The Underlying Universal Statistical Structure of Natural Datasets

Noam Levi & Yaron Oz

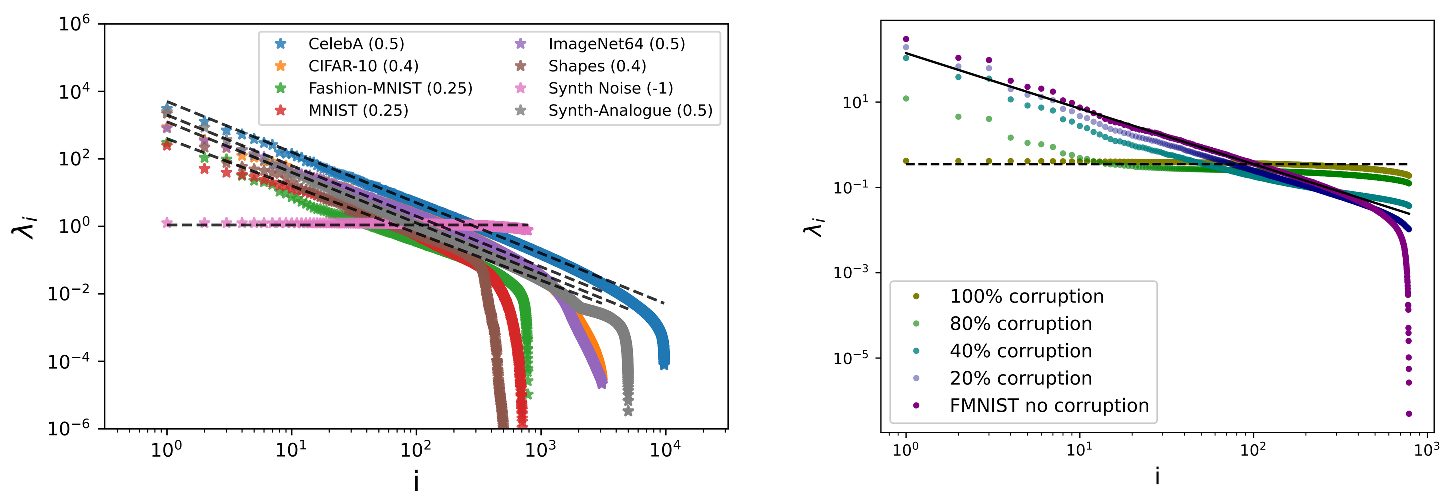
Check out our paper:

noam.levi@epfl.ch

Overview

TL; DR: We analogize the covariance matrices $\Sigma_M = XX^T/M$ of natural datasets to physics Hamiltonians. We apply tools from Random Matrix Theory to show that the analogous system belongs in the Quantum Chaotic universality class.

Power law spectra are ubiquitous and related to neural scaling laws (NSL)



Power laws are ubiquitous in physical systems near criticality:

At a critical point: $\langle\phi(x)\phi(y)\rangle \propto \frac{1}{r^{d-2+\eta}} = \frac{1}{r^{2\Delta+\eta}}$

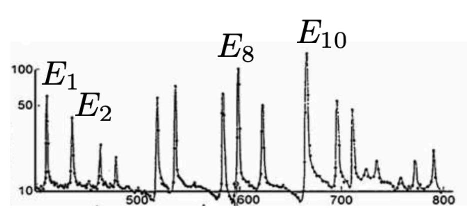
Translation invariant correlations produce the same spectrum

$$\Sigma^{\text{Toe}}_{ij} = S, \qquad T = \mathbb{1}_{ij} + c|i-j|^{\alpha} = U^{\dagger}SV, \qquad \alpha, c \in \mathbb{R}.$$

Features are correlated - correlation exponent determines generalization under NSL

What about local statistics? Enter Universality

Resonances in heavy nuclei: $E_1 < E_2 < E_3 < \ldots < E_N$



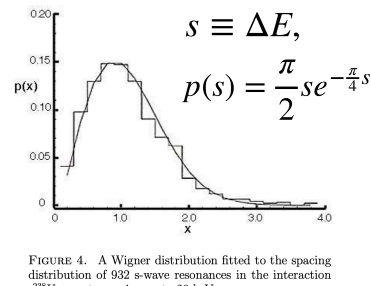
[Firk, Lynn, & Moxon, '60]

Level Spacing: $s_i \equiv E_{i+1} - E_i, \quad i = 1,2,\ldots$

What is the distribution of $\{s_i\}_{i<N}$?

