UNIVERSITÀ
DI PAVIA

Dipartimento di Fisica

Corso di laurea magistrale in Scienze Fisiche

Influence Functions and Convergence of Generative Models

Relazione per la laurea di:
Cecilia Bombari

Relatore:
Prof. Sebastian Goldt

Correlatore UniPV:
Prof. Giacomo Livan

Anno accademico 2025 – 2026

Abstract (English)

Diffusion models have shown strong empirical performance in generative tasks, such as image generation, but it is still not fully understood how they generalize to unseen data, rather than simply memorizing the training set. This is an important question nowadays, with direct consequences for privacy and copyright. This thesis aims to study the generalization properties of diffusion models using tools from statistical physics and statistical learning theory, within a theoretically tractable Gaussian framework. We define generalization through a bound on the Kullback-Leibler (KL) divergence, which measures the discrepancy between the distribution learned by the empirical model and the true distribution from which the data are drawn. To control this divergence, we introduce a Leave-One-Out (LOO) influence function derived from the regularized sample covariance. We show that a generalized version of the KL divergence, which we call the *excess* KL divergence, is rigorously upper-bounded by the LOO influence density. Furthermore, by evaluating the thermodynamic limit via the Marchenko-Pastur spectral measure, we derive matching upper and lower bounds for the excess KL divergence, establishing that both quantities are of the same order across the whole range of sample complexities. We then connect these bounds to the convergence overlap factor q, first introduced in the Kadkhodaie and Guth [1] framework and later formalized by Goldt and Maillard [2]. This extends our analysis from a leave-one-out perturbation to a macroscopic comparison between models trained on disjoint data subsets, providing an exact analytical description of the memorization-to-generalization transition.

My GitHub Repository: `https://github.com/Bombaric/Thesis`

Abstract (Italiano)

Nonostante gli ottimi risultati empirici ottenuti dai modelli di diffusione nei task generativi (come, ad esempio, la generazione di immagini), non è ancora del tutto chiaro come questi riescano a generalizzare su dati mai visti, piuttosto che limitarsi a memorizzare il training set. Si tratta di una questione cruciale, con dirette conseguenze, ad esempio, in materia di privacy e copyright. Questo lavoro affronta le proprietà di generalizzazione dei modelli di diffusione attraverso tecniche di fisica statistica e di statistical learning theory, all'interno di un framework Gaussiano trattabile analiticamente. Definiamo la generalizzazione attraverso un limite sulla divergenza di Kullback-Leibler (KL), che quantifica la discrepanza tra la distribuzione appresa dal modello empirico e la vera distribuzione generatrice dei dati. Per controllare tale divergenza, introduciamo una funzione di influenza Leave-One-Out (LOO) derivata dalla matrice di covarianza campionaria regolarizzata. Mostriamo che una versione generalizzata della divergenza KL, denominata *excess Kullback-Leibler divergence*, è rigorosamente limitata superiormente dalla densità di influenza LOO. Inoltre, valutando il limite termodinamico per n e d, tramite la misura spettrale di Marchenko-Pastur, ricaviamo dei limiti (upper e lower bounds) per la KL, dimostrando che divergenza e influenza sono dello stesso ordine di grandezza per qualsiasi complessità campionaria. Infine, colleghiamo questi risultati al fattore di overlap q, introdotto nel framework di Kadkhodaie e Guth [1] e formalizzato successivamente da Goldt e Maillard [2]. Questo passaggio estende l'analisi da un contesto di perturbazione leave-one-out al confronto macroscopico tra modelli addestrati su sottoinsiemi di dati disgiunti, fornendo una descrizione analitica esatta della transizione da memorizzazione a generalizzazione.

La mia Repository GitHub: `https://github.com/Bombaric/Thesis`

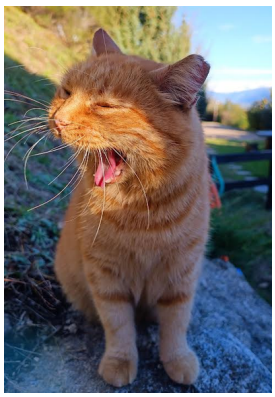
Contents

Chapter 1

Introduction

It's easy to destroy, but hard to create.
– Pearl S. Buck

Imagine a vast collection of cat photos. While the cats vary in colour, breed, and posture, they share a common structure: the texture of their fur, the ears, the almond shape of their eyes. In the language of statistics, these common features define an underlying probability distribution $p_{\text{cat}}(\mathbf{x})$ [3,4]. The natural goal is to sample new images from this distribution—cats that never appeared in the original collection, yet are visually indistinguishable from the real ones.



A yawning cat from my photo gallery.

Generative modelling provides the formal framework for this task. Its goal is to construct a sampler for an unknown target distribution $p^*(\mathbf{x})$ given a finite set of i.i.d. examples $\mathbf{x}_i$, that live in $\mathbb{R}^d$ [4], where d corresponds to the number of pixels of the photo we are testing. Diffusion models approach this by decomposing the problem of sampling from p^* into a sequence of incremental denoising steps [4]. By learning to reverse a process that progressively adds Gaussian noise to data, these models learn to transform an isotropic Gaussian back into the structured target distribution [3, 4].

