln\frac{2\pi}{\beta} - \frac{N}{2}\sum_{\mu=1}^{M}\int d\lambda \ln(\vartheta_\mu - \lambda)\rho_{\mathbf{A}}^{\infty}(\lambda). \tag{277}$$

The N-dependence now cancels out, and we are left an approximate expression due to the remaining dependence of the continuum limiting density $\rho_{\mathbf{A}}^{\infty}(\lambda)$ (for $\mathbf{A} = \mathbf{A}_2$)

$$\lim_{N\gg 1} \frac{2}{N}\ln \mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}] = M\ln\frac{2\pi}{\beta} - \sum_{\mu=1}^{M}\int d\lambda \ln(\vartheta_\mu - \lambda)\rho_{\mathbf{A}}^{\infty}(\lambda). \tag{278}$$

This completes the derivation of Eqn. 254; we have an expression for the expected value of $\mathbb{Z}^{IZ}$, evaluated in the large-N (continuum) limit.

A.6.3 Step 2: The Saddle Point Approximation (SPA): Explicitly forming the Large Deviation Principle (LDP)

We now evaluate $\mathbb{E}_{\mathbf{A}}[\mathbb{Z}^{IZ}]$ in Eqn. 255 as $\mathbb{E}_{\mathbf{A}_2}[\mathbb{Z}^{IZ}]$ to establish Eqn. 256, .

Using the LDP (and following similar approaches in spin glass theory [92]), below we will show that we can write the expected value of $\mathbb{Z}^{IZ}$ in terms of $d\mu(\mathbf{X})$ now (which is equivalent to $d\mu(\mathbf{A}_1)$) and in the large-N approximation, as

$$\lim_{N\gg 1} \mathbb{E}_{\mathbf{A}_1}[\mathbb{Z}^{IZ}] = \int \exp\left(\beta N \operatorname{Tr}\left[\mathbb{G}_{\mathbf{A}}(\mathbf{X})\right] - NI(\mathbf{X}) + o(N)\right) d\mu(\mathbf{X}) \tag{279}$$

where $I(\mathbf{X})$ is Rate Function, defined below, and $\mathbb{G}_{\mathbf{A}}(\mathbf{X})$ is what we are eventually solving for.

Step 2.0 We start with the *expected* Partition Function

$$\mathbb{E}_{\mathbf{A}}[\mathbb{Z}^{IZ}] = \int d\mu(\mathbf{A}) \int d\mu(\mathbf{W}) \exp\Big[\tfrac{\beta}{2}\operatorname{Tr}\left[\mathbf{W}^\top \mathbf{A}\,\mathbf{W}\right] - \tfrac{\beta}{2}\operatorname{Tr}\left[\mathbf{W}\mathbf{D}\mathbf{W}^\top\right]\Big]. \tag{280}$$

The average over $\mathbf{A}$ affects only the first exponential; applying the SPA, we **define** a matrix function $\mathcal{G}(\mathbf{X})$, depending solely on $\mathbf{X} = \frac{1}{N}\mathbf{W}^\top\mathbf{W}$, by

$$\int d\mu(\mathbf{A})\, \exp\Big[\tfrac{\beta}{2}\operatorname{Tr}\left[\mathbf{W}^\top \mathbf{A}\,\mathbf{W}\right]\Big] = \exp\Big[\tfrac{\beta N}{2}\operatorname{Tr}\left[\mathcal{G}(\mathbf{X})\right]\Big]. \tag{281}$$

which will be valid in the large-N approximation below.

Inserting (281) into (280) gives

$$\mathbb{E}_{\mathbf{A}}[\mathbb{Z}^{IZ}] = \int d\mu(\mathbf{W}) \exp\Big[\tfrac{\beta N}{2}\operatorname{Tr}\left[\mathcal{G}(\mathbf{X})\right] - \tfrac{\beta}{2}\operatorname{Tr}\left[\mathbf{W}\mathbf{D}\mathbf{W}^\top\right]\Big]. \tag{282}$$

The now need to determine an explicit form for $\mathcal{G}(\mathbf{X})$. We introduce a new change of measure, $d\mu(\mathbf{W}) \to d\mu(\mathbf{X})$. Then, we show this lets us express $\mathbb{E}_{\mathbf{A}_1}[\mathbb{Z}^{IZ}]$ as $\mathbb{E}_{\mathbf{X}}[\mathbb{Z}^{IZ}]$ and to express it using the LDP. Next, we apply a SPA to solve for $\mathcal{G}^{max} = max\ \mathbb{G}_{\mathbf{A}}$. Importantly, we also show how to incorporate the inverse-Temperature β, which is new.

