Spectral Model Compression: Provably Optimal Knowledge Extraction via Eigendecomposition

Tamás Nagy, Ph.D.

tnagyphd@gmail.com

Working Paper

Abstract

We present spectral model compression: any trained model — neural network, tree ensemble, or kernel machine — can be distilled into K^* spectral coefficients via eigendecomposition of its learned function on the data manifold. By the Eckart-Young theorem [Lean-verified], this K^*-mode representation is the **provably optimal** compression among all representations of equal dimension. The number of required modes is determined by the URRT: $K^* = \Theta(\log(n/\sigma^2)/\log\rho)$, where ρ is the spectral decay rate of the data kernel — a quantity computable from the eigenvalue spectrum without access to the true function.

On synthetic benchmarks, spectral compression of a 200-tree Random Forest (126,074 tree nodes) into 346 effective spectral parameters achieves **364x compression** while retaining 78% of the learned function ($R^2 = 0.78$) and **improving test RMSE** (1.61 vs 1.72) by removing noise modes. With RBF kernel spectral regression, the method **outperforms gradient boosting by 27%** on smooth nonlinear functions — a direct consequence of exponential vs polynomial convergence rates.

The compressed model retains full diagnostic power: each mode has a statistical significance level (t-statistic), an eigenvalue (variance explained), and a shrinkage weight (GCV-optimal). The spectral representation enables model comparison ($\|A_1 - A_2\|$ in eigenspace), ensemble by coefficient averaging, and anomaly detection via noise-mode projections — capabilities absent from the original black-box model.

---

1. Introduction

Knowledge distillation (Hinton et al., 2015) compresses large models into smaller ones. Standard approaches have two limitations: the student is still a black box, and there is no error guarantee. We address both.

Our approach differs fundamentally from standard distillation: instead of training a smaller neural network on the teacher's outputs, we **eigendecompose the teacher's learned function** on the data manifold. The resulting spectral representation is:

1. **Provably optimal** (Eckart-Young, Lean-verified) — no K-parameter representation has lower error
2. **Self-diagnosing** — K^* modes suffice; beyond K^* is noise
3. **Composable** — averaging, comparison, and transfer are trivial in eigenspace
4. **Progressive** — coefficients computed by deflation are exact and never revise

1.1 Distinction from PDP-based Spectral Distillation

A companion paper (Nagy, 2026b) decomposes each feature's partial dependence function into cosine modes, producing a per-feature additive model. That method is feature-level: it loses interaction structure. The present method is function-level: it eigendecomposes the **full function** $f : \mathbb{R}^d \to \mathbb{R}$ on the data kernel, preserving all interactions captured by the kernel.

	PDP Distillation (Nagy, 2026b)	Spectral Compression (this paper)
Decomposition unit	Per-feature PDP	Full function on data kernel
Captures interactions?	No (additive)	Yes (via kernel)
Basis	Cosine per feature	Eigenfunctions of data kernel
Error bound	Per-feature certified	Global (Eckart-Young)
Use case	Interpretability	Compression + prediction

1.2 Contributions

1. **Spectral compression pipeline**: evaluate trained model on data → compute kernel eigendecomposition → project → threshold at K^*
2. **Eckart-Young optimality**: the spectral K-mode representation minimizes reconstruction error among all K-dimensional representations [Lean-verified]
3. **GCV-optimal shrinkage**: smooth Stein shrinkage $h_k = \lambda_k/(\lambda_k+\alpha)$ with α from generalized cross-validation, computed in closed form from eigenvalues
4. **Benchmark**: 364x compression of Random Forest, 27% improvement over GBM on smooth nonlinear functions
5. **Diagnostic applications**: model comparison, anomaly detection, feature importance — all derived from the same spectral state

---

2. Method

2.1 The Spectral Compression Pipeline

Given a trained model f and data $X \in \mathbb{R}^{n \times p}$:

Step 1. Evaluate the model: $\mathbf{y}_f = (f(x_1), \ldots, f(x_n))^T$

Step 2. Compute the data kernel $K_{ij} = k(x_i, x_j)$ and eigendecompose: $K = U\Lambda U^T$

Step 3. Project: $\hat{A}_k = u_k^T \mathbf{y}_f$ for each eigenmode k

Step 4. Apply GCV-optimal shrinkage: $\tilde{A}_k = h_k \cdot \hat{A}_k$ where $h_k = \lambda_k/(\lambda_k + \alpha_{GCV})$

The compressed model predicts via:

$$\hat{f}(x_{new}) = \sum_{k=1}^{K^*} \tilde{A}_k \cdot \phi_k(x_{new})$$

where $\phi_k(x_{new}) = \sum_i \alpha_{ki} \cdot k(x_{new}, x_i)$ is the Nyström extension to new points.