1.1 The Forward Noising Process

The **forward process** progressively destroys the structure of a data point $\mathbf{x}_0 \sim p^*$ by injecting noise [3, 4]. In a discrete-time Gaussian diffusion, we build a sequence

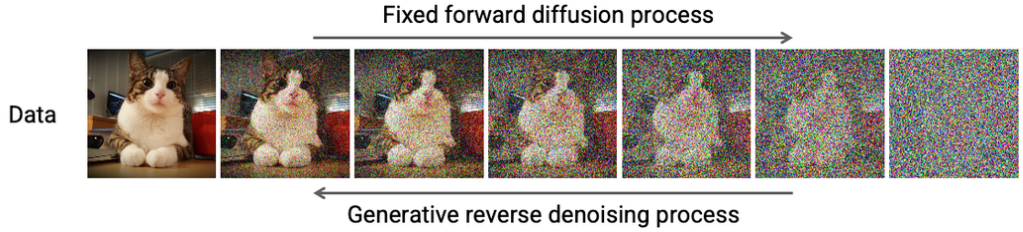


Figure 1.1: The forward noising process. Starting from a structured data point $\mathbf{x}_0$, further Gaussian perturbations gradually erase all structure until the distribution at $t = 1$ is approximately $\mathcal{N}(0, \sigma_q^2 \mathbb{I}_d)$. Figure from [5].

$\{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^d$ by adding independent Gaussian increments:

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0,\, \sigma_q^2\, \Delta t\, \mathbb{I}_d), \tag{1.1}$$

where $\Delta t = 1/T$ is the step size and σ_q^2 is the terminal variance [4]. Because a sum of independent Gaussian increments is again Gaussian, the marginal at any time t admits the closed form

$$p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t;\, \mathbf{x}_0,\, \sigma_t^2 \mathbb{I}_d), \qquad \sigma_t = \sigma_q \sqrt{t}, \tag{1.2}$$

or equivalently $\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon$, with $\epsilon \sim \mathcal{N}(0, \mathbb{I}_d)$. This single noise variable ϵ is what the denoiser ϵ_θ will later learn to predict (Section 1.2). As $t \to 1$ the accumulated noise renders p_1 nearly indistinguishable from an isotropic Gaussian (Figure 1.1).

1.2 Learning the Reverse Process

When the noise increment σ is small, the reverse conditional $p(\mathbf{x}_{t-\Delta t} \mid \mathbf{x}_t)$ is nearly Gaussian [4] (Figure 1.2), which makes it possible to learn by regression. The dynamics of this reverse process are governed by the **score function** $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, the gradient of the log-density pointing toward regions of higher probability [3].

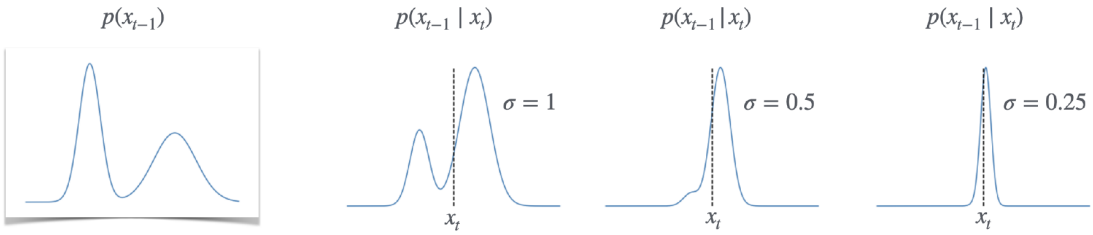


Figure 1.2: Reverse conditional distributions $p(x_{t-1} \mid x_t)$ for a fixed x_t and varying noise levels σ [4]. For small σ each conditional is well approximated by a Gaussian, making the reverse process tractable.

The conditional mean of the reverse step,

$$\mu_{t-\Delta t}(z) \;:=\; \mathbb{E}[\mathbf{x}_{t-\Delta t} \mid \mathbf{x}_t = z], \tag{1.3}$$

is the unique minimiser of the mean squared error, so

$$\mu_{t-\Delta t} = \operatorname*{argmin}_f \; \mathbb{E}_{\mathbf{x}_{t-\Delta t},\, \eta_t} \big\| f(\mathbf{x}_{t-\Delta t} + \eta_t) - \mathbf{x}_{t-\Delta t} \big\|_2^2. \tag{1.4}$$

This is a standard denoising problem: recover the clean signal $\mathbf{x}_{t-\Delta t}$ from its noisy version $\mathbf{x}_t = \mathbf{x}_{t-\Delta t} + \eta_t$ [4]. A neural network ϵ_θ, the **denoiser**, is trained to approximate the noise ϵ using a mean-squared-error objective [3, 4]:

$$\theta^* = \arg\min_\theta \; \mathbb{E}_{\mathbf{x}_0,\, \epsilon,\, t}\big[\, \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \big]. \tag{1.5}$$

Tweedie's formula then connects the optimal denoiser to the score [1, 4]:

$$\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \mathbf{x}_t + \sigma_t^2 \, \nabla \log p_t(\mathbf{x}_t). \tag{1.6}$$

Thus, looking at how we linked $\mathbf{x_0}$ to $\mathbf{x_t}$, we can state that:

$$\epsilon_\theta^* := \mathbb{E}[\epsilon \mid \mathbf{x}_t] = -\sigma_t \nabla \log p_t(\mathbf{x}_t). \tag{1.7}$$

Learning a high-dimensional generative model thus reduces to a supervised regression task [3, 4].