Step 2.1 To define the transformation $d\mu(\mathbf{W}) \to d\mu(\mathbf{X})$, where (recall) $\mathbf{X} = \frac{1}{N}\mathbf{W}^\top\mathbf{W}$, we use the (again) the integral representation of the Dirac delta-function $\delta(x)$:

$$\delta(x) := \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\hat{x}x} d\hat{x}. \tag{283}$$

This lets us express the transformation of measure $d\mu(\mathbf{W}) \to d\mu(\mathbf{X})$ (approximately) as

$$\begin{aligned} d\mu(\mathbf{W}) &:= \delta(\frac{1}{2}\operatorname{Tr}\left[N\mathbf{X} - \mathbf{W}^\top\mathbf{W}\right])d\mu(\mathbf{X}) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\frac{1}{2}\operatorname{Tr}[\hat{X}(N\mathbf{X}-\mathbf{W}^\top\mathbf{W})]} d\mu(\mathbf{X})d\mu(\hat{X}), \end{aligned} \tag{284}$$

where $\hat{X}$ is a scalar (or really matrix of scalars), and we have a 1/2 term for mathematical consistency below. [45]

Step 2.2 Next, we take a Wick Rotation, [46] $i\hat{\mathbf{X}} \to \hat{\mathbf{X}}$, so that the terms under the integral are all real (not complex), giving:

$$d\mu(\mathbf{W}) = \mathcal{N}_{\text{Wick}} \int_{-i\infty}^{i\infty} e^{-\frac{1}{2}\operatorname{Tr}[\hat{X}(N\mathbf{X}-\mathbf{W}^\top\mathbf{W})]} d\mu(\mathbf{X})d\mu(\hat{X}). \tag{285}$$

where $d\mu(\hat{X})$ is a measure over $\frac{M(M-1)}{2}$ Lagrange multipliers, and the normalization is

$$\mathcal{N}_{Wick} = (\frac{1}{2\pi i})^{\frac{M(M-1)}{2}}$$

.

[The minus sign has not been propagated through yet and may be confusing; I will fix shortly. The final result is not changed.]

Step 2.3 We now insert 285 into 282, which lets express $\mathbb{E}_{\mathbf{A}}[\mathbb{Z}^{IZ}]$ as an integral over the Teacher Correlation matrices

$$\begin{aligned} \mathbb{E}_{\mathbf{A}}[\mathbb{Z}^{IZ}] &= \mathcal{N}_{\text{Wick}} \int_{\mathbf{X}}\int_{-i\infty}^{i\infty} e^{N\frac{\beta}{2}\operatorname{Tr}[\mathcal{G}(\mathbf{X})]+\frac{N}{2}\operatorname{Tr}[\hat{X}\mathbf{X}]} e^{-\frac{1}{2}\operatorname{Tr}[\hat{X}\mathbf{W}^\top\mathbf{W}]} e^{\frac{\beta}{2}\operatorname{Tr}[\mathbf{W}\mathbf{D}\mathbf{W}^\top]} d\mu(\hat{X})d\mu(\mathbf{X}) \\ &= \mathcal{N}_{\text{Wick}} \int_{\mathbf{X}}\int_{-i\infty}^{i\infty} e^{N\frac{\beta}{2}\operatorname{Tr}[\mathcal{G}(\mathbf{X})]+\frac{N}{2}\operatorname{Tr}[\hat{X}\mathbf{X}]} e^{-\frac{1}{2}\operatorname{Tr}[\mathbf{W}\hat{X}\mathbf{W}^\top]+\frac{\beta}{2}\operatorname{Tr}[\mathbf{W}\mathbf{D}\mathbf{W}^\top]} d\mu(\hat{X})d\mu(\mathbf{X}) \\ &= \mathcal{N}_{\text{Wick}} \int_{\mathbf{X}}\int_{-i\infty}^{i\infty} e^{N\frac{\beta}{2}\operatorname{Tr}[\mathcal{G}(\mathbf{X})]+\frac{N}{2}\operatorname{Tr}[\hat{X}\mathbf{X}]} e^{\frac{1}{2}\operatorname{Tr}[\mathbf{W}(\beta\mathbf{D}-\hat{X})\mathbf{W}^\top]} d\mu(\hat{X})d\mu(\mathbf{X}). \end{aligned} \tag{286}$$

