What Is in Training?

Recoverable complexity, generalization, and overfitting risk.

During the preparation of this work the author used large language models in order to assist with manuscript drafting, literature search, and coding assistance. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Dr. Tamás Nagy

tnagyphd@gmail.com

Draft

Abstract

Modern training optimizes losses that are local, task-specific, and often blind to the amount of recoverable structure present in the data. We propose a complementary view: the spectral quality parameter $\rho > 1$ should be understood as a **structural fitness metric**. It measures not whether a model matches each training label or parameter entry, but whether it captures the correct degree of compressibility in the underlying object. This makes ρ especially relevant in partially non-identifiable inverse problems, spectral learning, and model selection under noise. Our central thesis is that ρ should be interpreted as a measure of **recoverable complexity**: if the energy envelope obeys $E^{\downarrow}_{(k)} \approx C\rho^{-k}$, then larger ρ implies fewer relevant modes, lower effective dimension, and therefore less room for variance-driven overfitting. We argue for six claims. First, ρ measures recoverable complexity, not raw performance. Second, high ρ implies a smaller structurally justified model class at fixed target accuracy. Third, matching ρ can reduce overfitting pressure because it penalizes structurally implausible complexity. Fourth, ρ is best used as a regularizer, model-selection statistic, or early-stopping diagnostic, not as a standalone objective. Fifth, ρ is not a universal training loss and is not overfitting-free by definition. Sixth, the right scientific question is often not "did we recover the exact object?" but "did we recover the same amount of structure?" We formalize these distinctions, connect ρ to rate-distortion, effective dimension, and representation size, relate the program to an existing proved spectral overfitting bound, and propose an experimental roadmap for validating when ρ-aware training truly improves generalization.

---

1. The Problem

1.1 Training Metrics Usually Measure the Wrong Thing

Most training pipelines optimize one of:

- prediction loss,
- parameter distance,
- calibration error,
- held-out accuracy,
- likelihood.

These are useful, but they answer a narrow question: *how well did the model fit the observed target?*

They do **not** directly answer:

- how much real structure is present in the data,
- how much structure the model has extracted,
- whether the model's complexity matches the object's true compressibility,
- whether low error came from signal capture or from noise fitting.

This distinction becomes critical in ill-posed or partially non-identifiable problems. Two models can fit the data equally well while implying very different internal objects. In such settings, entrywise closeness is often the wrong scientific target.

1.2 The Structural Question

The structural question is:

> Did the training process recover the correct **amount of compressible structure**?

We propose that ρ answers this question.

---

2. Definition

2.1 Spectral Quality

Let E_k denote a spectral energy profile, singular value sequence, or mode-energy sequence derived from data, a learned representation, or a recovered object. Let $E^{\downarrow}_{(k)}$ be its decreasing rearrangement. Define spectral quality by the decay law

$$E^{\downarrow}_{(k)} \approx C\rho^{-k}, \qquad \rho > 1.$$

Equivalently, if $\log E^{\downarrow}_{(k)}$ is approximately linear in k, then

$$\log E^{\downarrow}_{(k)} \approx \log C - k \log \rho.$$

So $-\log \rho$ is the slope of the spectral envelope in log scale.

2.2 Interpretation

- large ρ: energy concentrates rapidly in a few modes,
- moderate ρ: structure exists but is more distributed,
- $\rho \approx 1$: weak or non-compressible structure.

Thus ρ is not a local fit statistic. It is a one-number summary of **structural compressibility**.

2.3 Representation Size

By the Universal Spectral Representation logic,

$$N(\varepsilon) = \Theta\left(\frac{\log(1/\varepsilon)}{\log \rho}\right).$$

So ρ directly controls how many modes are needed for an ε-accurate representation.

This is the first reason ρ matters in training:

> ρ tells us what level of model complexity is structurally justified.

2.4 Recoverable Complexity

The key strengthening is to interpret ρ not only as smoothness, but as a bound on the amount of structure that can be stably recovered.