1.2.1 Stochastic Sampling: DDPM

The **Denoising Diffusion Probabilistic Model (DDPM)** [3, 4] is a popular sampler in diffusion models. It couples two Markov chains: a forward chain that corrupts data with noise, and a reverse chain that learns to undo it step by step. Starting from $\mathbf{x}_1 \sim \mathcal{N}(0, \sigma_q^2 \mathbf{I})$, the reverse chain applies the learned denoising step while re-injecting a small amount of noise at each stage,

$$\mathbf{x}_{t-\Delta t} = \mu_\theta(\mathbf{x}_t, t) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0,\, \sigma_q^2 \, \Delta t \, \mathbf{I}). \tag{1.8}$$

The stochastic injection prevents the chain from collapsing to a single mode and allows the model to explore the full target distribution. The training objective follows from a variational lower bound on the log-likelihood but reduces in practice to the same noise-prediction MSE loss [3, 4].[1]

[1] Computing $\log p(\mathbf{x}_0)$ directly is intractable, since it requires integrating over all possible noisy trajectories $\mathbf{x}_1, \ldots, \mathbf{x}_T$. The standard remedy is to maximise instead a tractable lower bound, the **ELBO** (Evidence Lower BOund), obtained by introducing the forward process $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$ as a variational distribution.

1.3 Generalization in Diffusion Models

The question of whether a diffusion model genuinely learns the data distribution, or merely memorizes the training set, is harder to pose precisely than it first appears. In supervised learning, **memorization** and **generalization** are clearly distinguished by train and test error. For generative models, this distinction breaks down: achieving small denoising error on held-out images does not guarantee that the model generates diverse, high-quality samples from the true distribution [2].

1.3.1 Why is generalization non-trivial here?

This difficulty reflects a non-alignment with what is, by now, a comparatively well-understood chapter of supervised-learning theory [2]. In supervised learning, it was long observed that deep networks continue to generalize well even once they have been trained to (near) zero training loss [6]—a finding that initially seemed to contradict classical statistical learning theory, which predicts that a model achieving zero training error has overfit the training data and should generalize poorly. This apparent paradox has since been largely looked into by studies on kernel methods and small neural networks exhibiting so-called **benign overfitting**.

This does not carry over cleanly to generative models. Here, achieving (near) zero training loss typically corresponds to **explicit memorization**: the model's samples become close reproductions of individual training examples, or small perturbations of them. It is thus not yet established whether overparameterized generative models enjoy an analogous overparametrized good behaviour—and, it is not even obvious how one should define "generalization" for a generative model in the first place, since there is no natural analogue of a labelled test point against which to measure error.

1.3.2 Some conjectures on generalization in literature

Before turning to a specific formal framework, it is worth surveying how the field has tried to pin down *why* diffusion models generalize at all. Marion and Wu organize existing explanations for why practical diffusion models nevertheless avoid memorization into three broad, not-mutually-exclusive families [7].

Capacity limitations. The earliest and most intuitive account attributes the transition to generalization to the network simply lacking the capacity to memorize a large training set relative to its number of parameters. Gu et al. [8] study memorization empirically along these lines, relating the onset of memorization to the ratio between dataset size and model capacity. This family of explanations is very simple, but, as discussed below, it is also the one most directly challenged by later work.

Implicit regularization from optimization. A second family of explanations sees the generalization in the training dynamics themselves. Bonnaire, Urfin, Biroli and

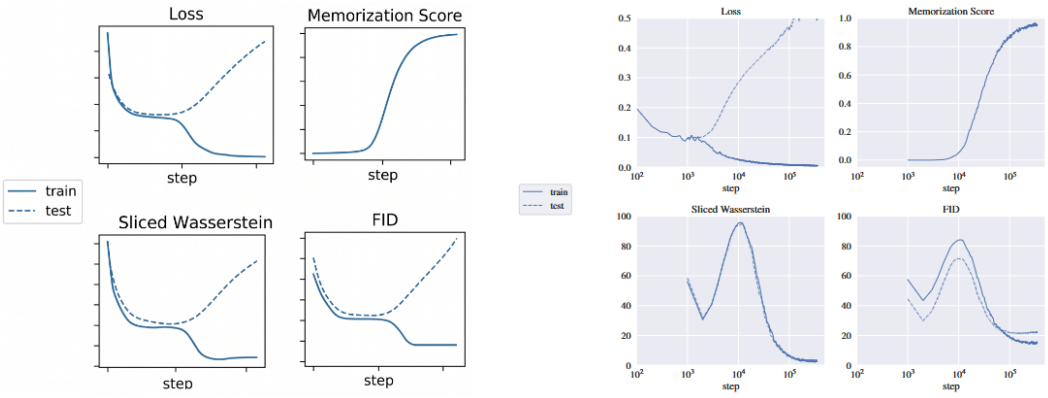


Figure 1.3: Predicted (left) versus actual (right) evolution of training metrics. In particular, the double descent of the FID metric is unexpected. [7]