Step 2.4 We can now rearrange terms to make this expression look like the Eqn. 279

In Large Deviations Theory, the Rate Function is defined by the Legendre Transform,

$$\mathcal{I}(\mathbf{X}) = \sup_{\check{\mathbf{X}}}\left[Tr\frac{1}{2}\mathbf{X}^\top\check{\mathbf{X}} - \ln\mathbb{M}(\check{\mathbf{X}})\right], \tag{287}$$

[45] The full change of measure would require a delta function constraint for each matrix element $X_{i,j}$, i.e., $\delta\left(\frac{1}{2}N\left(X_{i,j} - [\mathbf{W}^\top\mathbf{W}]_{i,j}\right)\right)$. Here, we assume the Trace constraint is sufficient for our level of rigor.

[46] The Wick rotation converts an oscialltory integral into an exponentially decaying one which should be well defined. Technically, this is an analytic continuation which needs to be checked, but following standard practice in physics we will assume the resulting integral is analytic and therefore well defined and we will proceed onwards.

where $\mathbb{M}(\check{\mathbf{X}})$ is the Moment Generating Function, $\ln \mathbb{M}(\check{\mathbf{X}})$, is the Cumulant Generating Function, and and $\check{\mathbf{X}}$ is a (matrix of) *Lagrange Multiplier*(s). $\mathbb{M}(\check{\mathbf{X}})$ is defined in terms of the (unnormalized) density $p(\mathbf{x})$ as

$$\mathbb{M}(\check{\mathbf{X}}) = \exp\left(\frac{1}{2}\mathbf{x}^\top \check{\mathbf{X}}\mathbf{x}\right), p(\mathbf{x})d\mathbf{x} \tag{288}$$

which, in term, is defined in terms of the source matrix $\mathbf{D}$,

$$p(\mathbf{x}) = \exp\left(-\tfrac{1}{2}\mathbf{x}^\top \beta \mathbf{D}\mathbf{x}\right). \tag{289}$$

The moment generating function $\mathbb{M}(\check{\mathbf{X}})$ is then given by

$$\mathbb{M}(\check{\mathbf{X}}) = \int \exp\left(-\frac{1}{2}\mathbf{x}^T(\beta\mathbf{D} - \check{\mathbf{X}})\mathbf{x}\right) d\mathbf{x} = (2\pi)^{\frac{M}{2}} \det\left(\beta\mathbf{D} - \check{\mathbf{X}}\right)^{-\frac{1}{2}}. \tag{290}$$

Step 2.5 The Saddle Point Approximation (SPA) can be used to solve for $\mathcal{I}(\check{\mathbf{X}})$ by solving for the stationary conditions

$$\frac{\partial}{\partial \check{\mathbf{X}}} I(\mathbf{X}, \check{\mathbf{X}}) = 0. \tag{291}$$

First, let us compute $\ln \mathbb{M}(\check{\mathbf{X}})$ as:

$$\ln \mathbb{M}(\check{\mathbf{X}}) = \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(\beta\mathbf{D} - \check{\mathbf{X}}\right). \tag{292}$$

Substituting this into the expression for the Legendre transform, we obtain:

$$I(\mathbf{X}, \check{\mathbf{X}}) = \sup_{\check{\mathbf{x}}}\left[\frac{1}{2}\operatorname{Tr}\left[\mathbf{X}\check{\mathbf{X}}\right] - \frac{M}{2}\ln(2\pi) + \frac{1}{2}\ln\det\left(\beta\mathbf{D} - \check{\mathbf{X}}\right)\right]. \tag{293}$$

The supremum of this expression is attained at the value of $\check{\mathbf{X}}$ that satisfies:

$$\frac{\partial}{\partial \check{\mathbf{X}}}\left[\frac{1}{2}\operatorname{Tr}\left[\mathbf{X}\check{\mathbf{X}}\right] + \frac{1}{2}\ln\det\left(\beta\mathbf{D} - \check{\mathbf{X}}\right)\right] = 0. \tag{294}$$

Taking the derivative, we obtain

$$\frac{1}{2}\mathbf{X} + \frac{1}{2}(\beta\mathbf{D} - \check{\mathbf{X}})^{-1} = 0, \tag{295}$$

which simplifies to:

$$\mathbf{X} = (\beta\mathbf{D} - \check{\mathbf{X}})^{-1} \quad \Rightarrow \quad \check{\mathbf{X}} = \beta\mathbf{D} - \mathbf{X}^{-1}. \tag{296}$$