Define the effective spectral dimension at tolerance $\tau > 0$ by

$$K_{\text{eff}}(\tau) = \min\left\{K : \sum_{k>K} \left(E^{\downarrow}_{(k)}\right)^2 \leq \tau^2\right\}.$$

If the envelope satisfies

$$E^{\downarrow}_{(k)} \leq C\rho^{-k},$$

then its tail obeys

$$\sum_{k>K} \left(E^{\downarrow}_{(k)}\right)^2 \leq \frac{C^2\rho^{-2(K+1)}}{1-\rho^{-2}}.$$

Therefore

$$K_{\text{eff}}(\tau) = O\left(\frac{\log(1/\tau)}{\log \rho}\right).$$

This gives the structural chain motivating the paper:

larger $\rho \Longrightarrow$ faster spectral decay $\Longrightarrow$ smaller effective dimension $\Longrightarrow$ less recoverable complexity $\Longrightarrow$ less room for

This is not yet a universal theorem of generalization, but it is a precise mathematical program.

---

3. What Is in Training

3.1 A Structural Fitness Metric

We define a **structural fitness metric** as a statistic that measures whether a learned object has the right large-scale structure, even when many different microscopic realizations fit the observed data.

In this sense, ρ is a structural fitness metric because it asks whether the model and the data occupy the same compressibility regime.

Examples:

- in matrix or tensor learning, whether the learned representation has the same spectral decay as the data;
- in inverse problems, whether the recovered operator induces the same structural envelope as the true or observed object;
- in model distillation, whether the student retains the teacher's compressible structure;
- in training diagnostics, whether additional capacity is extracting signal or flattening into noise modes.

3.2 Not a Replacement for Task Loss

This does **not** mean that ρ replaces prediction loss, likelihood, or calibration error.

A model can have the right ρ and still be wrong locally.

Therefore:

$$\rho \text{ is a structural metric, not a complete task metric.}$$

The correct use is complementary, not substitutional.

3.3 Not a Universal Scalar for Truth

Many distinct objects can share the same ρ. Therefore:

- matching ρ does not identify the object,
- matching ρ does not guarantee semantic correctness,
- matching ρ does not guarantee pricing accuracy, classification accuracy, or causal validity.

It only says:

> the learned object has the same level of compressible structure.

That is already extremely useful, but it is not everything.

---

4. Why Helps With Overfitting

4.1 The Basic Mechanism

Overfitting occurs when the model allocates capacity to directions unsupported by stable structure.

In spectral language, this means:

- too many modes retained,
- too much energy placed into weak modes,
- a decay profile flatter than the data justifies.

Since ρ controls the speed of spectral decay, it provides a direct proxy for how quickly additional modes stop being worth modeling.

If the data has high ρ, a compact representation is justified. If the data has $\rho \approx 1$, any claim of highly stable fine structure is suspicious.

4.2 From to Generalization Pressure

The real point is not that ρ magically prevents overfitting. The point is that ρ changes the complexity regime in which overfitting becomes likely.

If the recoverable tail decays geometrically, then the learner only needs a logarithmic number of modes in the target precision. The variance burden of estimation is therefore paid over a much smaller active subspace. In contrast, when $\rho \approx 1$, the tail is flat, the effective dimension grows rapidly, and the learner must estimate many weak modes. That is exactly the regime in which noise fitting becomes attractive.

So the hypothesis is not:

> fit ρ and overfitting disappears.

It is:

> large ρ implies a smaller structurally justified model class, and a smaller justified model class should generalize better at fixed sample size.

4.3 Why This Is Promising

Suppose a learner keeps adding modes until training loss improves. If those new modes only flatten the decay tail without improving holdout structure, then the model is fitting noise. A ρ-aware criterion can resist this by favoring representations whose complexity remains consistent with the observed spectral envelope.

This suggests a general principle:

> Overfitting is often not "too much error reduction," but "too little structural decay."

4.4 What We Can Safely Claim

We can safely claim:

- matching ρ can reduce overfitting pressure,
- matching ρ can regularize complexity,
- matching ρ can improve model selection,
- matching ρ can detect when parameter RMSE is the wrong target.