Mézard [9] and Favero, Sclocchi and Wyart [10] show that even heavily overparameterized diffusion models—with more than enough capacity to memorize—can still generalize well if training is stopped early enough, and that the time needed to transition into the memorization regime grows with dataset size.

Wu, Marion, Biau and Boyer [11] show that the choice of learning rate itself acts as an implicit regularizer in denoising score matching, with large learning rates actively preventing memorization, which states something new with respect to the previous dimensional conjecture.

Architectural inductive biases. A third family, to which KGSM belongs and which we discuss in detail in Section 1.4, attributes generalization to structural properties of the network architecture itself[2].

An open methodological question. Marion and Wu [7] also raise a more skeptical point. In practice, generalization in diffusion models is usually measured with the **Fréchet Inception Distance (FID)**, a score that compares the statistics of generated images to the statistics of a reference image set in a fixed feature space [7] (see Figure 1.3). The common proxy for generalization is the *FID gap*: the difference between FID computed against the training set and FID computed against a held-out test set. A small gap is usually read as evidence that the model has generalized.

Marion and Wu argue that this proxy is not reliable: a small FID gap does not guarantee that the model is close to the true data distribution, and networks that clearly have enough capacity to memorize their training set often simply do not, for reasons current theory still cannot fully explain. This is a real gap in the field: most existing accounts explain *why* memorization is avoided in specific settings, but none of

[2]for example, as in the KGSM experiment, the result bias toward smooth, geometry-adaptive functions.

them gives a direct, computable measure of how close the learned distribution actually is to the true one. Marion and Wu's proposal is to shift the question itself, from why diffusion models fail to memorize, to what exactly they learn before memorization sets in—which in practice means looking for *operational*, *directly computable quantities that track generalization*, instead of proxies like the FID gap.

1.4 The KGSM Experiment

The transition from memorization to generalization was established empirically by Kadkhodaie, Guth, Simoncelli, and Mallat (KGSM) [1]. Their experiment is direct: train two diffusion models – using UNets convolutional neural networks – independently on **disjoint subsets** of the same dataset, then generate samples from both using the same noise seed, and measure the **cosine similarity**

$$S_C(\mathbf{A}, \mathbf{B}) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1.9}$$

between the output images and the input and output of the same model.

For small sample size n, the two models produce entirely different outputs—each reproducing images close to its own training set (*memorization regime*). Past a dataset-dependent threshold, the outputs converge: the cosine similarity between the two models' samples approaches the saturation level 1, while the similarity of each sample to its nearest training image decreases (Figure 1.5). At $n = 10^5$ face images the two models generate nearly identical samples, which no longer resemble any individual training image [1]. As we will see in Section 1.4.1, this empirical convergence is exactly what Maillard and Goldt later formalized as the order parameter q.

1.4.1 Generalization with order parameters

A recent theoretical analysis by Maillard and Goldt [2] puts the empirical transition observed by KGSM on a rigorous footing, in a tractable setting: **linear generative models trained on data drawn from a Gaussian distribution with a low-rank covariance structure.**, that is, $N(0, \mathbb{I}_d + \beta \mathbf{u}\mathbf{u}^\mathbf{T})$ with $\mathbb{I}_d, \mathbf{u} \in \mathbb{R}^d$.

Their analysis shows that generalisation decomposes into three *distinct* phenomena, each associated with its own order parameter and its own statistical distance between the true distribution $P^\star$ and the learned one $\hat{P}$.
At very small sample sizes, $n = \Theta(1)$, the model memorizes individual training examples. The memorization overlap

$$m := \mathbb{E}_z \max_\mu \frac{|\langle \mathbf{x}_\mu, \tilde{\mathbf{x}}(\mathbf{z}) \rangle|}{(\|\mathbf{x}_\mu\|\|\tilde{\mathbf{x}}(\mathbf{z})\|)}$$

—the maximum normalized inner product between a generated sample and any training point—decays as $m \simeq \sqrt{2 \log n/(n(1+\sigma^2))}$, where σ^2 is the regularization coefficient. m vanishes well before the second phenomenon sets in (see Figure 1.6).