Substituting $\check{\mathbf{X}} = \beta\mathbf{D} - \mathbf{X}^{-1}$ back into the expression for $I(\mathbf{X})$, we obtain:

$$I(\mathbf{X}) = \frac{1}{2}\left[\operatorname{Tr}\left[\mathbf{X}(\beta\mathbf{D} - \mathbf{X}^{-1})\right] - \frac{M}{2}\ln(2\pi) + \frac{1}{2}\ln\det\left(\mathbf{X}^{-1}\right)\right]. \tag{297}$$

$$\operatorname{Tr}\left[\mathbf{X}\beta\mathbf{D} - \mathbf{I}\right] = \operatorname{Tr}\left[\mathbf{X}\beta\mathbf{D}\right] - N, \tag{298}$$

$$\ln\det\left(\mathbf{X}^{-1}\right) = -\ln\det\left(\mathbf{X}\right), \tag{299}$$

we get:

$$I(\mathbf{X}) = \frac{1}{2}\left[\operatorname{Tr}\left[\mathbf{X}\beta\mathbf{D}\right] - \ln\det\left(\mathbf{X}\right) - M - M\ln(2\pi)\right]. \tag{300}$$

Finally, we express $I(\mathbf{X})$ in the form:

$$I(\mathbf{X}) = \frac{1}{2}\left[-M(1 + \ln(2\pi)) + \operatorname{Tr}\left[\mathbf{X}\beta\mathbf{D}\right] - \ln\det\left(\mathbf{X}\right)\right]. \tag{301}$$

[THE DERIVATION ABOVE FOR $\mathcal{G}(\mathbf{X})$ may have some TYPOS: CHECK]

Step 2.6

$$\beta\mathcal{G}(\mathbf{X}) = \mathbb{M}(1 + \ln 2\pi) + \beta \operatorname{Tr}\left[\mathbb{G}_{\mathbf{A}}(\mathbf{X})\right] - \operatorname{Tr}\left[\mathbf{X}\beta\mathbf{D}\right] + \ln\det\left(\mathbf{X}\right). \tag{302}$$

We restrict our solution to those where $\mathbf{X}$ and $\beta\mathbf{D}$ can be diagonalized simultaneously. In particular, this lets us write

$$\operatorname{Tr}\left[\mathbf{X}\beta\mathbf{D}\right] = \sum_{\mu=1}^{M} \beta\delta_\mu\lambda_\mu, \tag{303}$$

where $\beta\delta_\mu$ and λ_μ denote the eigenvalues of $\mathbf{X}$ and $\beta\mathbf{D}$, resp.

We can now write the maximum value of $\mathbb{G}_{\mathbf{A}}$, $\mathcal{G}^{max}$, as

$$\beta\mathcal{G}^{max} = M\left(1 + \ln\frac{2\pi}{\beta}\right) - \sum_{\mu=1}^{M} \min_{\beta\delta_\mu}\left[\beta\delta_\mu\lambda_\mu - \beta\mathbb{G}_{\mathbf{A}}(\lambda_\mu) + \ln\lambda_\mu\right]. \tag{304}$$

A.6.4 Expressing the Norm Generating Function ($\mathbb{G}_{\mathbf{A}}(\lambda)$) as the Integrated R-transform ($R(z)$) of the Correlation Matrix ($\mathbf{A}$)

Having completed both steps, let us combine Eqns. 254, 278 with 256 and 304. We follow the first arguments by Tanaka[82] (which follows Cherrier[119]).

$$M\ln(\frac{2\pi}{\beta}) - \sum_{\mu=1}^{M}\int \ln(\beta\delta_\mu - \lambda)\rho_{\mathbf{A}}^{\infty}(\lambda)d\lambda = M\left(1 + \ln\frac{2\pi}{\beta}\right) - \sum_{\mu=1}^{M}\min_{\beta\delta_\mu}\left[\beta\delta_\mu\lambda_\mu - \beta\mathbb{G}_{\mathbf{A}}(\lambda_\mu) + \ln\lambda_\mu\right]. \tag{305}$$

By cancelling the $\ln\frac{2\pi}{\beta}$ term from both sides, we obtain

$$-\sum_{\mu=1}^{M}\int \ln(\beta\delta_\mu - \lambda)\rho_{\mathbf{A}}^{\infty}(\lambda)d\lambda = M - \sum_{\mu=1}^{M}\min_{\beta\delta_\mu}\left[\beta\delta_\mu\lambda_\mu - \beta\mathbb{G}_{\mathbf{A}}(\lambda_\mu) + \ln\lambda_\mu\right]. \tag{306}$$

Since this is true for every μ, we can solve this for any arbitrary eigenvalue λ_μ.