We should **not** yet claim:

- that ρ alone eliminates overfitting,

- that ρ is always sufficient for generalization,
- that any ρ-matching procedure is automatically robust.

4.5 The Stronger Conjecture

The stronger and more interesting statement is:

Conjecture. *Whenever the target admits an error decomposition of the form*

$$E_{\text{gen}}(K) \leq E_{\text{approx}}(K;\rho) + E_{\text{est}}(K;n,\sigma),$$

with E_{approx} decaying geometrically in K at rate ρ, the optimal complexity and generalization risk are controlled primarily by ρ, sample size, and noise level rather than by the ambient parameter dimension.

That is the real generalization-theory program behind this paper.

---

5. What Does Not Solve

5.1 Is Not Overfitting-Free by Definition

It is tempting to say:

> "If we fit ρ, we avoid overfitting by definition."

This is too strong.

Why?

1. Different objects can share the same ρ.
2. Noise can distort the spectral envelope.
3. A model can match ρ globally while missing important local structure.
4. A single scalar cannot capture full geometry, signs, phase, or semantics.

Therefore ρ is not overfitting-free by definition. At most, it is **overfitting-resistant** when used properly.

5.2 Pure -Fitting Is Dangerous

If one optimizes only

$$\min_{\theta} |\rho_{\text{model}}(\theta) - \rho_{\text{data}}|,$$

the learner may produce objects with the correct compressibility and the wrong content.

So pure ρ-fitting is generally a bad objective.

5.3 The Better Role

The better role is:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_\rho \mathcal{L}_\rho + \lambda_{\text{stab}} \mathcal{L}_{\text{stability}} + \lambda_{\text{simp}} \mathcal{L}_{\text{complexity}},$$

where

$$\mathcal{L}_\rho = |\rho_{\text{model}} - \rho_{\text{data}}|$$

or a related penalty on the mismatch of spectral envelopes.

In words:

- task loss keeps the model useful,
- ρ keeps the model structurally honest,
- stability penalties keep it physically or statistically admissible,
- complexity penalties keep it parsimonious.

---

6. A Training Taxonomy for

6.1 as a Pre-Training Diagnostic

Before fitting any model, estimate ρ from the data.

Interpretation:

- high ρ: use compact models, aggressive compression, low-rank methods;
- moderate ρ: structure exists, but regularization matters;
- $\rho \approx 1$: the domain may be inherently weakly predictable.

This tells us what kind of training problem we are entering.

6.2 as a Capacity Selector

Use ρ to choose representation size:

$$K^*(\varepsilon) \approx \frac{\log(1/\varepsilon)}{\log \rho}.$$

This gives a principled alternative to blind hyperparameter sweeps.

6.3 as a Regularizer

Add a penalty when learned structure is flatter than observed structure. This discourages the learner from inventing weak modes with no stable support.

6.4 as an Early-Stopping Signal

Track ρ during training.

If training loss keeps falling while the learned spectrum flattens toward $\rho \approx 1$, that is a warning sign that the model is consuming noise.

6.5 as a Model-Selection Statistic

Given two models with similar validation performance, prefer the one whose spectral quality better matches the data and uses fewer structurally justified modes.

6.6 as an Inverse-Problem Objective

In inverse problems, exact parameter recovery may be impossible or ill-posed. There, ρ can be more scientifically meaningful than entrywise closeness.

This is exactly the situation in implied-generator recovery, where preserving holdout pricing and panel-level spectral quality is often more meaningful than matching the true matrix entry by entry.

---

7. The Overfitting Hypothesis

7.1 Main Hypothesis

We propose the following:

Hypothesis A. *When the target object is compressible and partially non-identifiable, ρ-aware training generalizes better than parameter-focused training at fixed sample size.*

This is plausible because parameter loss is sensitive to coordinate choice, while ρ tracks the object's structural complexity.

7.2 Stronger Conjecture

Conjecture B. *For model classes whose error decomposes into approximation error plus spectral tail error, constraining ρ reduces the variance term of generalization by limiting unsupported tail complexity.*

This is the mathematically interesting statement.

It is not yet proved here, but it is the right theorem to target.