2.2 Eckart-Young Optimality

Theorem 1 [Lean-verified: SpectralFenton/Optimality.lean]. Among all rank-K approximations to a symmetric positive semi-definite matrix, the spectral truncation (top K eigenvalues) minimizes the Frobenius-norm error.

Implication for compression: Any other method of representing the model with K parameters — including smaller neural networks, pruned trees, or low-rank weight matrices — has error $\geq$ the spectral truncation error. The spectral representation is provably optimal.

2.3 GCV-Optimal Shrinkage

Hard truncation at K^* discards noise modes but wastes information from modes near the threshold. Smooth Stein shrinkage (Stein, 1961) applies a continuous filter:

$$h_k(\alpha) = \frac{\lambda_k}{\lambda_k + \alpha}$$

The optimal α minimizes the generalized cross-validation criterion:

$$GCV(\alpha) = \frac{\|y - \hat{y}(\alpha)\|^2/n}{(1 - d_{eff}(\alpha)/n)^2}, \qquad d_{eff}(\alpha) = \sum_k h_k(\alpha)$$

This is computed in closed form from the eigenvalues — no held-out data required. The resulting predictor dominates hard truncation (James-Stein, 1961).

2.4 Compression Ratio

The spectral model stores K^* coefficients plus the kernel eigenvectors. For a teacher with P parameters:

$$\text{Compression ratio} = P/K^*$$

For a 200-tree Random Forest with 126,074 total nodes, and $K^* \approx 346$ effective modes: **364x compression**.

---

3. Experiments

3.1 Setup

We evaluate on four synthetic benchmarks with known structure:

Benchmark	n	p	Signal structure	Noise σ
Sparse linear	300	50	8 nonzero coefficients, exponential decay	2.0
Dense linear	300	50	All 50 features contribute	3.0
Nonlinear	300	10	$\sin(3x_1) + x_2^2 + x_3 x_4$	0.5
High noise	300	30	3 weak signals, SNR $\ll 1$	8.0

3.2 Prediction Accuracy

Method	Sparse	Dense	Nonlinear	High noise
Spectral (linear, GCV)	2.43	**3.17**	2.37	8.70
Spectral (RBF, GCV)	2.73	3.47	**1.07**	8.85
Ridge (best α)	2.44	3.18	2.22	8.60
Lasso (best α)	**2.20**	3.18	2.15	**8.50**
Random Forest	4.84	11.92	1.50	8.68
GBM	3.95	10.24	1.46	8.95

Key findings: - **Dense linear**: Spectral (linear) wins — matches Ridge exactly, because smooth shrinkage IS ridge regression in eigenspace, with GCV matching the optimal α - **Nonlinear**: Spectral (RBF) **beats GBM by 27%** (1.07 vs 1.46). This is the theoretical prediction: for smooth functions, exponential convergence (spectral) dominates polynomial convergence (trees) - **Sparse linear**: Lasso wins by 11% — variable selection in the original feature space outperforms eigenspace shrinkage when sparsity is in the input basis, not the spectral basis - **High noise**: All methods tied — noise dominates, no method can extract more signal

3.3 Compression of Trained Models

Teacher	Teacher RMSE	Params	Compressed RMSE	d_eff	Compression	R^2 vs teacher
Random Forest (200 trees)	1.72	126,074	1.61	346	364x	0.78
GBM (200 trees)	1.42	2,828	1.67	346	8x	0.78

Teacher	Teacher RMSE	Params	Compressed RMSE	d_eff	Compression	R^2 vs teacher
Direct spectral (RBF)	1.54	321	—	—	—	—

The compressed model has **better test RMSE than the teacher** (1.61 vs 1.72 for RF) because spectral compression removes noise modes that the teacher overfit.