Dropping the μ subscript, we have the following identity:

$$\min_{\delta}\left[\beta\delta\lambda - \beta\mathbb{G}_{\mathbf{A}}(\lambda) + \ln\lambda\right] = 1 - \int \ln(\beta\delta - \lambda)\rho_{\mathbf{A}}^{\infty}(\lambda)d\lambda. \tag{307}$$

We need to invert 307 in order to find $\beta\mathbb{G}_{\mathbf{A}}(\lambda)$. If we choose the eigenvalues of $\mathbf{D}$ such that $\beta\delta_\mu > \lambda_{max}$ for all μ, then this relation is concave and therefore invertible via a Legendre transform.

This gives

$$\beta\mathbb{G}_{\mathbf{A}}(\lambda) = \beta\delta(\lambda)\lambda - \int \ln[\beta\delta(\lambda) - \lambda]\rho_{\mathbf{A}}^{\infty}(\lambda)d\lambda - \ln\lambda - 1, \tag{308}$$

where we need to define $\beta\delta(\lambda)$, which (not to be confused with the Dirac delta-function), describes the functional dependence between the eigenvalues of the source matrix $\mathbf{D}$ and the Student Correlation Matrix $\mathbf{A}$.

$\mathbb{G}_{\mathbf{A}}(\lambda)$ is computed by minimizing over δ, ensuring the relationship holds for the entire spectrum. So let us take the derivative of $\beta\mathbb{G}_{\mathbf{A}}$ w/r.t λ. Term by term, this gives:

$$\frac{d}{d\lambda}\beta\delta(\lambda)\lambda = \beta\delta(\lambda) + \frac{d\beta\delta(\lambda)}{d\lambda}\lambda \tag{309}$$

$$\frac{d}{d\lambda}\ln\lambda = \frac{1}{\lambda} \tag{310}$$

$$\frac{d}{d\lambda}\int \ln[\beta\delta(\lambda)-\lambda]\rho_{\mathbf{A}}^{\infty}(\lambda)d\lambda = \int \frac{d}{d\lambda}\ln[\beta\delta(\lambda)-\lambda]\rho_{\mathbf{A}}^{\infty}(\lambda)d\lambda \tag{311}$$
$$= \int \frac{d\beta\delta(\lambda)}{d\lambda}\frac{\rho_{\mathbf{A}}^{\infty}(\lambda)}{\beta\delta(\lambda)-\lambda}d\lambda$$
$$= \frac{d\beta\delta(\lambda)}{d\lambda}\int \frac{\rho_{\mathbf{A}}^{\infty}(\lambda)}{\beta\delta(\lambda)-\lambda}d\lambda$$

We can now simplify by defining $\delta(\lambda)$ implicitly by the integral relation

$$\lambda = \int \frac{\rho_{\mathbf{A}}^{\infty}(\lambda)}{\beta\delta(\lambda)-\lambda}d\lambda. \tag{312}$$

Combining terms, this gives

$$\frac{d\beta\mathbb{G}_{\mathbf{A}}(\lambda)}{d\lambda} = \beta\delta(\lambda) - \frac{1}{\lambda}, \tag{313}$$

Inverting the derivative, we obtain an integral equation for $\beta\mathbb{G}_{\mathbf{A}}(\lambda)$

$$\beta\mathbb{G}_{\mathbf{A}}(\lambda) = \int_0^{\lambda}\left(\beta\delta(z) - \frac{1}{z}\right)dz. \tag{314}$$

Notice since $\beta\delta(\lambda) \approx \frac{1}{\lambda}$ for $\lambda \ll 1$, then as $\mathbb{G}_{\mathbf{A}}(0) = 0$ and we set the lower integrand to 0 (for now). Even though Tanaka's original proof assumes an analytic continuation without branch cuts, a heavy-tailed spectrum merely shifts the lower limit of the R-transform integral, so the expression for $\beta\mathcal{G}(\lambda)$ continues to hold.