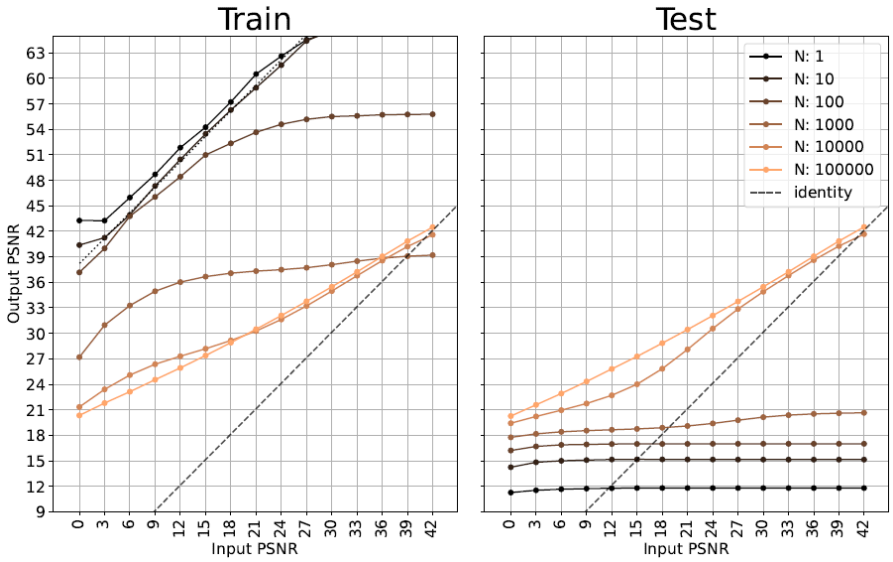


Figure 1.4: Train and test PSNR (signal to noise ratio) curves for a UNet denoiser trained on face images, for training sets of size N. As N increases, test performance improves while training performance degrades; at $N = 10^5$ the two curves are essentially identical, indicating the model has left the memorization regime. From [1].

At linear sample complexity, $n \asymp d$, independently trained generative models converge to the same output when started from the same latent variable. This convergence is captured by the overlap

$$q := \mathbb{E}_z \left[\frac{|\langle \tilde{x}^{(0)}(z), \tilde{x}^{(1)}(z) \rangle|}{(\|\tilde{x}^{(0)}\| \|\tilde{x}^{(1)}\|)} \right]$$

between samples produced by two independently trained models from the same noise seed z—exactly the population-level expectation of the cosine similarity S_C of (1.9) that KGSM measured empirically in Section 1.4.
Maillard and Goldt show that q rises continuously from 0 to 1 in the regime $n \asymp d$, and that it controls the rescaled *excess*[3] KL divergence

$$\mathcal{D} := \frac{1}{d} \mathbb{E}\, d_{\mathrm{KL}}[\mathcal{N}(0, \hat{\Sigma}) \,\|\, \mathcal{N}(0, \Sigma^\star + \sigma^2 I)] \tag{1.10}$$

where $\hat{\Sigma}$ is the empirical covariance matrix and Σ^* is the theoretical one. through the bounds

$$-\tfrac{1}{2} \log q \;\le\; \mathcal{D} \;\le\; 2\big[1 + \log(1 + \sigma^{-2})\big](1 - q). \tag{1.11}$$

As $q \to 1$, both sides go to zero, so convergence is equivalent to the KL divergence vanishing.

[3]So called because the theoretical distribution is shifted in excess of the true one, as a consequence of the regularized covariance estimator; see Section 3.1.1 for the precise definition.

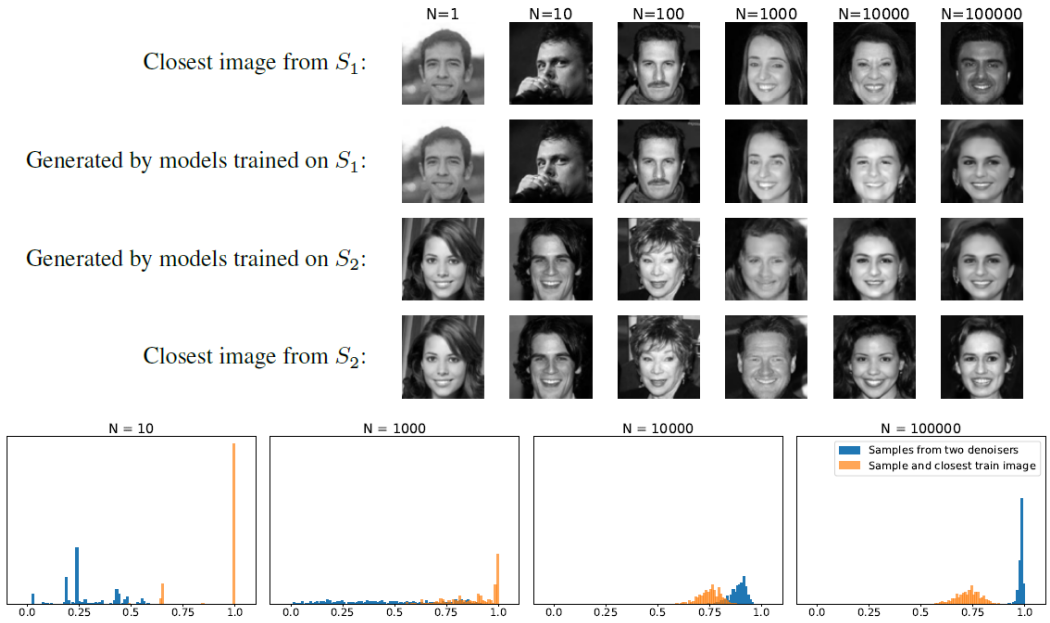


Figure 1.5: Convergence of model variance. Diffusion models are trained on disjoint subsets S_1 and S_2 of a face dataset. The distribution of cosine similarity between samples from the two models (blue) shifts rightward with increasing N, showing vanishing model variance. Conversely, the distribution of cosine similarity between a generated sample and its nearest training image (orange) shifts leftward. From [1].