7.3 A Special Case Already Exists

This paper is not starting from zero. A concrete spectral overfitting result already exists in the companion draft topics/publications/SPECTRAL_OVERFITTING.md.

In that setting, the generalization error for a K-mode spectral estimator obeys

$$E_{\text{gen}}(K) \leq C^2 \rho^{-2K} + \frac{K\sigma^2}{n},$$

and the optimal complexity is

$$K^* = \Theta\left(\frac{\log(n/\sigma^2)}{\log \rho}\right).$$

This is exactly the mechanism proposed here:

- ρ determines the decay of approximation error,
- the decay determines the effective usable dimension,
- the effective dimension determines the variance burden,
- and the variance burden controls overfitting risk.

So the present paper should be read as the **general conceptual wrapper** around a special-case theorem that is already substantially worked out.

7.4 What Would "Per Definition" Mean?

To say that ρ avoids overfitting **by definition** would require a theorem of the form:

$$\rho\text{-matched} \implies \text{generalization bound.}$$

We do not currently have that theorem in this generality.

So the honest position is:

> ρ is not overfitting-free by definition; it is a candidate structural control variable whose link to overfitting can and should be formalized.

---

8. Relation to Existing Concepts

8.1 Bias-Variance

Classical bias-variance says complexity must be traded off.

The ρ view says complexity is not arbitrary; it should be calibrated to the object's spectral compressibility.

8.2 Minimum Description Length

MDL favors models with shorter descriptions.

The ρ framework provides a concrete spectral mechanism for description length:

$$N(\varepsilon) \sim \frac{\log(1/\varepsilon)}{\log \rho}.$$

So larger ρ means shorter valid description.

8.3 Rate-Distortion

If $R(D)$ denotes rate-distortion, then the spectral point of view suggests

$$\rho \approx e^{1/R(D)}.$$

This means ρ is not only a smoothness parameter; it is an inverse measure of the rate needed to describe the object at a given distortion scale.

8.4 Shrinkage and Effective Dimension

Shrinkage methods already exploit the idea that small modes should be suppressed. The ρ view unifies this:

- shrinkage is local mode control,
- ρ is global envelope control,
- together they define structural regularization.

---

9. Current Evidence

9.1 Pattern Theory Evidence

The spectral pattern framework already suggests:

- high ρ corresponds to strong, stable patterns,
- low ρ corresponds to weak structure and overfitting danger,
- representation size grows as $\rho \to 1$.

This is evidence that ρ is structurally meaningful.

9.2 Implied Generator Evidence

In the implied-generator inverse problem, the exact generator matrix is partially non-identifiable. Yet holdout pricing can remain accurate when the recovered panel preserves ρ_{spec}.

This is strong evidence that:

- parameter closeness is not always the right metric,
- structural closeness can be more meaningful,
- ρ can function as a fitness statistic for inverse recovery.

We now also have a first explicit **ρ-aware solver term** inside the package-level implied-generator calibration runtime. In src/spectral_fenton/implied_generator.py, the generator fit can include the additional residual

$$\sqrt{\lambda_\rho}\,(\rho_{\text{fit}} - \rho_{\text{target}}),$$

where ρ_{target} is estimated from the recovered calibration coefficient panel and ρ_{fit} is computed from the semigroup-generated panel of the candidate generator.

This gives a direct numerical test of the thesis of this paper: what happens when ρ is moved from a reporting metric into an actual fitness term?

On the current market-like Heston-style split benchmark (examples/implied_generator_vol_surface.py), the baseline fit gives

$$\mathrm{RMSE}_{\mathrm{holdout}} = 0.0884, \qquad |\rho_{\mathrm{fit}} - \rho_{\mathrm{target}}| = 0.2115.$$

Adding a small ρ penalty with weight 10^{-4} changes this to

$$\mathrm{RMSE}_{\mathrm{holdout}} = 0.1230, \qquad |\rho_{\mathrm{fit}} - \rho_{\mathrm{target}}| = 0.0526.$$

So the ρ-aware solver does exactly what it is asked to do: it makes the recovered dynamics structurally closer in spectral-quality terms. But it does **not** automatically improve holdout pricing. In this benchmark, the price error actually gets worse as the ρ term is strengthened.