To further connect these to the R-transform $R_{\mathbf{A}}(z)$, we recall that the Cauchy-Stieltjes (or just Cauchy See 133) transform $\mathcal{C}_{\mathbf{A}}(z)$ is given by:

$$\mathcal{C}_{\mathbf{A}}(z) = \int \frac{\rho_{\mathbf{A}}(\lambda)}{z-\lambda}d\lambda. \tag{315}$$

The relationship between the Cauchy transform and the R-transform is then expressed as:

$$\mathcal{C}_{\mathbf{A}}\left(R_{\mathbf{A}}(z) + \frac{1}{z}\right) = z, \tag{316}$$

which implies:

$$\beta\mathbb{G}_{\mathbf{A}}(\lambda) = \int_0^{\lambda} R_{\mathbf{A}}(z)dz. \tag{317}$$

WLOG, as mentioned earlier, we can replace the lower bound on λ from $0 \to \lambda_{min}^{ECS}$ to obtain

$$\beta\mathbb{G}_{\mathbf{A}}(\lambda) = \int_{\lambda_{min}^{ECS}}^{\lambda} R_{\mathbf{A}}(z)dz. \tag{318}$$

where λ_{min}^{ECS} corresponds to the start of the Effective Correlation Space (`ECS`).

A.6.5 Selecting $\mathbf{A} \coloneqq \mathbf{A}_1$ instead of $\mathbf{A}_2$

In principle, we could have selected $\mathbf{A} \coloneqq \mathbf{A}_1 = \frac{1}{N}\mathbf{S}^{\top}\mathbf{S}$ for the Student Correlation matrix, thereby avoiding the discussion on the Duality of Measures altogether. Doing this, however, would make $\mathbf{A}$ $M \times M$, thereby require defining the Source Matrix $\mathbf{D}$ as an $N \times N$ matrix, with presumably $N - M$ zero eigenvalues. This would cause $\mathbf{D}$ to violate the condition $\vartheta_{\mu} > \beta\lambda$ for all eigenvalues λ of $\mathbf{A}_1$. In this case, it would be challenging to define the large-N limit.

A.7 The Inverse-Wishart (IW) Model

In this section, we rederive the integral $G(\lambda)[IW]$ for the Inverse Wishart (IW) model, focusing on the branch cut starting at $z = \kappa/2$ and extending to infinity. This branch cut corresponds to the support of the ESD in this region. We will:

1. Explain the presence of the branch cut and its implications.
2. Show that $R(z)[IW]$ becomes complex along this branch cut because the term under the square root becomes negative.
3. Perform the integral $G(\lambda)[IW]$, showing all steps.
4. Compute the modulus $|G(\lambda)[IW]| = \sqrt{G(\lambda)[IW]^* G(\lambda)[IW]}$ to obtain a real-valued estimate, analogous to a probability estimate.

[Note: Below, we may want the Real part of $G(\lambda)[IW]$ and not the modulus. Stil thinking on this. Also, we may want to add additional parameters to the IW model to account for the dimension of the matrix. This changes things bit.]

A.7.1 The Branch Cut in the IW Model

The R-transform for the IW model is given by:

$$R(z)[IW] = \frac{\kappa - \sqrt{\kappa(\kappa - 2z)}}{z}, \tag{319}$$

where $\kappa > 0$ is a parameter related to the dimensions of the random matrices under consideration. The function $\sqrt{\kappa(\kappa - 2z)}$ introduces a branch point at $z = \kappa/2$ because the argument of the square root becomes zero at this point:

$$\kappa - 2z = 0 \quad \Rightarrow \quad z = \frac{\kappa}{2}. \tag{320}$$

For $z > \kappa/2$, the argument $\kappa - 2z$ becomes negative, and thus the square root becomes imaginary. This leads to a branch cut starting at $z = \kappa/2$ and extending to $z = \infty$ along the real axis. This branch cut affects the analyticity of $R(z)[IW]$, and it must be carefully considered in the integral $G(\lambda)[IW]$.

A.7.2 $R(z)[IW]$ is Complex Along the Branch Cut

For $z > \kappa/2$, we have:

$$\kappa - 2z < 0 \quad \Rightarrow \quad \sqrt{\kappa(\kappa - 2z)} = \sqrt{-\kappa(2z - \kappa)} = i\sqrt{\kappa(2z - \kappa)}. \tag{321}$$

Therefore, $R(z)[IW]$ becomes complex:

$$R(z)[IW] = \frac{\kappa - i\sqrt{\kappa(2z - \kappa)}}{z} = \frac{\kappa}{z} - i\frac{\sqrt{\kappa(2z - \kappa)}}{z}. \tag{322}$$

This expression shows that $R(z)[IW]$ has both real and imaginary parts when $z > \kappa/2$.