The energies are Eigenvalues of a quantum Hamiltonian $H\psi_n = E_n\psi_n$

Wigner's surmise (1956): The interactions are so complex we might as well regard each entry as some randomly chosen number → the statistical symmetries and scaling dictates everything



Insight: the details of H are not important (universality). Replace it by a random H

Two main classes: Integrable or Chaotic. Which one does natural Σ_M belong to? Can be read from the level spacing distribution!

Natural data spectra converge to the Gaussian Orthogonal Ensemble (GOE)

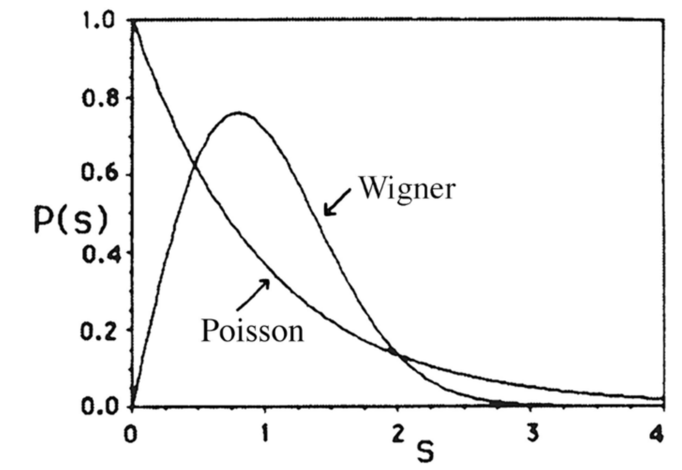
Metrics

Level Spacing: $s_i \equiv \lambda_{i+1} - \lambda_i, \quad i = 1,2,\ldots,d$

Ratio of Level Spacing: $r_i \equiv \frac{\min(s_i, s_{i+1})}{\max(s_i, s_{i+1})},$

Spectral form factor:

$$K(\tau) = \simeq \frac{1}{Z}\langle \Big|\sum_i \rho(\lambda_i) e^{-2\pi i \lambda_i \tau}\Big|^2\rangle, \quad Z = \sum_i |\rho(\lambda_i)|^2$$



Universality Classes

Chaotic/Ergodic/GOE

GOE:

$$P_{GOE}(H) = \frac{1}{Z_{GOE}(d)} e^{-\frac{1}{2}\text{Tr}H^2}$$

$$H_{ij} \sim \mathcal{N}(0,1)$$

RMT Predictions for spacings:

$$P_{\text{GOE}}(s) = \frac{\pi}{2} s e^{-\pi s^2/4}$$

$$P_{\text{GOE}}(r) = \frac{27}{4}\frac{r + r^2}{(1 + r + r^2)^{5/2}}\Theta(1-r)$$

$$\langle r\rangle_{\text{GOE}} = 4 - 2\sqrt{3}$$

Integrable

Poisson Ensemble:

$$P_{poiss}(H) = \frac{1}{Z_{Poiss}(d)} e^{-\frac{1}{2}\text{Tr}H^2}$$

$$H_{diag} \sim \mathcal{N}(0,1) \text{ else } 0$$

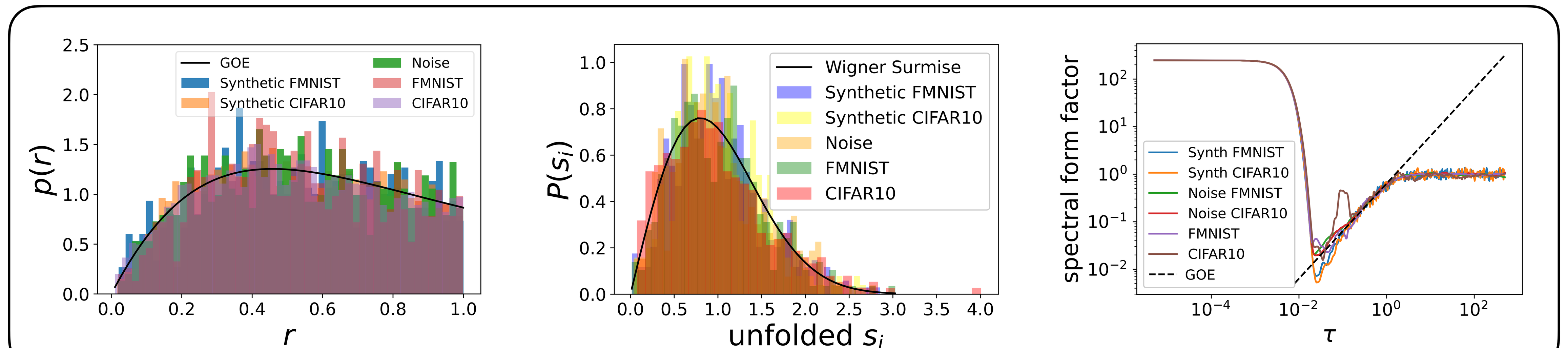
RMT Predictions for spacings:

$$P_{\text{poiss}}(s) = e^{-s}$$

$$P_{\text{poiss}}(r) = \frac{2}{(1+r)^2}\Theta(1-r)$$

$$\langle r\rangle_{\text{poiss}} = 0.386$$

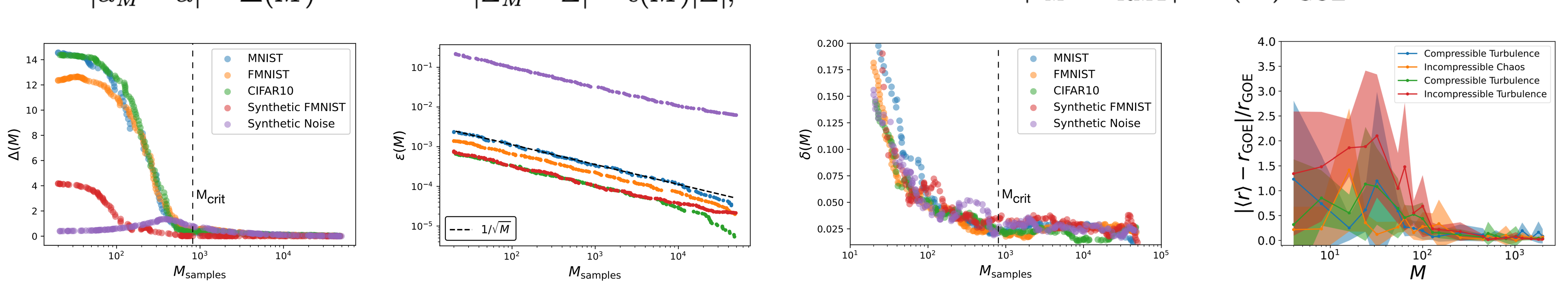
****Black curves indicate theoretical predictions.***



Real world covariance matrices converge to the GOE universality class independently of the global spectrum- same as sampling from a Gaussian distribution!

How many samples do we need to reach the RMT regime? (Ergodicity)

$|\alpha_M - \alpha| = \Delta(M)$ $\qquad |\Sigma_M - \Sigma| = \epsilon(M)|\Sigma|,$ $\qquad |r_M - r_{\text{RMT}}| = \delta(M) r_{\text{GOE}}$



It requires roughly $M \sim d$ to reach the RMT as well as the scaling regime

Spectral density and Stieljes Transform

Stieljes Transform relates the spectral density and the ensemble average

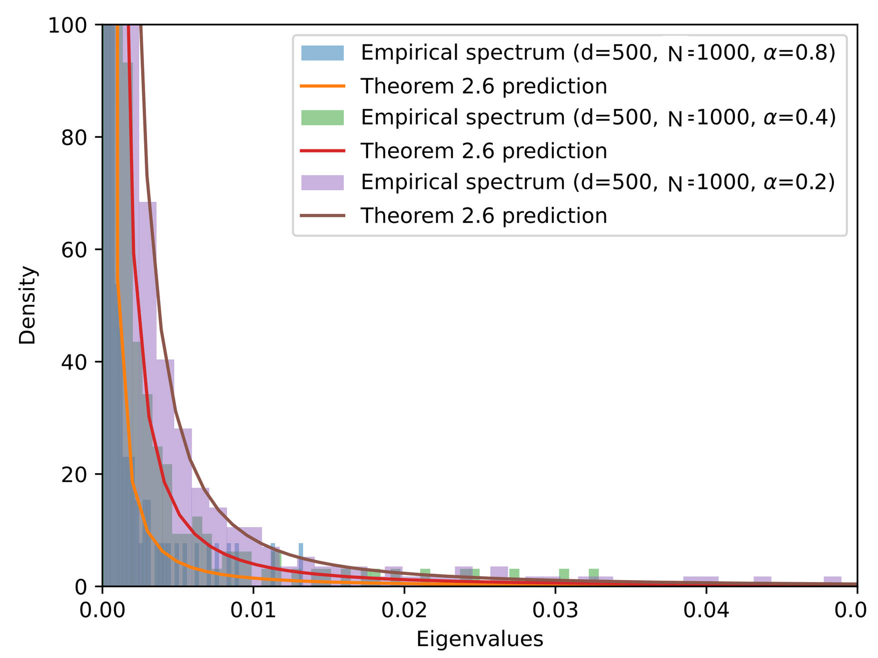
$$G(z) = \int \frac{\rho_{\Sigma_N}(t)}{z-t}dt = -\frac{1}{N}\mathbb{E}\left[\text{tr}(\Sigma_N - zI)^{-1}\right], \qquad \rho_{\Sigma_N}(\lambda) = -\frac{1}{\pi}\lim_{\epsilon\to 0^+}\Im G(\lambda + i\epsilon),$$