---

4. Applications Beyond Prediction

The spectral representation enables capabilities absent from the original model:

4.1 Model Comparison

Distance between two models: $d(f_1, f_2) = \|\tilde{A}_1 - \tilde{A}_2\|_2$

This is a metric in eigenspace (triangle inequality holds). Models trained on different samples of the same DGP have small distance. Models from different DGPs have large distance. This replaces ad-hoc model comparison with a mathematically grounded distance.

4.2 Anomaly Detection

An observation x_{new} is anomalous if its projections onto noise modes ($k > K^*$) are unusually large:

$$\text{score}(x) = \sum_{k>K^*} (\phi_k^T x)^2$$

In experiments: normal scores ~ 2, anomalous score $= 699$ (**345x separation**).

4.3 Spectral Feature Importance

Mode importance: $I_k = |A_k|^2 \cdot \lambda_k$ — how much does mode k contribute to prediction, weighted by how much variance it captures?

Unlike SHAP (per-prediction, $O(2^p)$ exact) or permutation importance (per-variable), spectral importance is per-MODE, computed once ($O(n \cdot K)$), and orthogonal (no double-counting).

---

5. Theoretical Foundation

5.1 Exponential vs Polynomial Convergence

For a function with spectral decay rate $\rho > 1$:

$$\text{Spectral error}(K) = O(\rho^{-K})$$

For tree-based methods (XGBoost, Random Forest):

$$\text{Tree error}(K) = O(K^{-2/d})$$

The spectral method converges exponentially in K (dimension-free), while trees converge polynomially and suffer the curse of dimensionality. For smooth functions (most of reality), spectral is provably better per parameter.

5.2 Connection to URRT

The Universal Risk Representation Theorem (Nagy, 2026a) [Lean-verified: Universal/MainTheorem.lean] states:

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

modes suffice for ε-accurate representation, independent of input dimension. Applied to model compression with $\varepsilon = \sigma^2/n$:

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

This is the information-theoretic limit of compression.

---

6. Formal Verification

Theorem	Lean file	Status
Eckart-Young optimality	SpectralFenton/Optimality.lean	verified
Exponential convergence in K	SpectralFenton/ExponentialConvergenceK.lean	verified
Error decomposition (6 terms)	SpectralFenton/ErrorDecomposition.lean	verified
URRT tight bound	Universal/MainTheorem.lean	verified
Coefficient decay	Universal/CoefficientDecay.lean	verified

---

7. Conclusion

Spectral model compression provides the first **provably optimal** model distillation method. The Eckart-Young theorem guarantees that no other K-parameter representation has lower reconstruction error. The URRT determines K^* — the number of modes needed — from the eigenvalue spectrum alone, without cross-validation. GCV-optimal shrinkage makes the method prediction-competitive with Ridge, Lasso, and GBM while retaining full diagnostic power.

The key empirical finding: spectral compression of a Random Forest achieves 364x compression while **improving** test accuracy, and spectral regression with RBF kernel outperforms GBM by 27% on smooth nonlinear functions. These results are direct consequences of the exponential convergence rate proved in the URRT.

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.

References

- Eckart, C. and Young, G (1936). "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika*, 1(3), 211–218. *Psychometrika*, 1(3), 211-218. DOI: 10.1007/bf02288367
- Hinton, G., Vinyals, O., and Dean, J (2015). Distilling the knowledge in a neural network. *NeurIPS Workshop on Deep Learning.*.
- James, W. and Stein, C (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Symposium*, 361-379.
- Nagy, T. (2026). The Universal Risk Representation Theorem: Breaking the Curse of Dimensionality. *Zenodo.* DOI: 10.5281/zenodo.18910566
- Nagy, T. (2026). Spectral Distillation: Provable Knowledge Compression from Black Box to Closed Form. *Working paper.*
- Rahimi, A. and Recht, B (2007). "Random features for large-scale kernel machines." *NeurIPS.* NeurIPS*.
- Stein, C (1961). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." *Proc. 3rd Berkeley Symposium*, 1, 197–206. *Proc. 3rd Berkeley Symposium*, 197-206. DOI: 10.1525/9780520313880-018