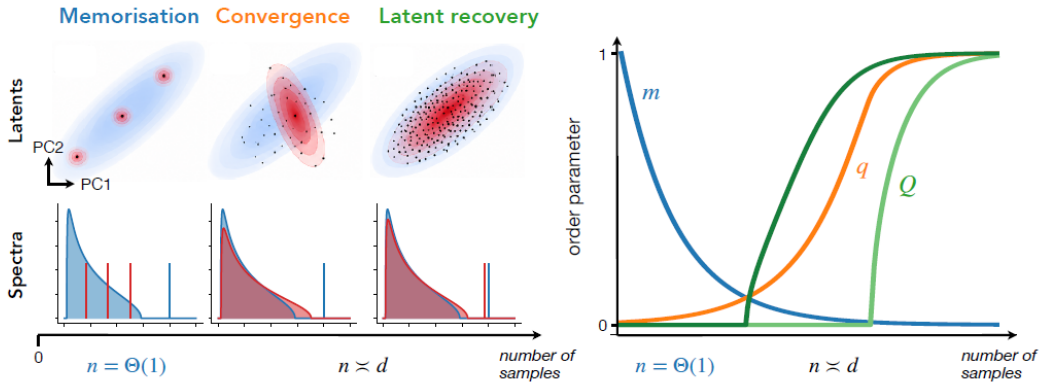


Figure 1.6: Left: the principal components (top) and the covariance spectra (bottom) of the true data distribution (blue) and a distribution learnt from n samples (red) in d dimensions. Right: three order parameters in function of the number of samples n. From [2].

The third phenomenon is *latent recovery*: consistent estimation of the principal latent directions of the data, measured by the subspace overlap

$$Q := |\langle v_{\max}(\hat{\Sigma}^{(0)}), v_{\max}(\hat{\Sigma}^{(1)})\rangle|$$

Unlike q, this quantity undergoes a sharp phase transition of the Baik–Ben Arous–Péché (BBP) type[4]: Q is identically zero below a threshold Q_{BBP} and jumps to a non-trivial value above it. Instead of the KL divergence, latent recovery governs the maximum-sliced (MS) distance between $P^\star$ and $\hat{P}$[5]. A key point is that q and Q are independent: models can reach $q \approx 1$ while Q remains near zero, and vice versa (Figure 1.6).

1.4.2 Algorithmic Stability

The KGSM experiment is a clear example of algorithmic stability [2]. In classical statistical learning theory, a model is stable if its output changes little when a single training point is removed; classic results by Vapnik and Chervonenkis show that stability of this kind implies little generalization error, as treated in [12]. The KGSM test is a more extreme version: instead of removing one point, the entire training set is replaced by an independent draw of the same size, and the question is whether the learned function remains the same [1, 2]. If it does, the model is a function of the distribution, not of the particular data, which is precisely what $q \to 1$ encodes. Maillard and Goldt make this connection quantitative through Equation (1.11): convergence is equivalent to the KL divergence between the true and learned distributions becoming small.

1.4.3 Why Influence Functions?

Computing q requires training two full models from scratch for every configuration one wants to study.

Influence functions offer a more intuitive and computation-friendly approach. Given a single trained model, the leave-one-out (LOO) influence of a training point $\mathbf{x}_i$ is a first-order approximation of how the model's output would change if $\mathbf{x}_i$ were removed [13], identified in the notation $\hat{p}_{(-i)}$ and $\hat{\Sigma}_{(-i)}$. It is, in fact, defined as

$$\Delta_i = \ln \hat{p}(\mathbf{x}_i|\hat{\Sigma}) - \ln \hat{p}_{(-i)}(\mathbf{x}_i|\hat{\Sigma}_{(-i)}). \tag{1.12}$$

We also define the **average rescaled influence density** as:

$$\mathcal{I} := \frac{\overline{\Delta}}{d} = \frac{1}{d\,n}\sum_{i=1}^{n}\Delta_i \tag{1.13}$$

The influence density is computable from one trained model without retraining, and provides a per-sample decomposition of the model's dependence on its data. This raises the central question of the thesis:

[4]The BBP transition is further explained in appendix A.

[5]MS uses a smooth test function to measure the largest discrepancy between two multivariate distributions over all linear projections. [2]

Can the KL divergence—the gap between the true and learned distributions—be bounded directly in terms of a leave-one-out influence function computed on a single trained model?