Theorem [Silverstein, Bai 95'] for Wishart matrices:

$$G(z) = \frac{1}{\gamma}\tilde{G}(z) + \frac{1-\gamma}{\gamma z}, \qquad \tilde{G}(z) = \left(-z + \frac{1}{N}\text{tr}\left[\Sigma(I + \tilde{G}(z)\Sigma)^{-1}\right]\right)^{-1},$$

Solution is given by:

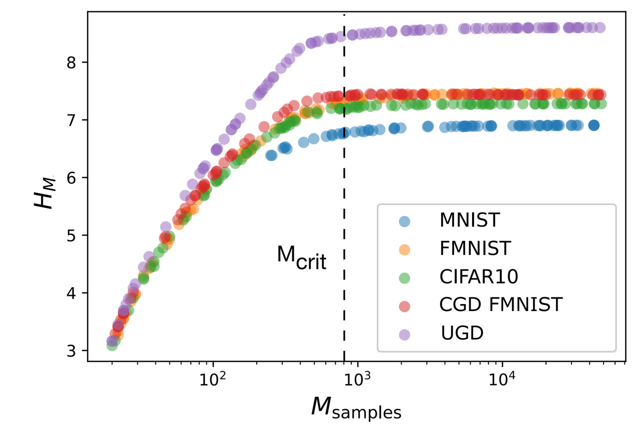
$$\tilde{G}(z) = \left[-z + \gamma c\left(\frac{\pi\csc\left(\frac{\pi}{1+\alpha}\right)}{(1+\alpha)d}\right)\left(\frac{1}{c\tilde{G}(z)}\right)^{\frac{\alpha}{1+\alpha}}\right]^{-1}$$

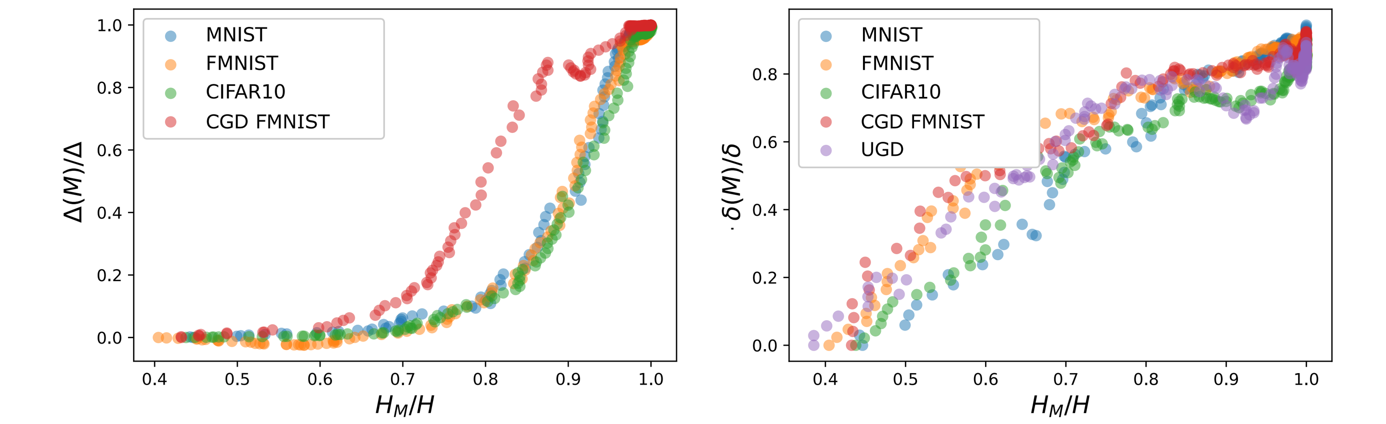


Relation to Shannon Entropy

Shannon Entropy: $H = -\sum_{i=1}^{n} p_i \log(p_i)$

Where the probabilities: $p_i = \lambda_i / \sum_{i=1}^{n_{\text{bulk}}} \lambda_i$





The saturation of entropy converges with the onset of the RMT regime!