A.7.3 Calculation of $G(\lambda)[IW]$

We aim to compute the integral:

$$G(\lambda)[IW] = \int_{z_0}^{\lambda} R(z)[IW], dz, \tag{323}$$

where $z_0 \geq \kappa/2$.

Integrating the Real Part. First, we consider the real part of $R(z)[IW]$:

$$\mathrm{Re}[R(z)[IW]] = \frac{\kappa}{z}. \tag{324}$$

The integral of this real part is:

$$G_{\mathrm{real}}(\lambda)[IW] = \int_{z_0}^{\lambda} \frac{\kappa}{z}, dz = \kappa \left[\ln z\right]_{z_0}^{\lambda} = \kappa \left(\ln \lambda - \ln z_0\right). \tag{325}$$

Integrating the Imaginary Part. Next, consider the imaginary part:

$$\mathrm{Im}[R(z)[IW]] = -\frac{\sqrt{\kappa(2z-\kappa)}}{z}. \tag{326}$$

If we let $u = 2z - \kappa$, then:

$$z = \frac{u+\kappa}{2}, \quad dz = \frac{du}{2}. \tag{327}$$

Substituting this into the imaginary part:

$$\mathrm{Im}[R(z)[IW]] = -\frac{\sqrt{\kappa u}}{\frac{u+\kappa}{2}} = -\frac{2\sqrt{\kappa u}}{u+\kappa}, \tag{328}$$

the integral becomes:

$$G_{\mathrm{imag}}(\lambda)[IW] = \int_{u_0}^{u_\lambda} -\frac{2\sqrt{\kappa u}}{u+\kappa} \cdot \frac{du}{2} = -\int_{u_0}^{u_\lambda} \frac{\sqrt{\kappa u}}{u+\kappa} du, \tag{329}$$

where $u_0 = 2z_0 - \kappa$ and $u_\lambda = 2\lambda - \kappa$. If we simplify the integrand:

$$\sqrt{\kappa u} = \sqrt{\kappa}\sqrt{u}, \tag{330}$$

then the integral becomes:

$$G_{\mathrm{imag}}(\lambda)[IW] = -\sqrt{\kappa} \int_{u_0}^{u_\lambda} \frac{\sqrt{u}}{u+\kappa}, du. \tag{331}$$

This integral can be evaluated using standard integral formulas. We will compute it step by step.

Evaluating the Integral. Consider the integral:

$$I = \int \frac{\sqrt{u}}{u+\kappa}, du. \tag{332}$$

We can use the following integral formula:

$$\int \frac{\sqrt{u}}{u+a}, du = 2\sqrt{u} - 2a\tan^{-1}\left(\frac{\sqrt{u}}{\sqrt{a}}\right) + C, \tag{333}$$

where $a > 0$ and $u > 0$. Applying this formula, we get:

$$I = 2\sqrt{u} - 2\kappa\tan^{-1}\left(\frac{\sqrt{u}}{\sqrt{\kappa}}\right) + C. \tag{334}$$

Therefore, the imaginary part of $G(\lambda)[IW]$ is:

$$\begin{aligned} G_{\text{imag}}(\lambda)[IW] &= -\sqrt{\kappa}\left[2\sqrt{u} - 2\kappa\tan^{-1}\left(\frac{\sqrt{u}}{\sqrt{\kappa}}\right)\right]{u_0}^{u\lambda} \\ &= -\sqrt{\kappa}\left(\left[2\sqrt{u_\lambda} - 2\kappa\tan^{-1}\left(\frac{\sqrt{u_\lambda}}{\sqrt{\kappa}}\right)\right] - \left[2\sqrt{u_0} - 2\kappa\tan^{-1}\left(\frac{\sqrt{u_0}}{\sqrt{\kappa}}\right)\right]\right) \\ &= -2\sqrt{\kappa}\left(\sqrt{u_\lambda} - \sqrt{u_0}\right) + 2\kappa^{3/2}\left(\tan^{-1}\left(\frac{\sqrt{u_\lambda}}{\sqrt{\kappa}}\right) - \tan^{-1}\left(\frac{\sqrt{u_0}}{\sqrt{\kappa}}\right)\right). \end{aligned} \tag{335}$$

Combining Real and Imaginary Parts. Combine the real and imaginary parts to obtain $G(\lambda)[IW]$:

$$G(\lambda)[IW] = G_{\text{real}}(\lambda)[IW] + iG_{\text{imag}}(\lambda)[IW]. \tag{336}$$