1.5 Thesis Contribution

We work in the same Gaussian framework as Maillard and Goldt [2], except that we do not include the rank-one spike, which is not relevant for defining $\mathcal{D}$ in this setting. Throughout, n denotes the number of i.i.d. training samples, d the ambient dimension of the data, and σ^2 the ridge (regularization) parameter added to the empirical covariance estimator: data are drawn from $\mathcal{N}(0, \mathbb{I}_d)$, and the generative model estimates the true covariance with the regularised sample estimator

$$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top + \sigma^2 \mathbb{I}_d.$$

The regularization parameter σ^2 is what produces the excess shift in the learned distribution discussed in Section 3.1.1.

We derive the LOO influence function Δ_i, which measures the change in the model's log-likelihood of $\mathbf{x}_i$ when that sample is removed from the training set.

We then take the thermodynamic limit $n, d \to \infty$ with the ratio $d/n \to \gamma$ held fixed, in order to reveal the spectral properties of $\hat{\Sigma}$ in the linear regime $n \asymp d$. In this limit, the eigenvalues of sample covariance matrices such as $\hat{\Sigma}$ no longer fluctuate around the true spectrum but converge to a deterministic limiting distribution: this is the content of the Marchenko–Pastur law, discussed in Appendix A.

Combining this with self-consistent equations from high-dimensional ridge regression, we evaluate the behaviour of $\mathcal{D}$ and $\mathcal{I}$. After showing that the upper bound on the Kullback–Leibler divergence is, in fact, exactly the average influence density $\mathcal{I}$, we derive a matching lower bound in terms of $\mathcal{I}$ as well, establishing that $\mathcal{D}$ and $\mathcal{I}$ are of the same order. These results are the single-model analogue of the bounds in (1.11).

The last logical step of this path is to connect the influence-function perspective back to the Maillard–Goldt framework, and hence to the KGSM experiment, by extending the LOO analysis to a comparison between models trained on different data subsets of size n. This extension suggests a relationship between $\mathcal{I}$ and the convergence overlap q, bridging the single-model and two-subset viewpoints.

Chapter 2

Influence Functions

We ourselves feel that what we are doing is just a drop in the ocean.
But the ocean would be less because of that missing drop.
– Mother Teresa

2.1 A Metric for Generalization

Conjecture 1 (Influence as a generalization proxy)**.** *Let $\hat{\Sigma}$ be the regularized empirical covariance estimator fitted on n i.i.d. samples in $\mathbb{R}^d$, let $\mathcal{D}$ be the rescaled excess Kullback–Leibler divergence* (1.10)*, and let $\mathcal{I} := \overline{\Delta}/d$ be the average rescaled leave-one-out influence density of $\hat{\Sigma}$, computable from a single trained model. Then*

$$\mathcal{D} \asymp \mathcal{I} \qquad \text{for every } \sigma^2 > 0 \text{ and every sample complexity } \alpha = n/d, \text{ with } n, d \to \infty$$

i.e. the two quantities vanish together, diverge together at the interpolation threshold, and are bounded by one another up to constants depending only on σ^2.

This conjecture proposes $\mathcal{I}$ as a cheap, single-model substitute for $\mathcal{D}$: unlike the convergence overlap q of Maillard and Goldt [2], which requires training two independent models to completion, $\mathcal{I}$ is computable in closed form from one fitted covariance estimator. We test the conjecture in the simplest setting where it can be verified *exactly*: isotropic Gaussian data with no low-rank spike. Chapters 2 and 3 establish it rigorously in this case, first via the upper bound $\mathcal{D} \leq \mathcal{I}$ (Theorem 3.2.1), and then via a matching lower bound (Section 3.3), showing $\mathcal{D}$ and $\mathcal{I}$ are of the same order for every regularization strength σ^2. Extending the conjecture to spiked covariance structures and non-linear denoisers is discussed in Chapter 5.