This is an important result. It supports the paper's main caution:

> ρ is a meaningful structural fitness term, but not a complete objective.

Equally importantly, the panel-level benchmark remains structurally stable across the sweep:

$$\rho_{\mathrm{spec,observed}} = 1.4304, \qquad \rho_{\mathrm{spec,recovered}} = 1.4416, \qquad |\Delta\rho_{\mathrm{panel}}| = 0.0112.$$

This means the global compressibility regime is already being recovered well at the panel level; the ρ penalty mostly acts on the internal semigroup fit. That is exactly the kind of phenomenon a future generalization theory must explain.

9.3 Knowledge Artifact Evidence

The Knowledge Artifact program already distinguishes:

- task fit,
- spectral entropy,
- effective dimension,
- smoothness of representation.

This is conceptually aligned with the claim that training quality should be evaluated structurally, not just pointwise.

---

10. Experimental Program

The claim that ρ is a useful training metric should be tested by four experiments.

10.1 Predictive Validity

Does better ρ matching correlate with lower holdout error at fixed train loss?

10.2 Noise Robustness

Under increasing observation noise, is ρ more stable than parameter RMSE?

10.3 Regularization Benefit

Compare:

1. task loss only,
2. task loss + ρ penalty,
3. task loss + standard complexity penalty,
4. pure ρ fitting.

The current implied-generator sweep suggests a subtler prediction than the original optimistic version:

- (2) improves structural fit to the target spectral envelope,
- (2) may or may not improve holdout task error,
- (4) fails or becomes unstable,
- the real question is whether there exists a regime where (2) gives a better structural-generalization tradeoff than (1).

10.4 Lens Robustness

Estimate ρ under different bases, kernels, or observation lenses. If ρ is truly structural, it should be much more stable than coordinate-level quantities.

10.5 Recoverable Complexity Study

For each dataset or inverse problem instance, estimate

$$K_{\text{eff}}(\tau) \quad \text{and} \quad \rho.$$

Then test whether:

1. larger estimated ρ predicts smaller effective dimension,
2. smaller effective dimension predicts lower holdout variance,
3. ρ explains generalization better than raw parameter count.

This directly tests the recoverable-complexity interpretation.

10.6 Theorem Roadmap

The clean theorem sequence would be:

1. **Tail theorem**: ρ-decay implies logarithmic effective dimension.
2. **Bias-variance theorem**: effective dimension controls estimation variance.
3. **Selection theorem**: ρ-aware regularization recovers near-optimal complexity.
4. **Robustness theorem**: under noise and lens changes, ρ is more stable than coordinate-level fit metrics.

If all four are established, then ρ becomes more than an interpretation. It becomes a genuine generalization variable.

11. The Right Claim

The right claim is **not**:

> ρ is a universal training loss.

Nor is it:

> fitting ρ eliminates overfitting by definition.

The right claim is:

> ρ is a general structural fitness metric. It measures recoverable complexity, provides a principled notion of justified complexity, and can reduce overfitting pressure when used as a regularizer, selector, or diagnostic.

This is already a substantial statement.

If proved rigorously, it would unify:

- generalization,
- model compression,
- inverse-problem identifiability,
- overfitting diagnostics,
- and rate-distortion geometry

under one spectral quantity.

---

12. Conclusion

Training usually asks: *how well did we fit the observed target?*

The deeper question is:

> *how much real structure did we recover?*

That is what ρ measures.

It is not a full replacement for task loss. It is not automatically overfitting-free. But it is a principled, interpretable, and potentially universal **structural metric** for training.

If this program is correct, then the future of training will not be organized around accuracy alone. It will be organized around a two-part discipline:

1. fit the task,
2. match the structure.

And ρ is the first serious candidate for the second quantity.

More precisely, ρ is the first serious candidate for a one-number measure of **recoverable complexity**. If that program succeeds, then overfitting will no longer be understood mainly as a pathology

of large parameter count. It will be understood as a mismatch between model complexity and the amount of structure reality actually allows us to recover.