Substituting the expressions:

$$\begin{aligned} G(\lambda)[IW] &= \kappa\left(\ln\lambda - \ln z_0\right) \\ &\quad + i\left(-2\sqrt{\kappa}\left(\sqrt{u_\lambda} - \sqrt{u_0}\right) + 2\kappa^{3/2}\left(\tan^{-1}\left(\frac{\sqrt{u_\lambda}}{\sqrt{\kappa}}\right) - \tan^{-1}\left(\frac{\sqrt{u_0}}{\sqrt{\kappa}}\right)\right)\right). \end{aligned} \tag{337}$$

Recall that $u = 2z - \kappa$, so:

$$\sqrt{u} = \sqrt{2z - \kappa}. \tag{338}$$

Therefore, we can write $G(\lambda)[IW]$ as:

$$\begin{aligned} G(\lambda)[IW] &= \kappa\ln\left(\frac{\lambda}{z_0}\right) - 2i\sqrt{\kappa}\left(\sqrt{2\lambda - \kappa} - \sqrt{2z_0 - \kappa}\right) \\ &\quad + 2i\kappa^{3/2}\left(\tan^{-1}\left(\frac{\sqrt{2\lambda - \kappa}}{\sqrt{\kappa}}\right) - \tan^{-1}\left(\frac{\sqrt{2z_0 - \kappa}}{\sqrt{\kappa}}\right)\right). \end{aligned} \tag{339}$$

A.7.4 Computing the Modulus $|G(\lambda)[IW]|$

To obtain a real-valued estimate, we compute the modulus of $G(\lambda)[IW]$:

$$|G(\lambda)[IW]| = \sqrt{\left(\text{Re}[G(\lambda)[IW]]\right)^2 + \left(\text{Im}[G(\lambda)[IW]]\right)^2}. \tag{340}$$

Calculating the Real Part Square. The real part is:

$$\mathrm{Re}[G(\lambda)[IW]] = \kappa \ln\left(\frac{\lambda}{z_0}\right). \tag{341}$$

Therefore,

$$(\mathrm{Re}[G(\lambda)[IW]])^2 = \kappa^2 \left(\ln\left(\frac{\lambda}{z_0}\right)\right)^2. \tag{342}$$

Calculating the Imaginary Part Square. The imaginary part is:

$$\mathrm{Im}[G(\lambda)[IW]] = -2\sqrt{\kappa}\left(\sqrt{2\lambda-\kappa} - \sqrt{2z_0-\kappa}\right) \qquad +2\kappa^{3/2}\left(\tan^{-1}\left(\frac{\sqrt{2\lambda-\kappa}}{\sqrt{\kappa}}\right) - \tan^{-1}\left(\frac{\sqrt{2z_0-\kappa}}{\sqrt{\kappa}}\right)\right). \tag{343}$$

Let's denote:

$$A = -2\sqrt{\kappa}\left(\sqrt{2\lambda-\kappa} - \sqrt{2z_0-\kappa}\right),\ B \quad = 2\kappa^{3/2}\left(\tan^{-1}\left(\frac{\sqrt{2\lambda-\kappa}}{\sqrt{\kappa}}\right) - \tan^{-1}\left(\frac{\sqrt{2z_0-\kappa}}{\sqrt{\kappa}}\right)\right). \tag{344}$$

Then,

$$(\mathrm{Im}[G(\lambda)[IW]])^2 = (A+B)^2 = A^2 + 2AB + B^2. \tag{345}$$

Computing the Modulus. The modulus is:

$$|G(\lambda)[IW]| = \sqrt{\left(\kappa \ln\left(\frac{\lambda}{z_0}\right)\right)^2 + (A+B)^2}. \tag{346}$$

Interpretation. While the expression for $|G(\lambda)[IW]|$ appears complex, it encapsulates the cumulative effect of both the real and imaginary components of $G(\lambda)[IW]$. This modulus provides a real-valued estimate that is meaningful in the context of probability estimates.

A.7.5 Summary

By integrating $R(z)[IW]$ directly, including its complex components, we have obtained an explicit expression for $G(\lambda)[IW]$ as a complex function. Computing the modulus $|G(\lambda)[IW]|$ gives us a real-valued function that accurately captures the contribution of the tail of the ESD in the IW model. This approach accounts for the complex nature of $R(z)[IW]$ along the branch cut $z > \kappa/2$, and it provides a meaningful point estimate for further analysis.