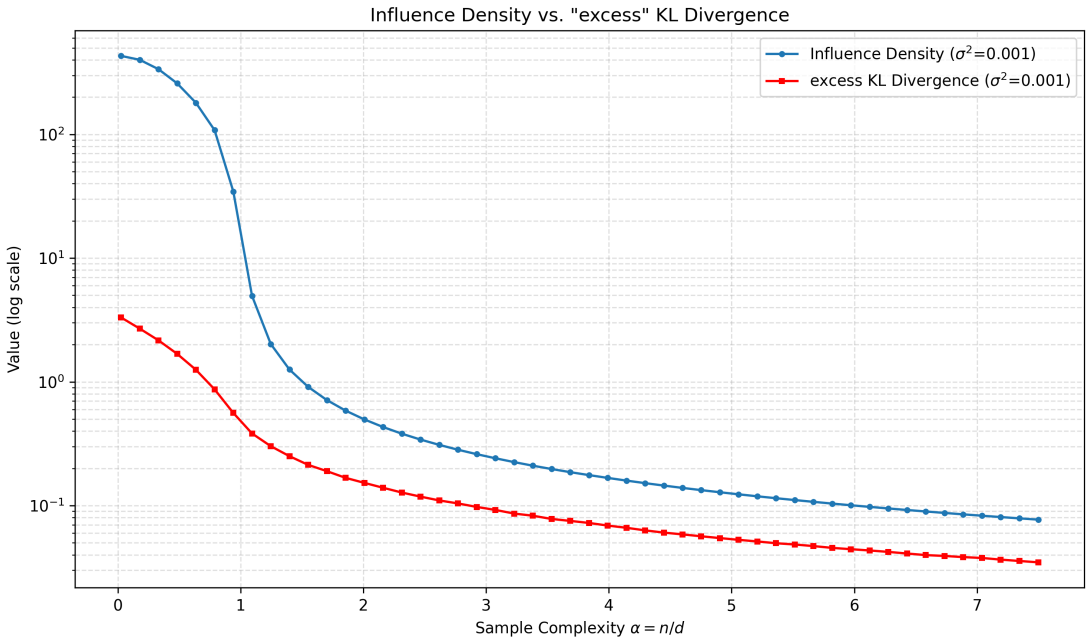


Figure 2.1: Result of the inequality simulated in `Python`. This result will be demonstrated throughout this work.

2.2 The Leave-One-Out (LOO) Influence Function

2.2.1 A brief history of influence functions

The notion of an influence function has a long history in statistics that predates its use here. Von Mises [14] first formalized the idea of measuring how an estimator's value responds to a small perturbation of the underlying distribution, an object that later became known as the (theoretical) influence function. Hampel, Ronchetti, Rousseeuw and Stahel [13] developed this into the foundation of robust statistics: an estimator whose influence function is unbounded is arbitrarily sensitive to a single contaminated observation, while a bounded influence function is the defining property of a robust estimator. Alongside this theoretical object, the classical literature also defines an *empirical* influence function, obtained not by perturbing the distribution but by directly comparing the estimator computed on the full sample to the estimator recomputed after removing, or replacing, a single observation. Our use of the LOO influence function follows this same motivation.

2.2.2 Derivation of the Leave-One-Out Influence Function

We begin by fixing the setting of our model:

- the number of generated samples n;

- the feature dimension d. Each sample $\mathbf{x}_i \in \mathbb{R}^d$;
- the regularization coefficient σ^2;
- all these samples are generated from a gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$.

Throughout, we work with purely Gaussian data, which is both easy to quantify and admits a closed-form Kullback-Leibler divergence between any two members of the family.

The resulting generative model is a **linear denoiser**—the simplest tractable case, and one that can be extended along the lines sketched in Chapter 5.

We then redefine the influence function directly from the structure of the covariance matrices themselves, which enter the Gaussian log-likelihood in two distinct roles: as a *quadratic form*, and as the argument of a logarithm through the *determinant*. The two matrices we compare, $\hat{\Sigma}_n$ and $\hat{\Sigma}_{(-i)}$, are related in the simplest possible way: one is exactly a rank-one perturbation of the other, differing only by the contribution of the single removed point $\mathbf{x}_i\mathbf{x}_i^\top$.

2.2.3 Definitions and Setup

We define the unnormalized, **regularized** sample covariance matrix S_n and its rigorous normalized version $\hat{\Sigma}_n$, consistent with the ridge estimator introduced in Section 1.3:

$$S_n = \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T + n\sigma^2 \mathbb{I}_d \implies \hat{\Sigma}_n = \frac{1}{n} S_n = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T + \sigma^2 \mathbb{I}_d \tag{2.1}$$

$$S_{(-i)} = S_n - \mathbf{x}_i \mathbf{x}_i^T \implies \hat{\Sigma}_{(-i)} = \frac{1}{n-1} S_{(-i)} \tag{2.2}$$

Note that removing the point $\mathbf{x}_i$ only removes its own rank-one contribution: the ridge term $n\sigma^2\mathbb{I}_d$ is unaffected, so $S_{(-i)} = S_n - \mathbf{x}_i\mathbf{x}_i^T$ holds exactly as in the unregularized case.

We define the leverage of the point $\mathbf{x}_i$ against the full model as the i-th diagonal element of the ridge hat matrix $\mathbf{H}$ [15][1] denoted as h_i:

$$\mathbf{H} := \mathbf{X} \left(\mathbf{X}^\mathsf{T}\mathbf{X} + n\sigma^2 \mathbb{I}_d \right)^{-1} \mathbf{X}^\mathsf{T} \tag{2.3}$$

$$h_i = \frac{1}{n} \mathbf{x}_i^T \hat{\Sigma}_n^{-1} \mathbf{x}_i = \mathbf{x}_i^T S_n^{-1} \mathbf{x}_i \tag{2.4}$$

The log-likelihood of a zero-mean Gaussian with Σ covariance matrix is given by:

$$\ln p(\mathbf{x}_i \mid \Sigma) = C - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i \tag{2.5}$$

[1] In statistics, the *projection* or *hat* matrix H maps the vector of response values (dependent variable values) to the vector of fitted values (or predicted values). It describes the influence each response value has on each fitted value. The diagonal elements of the projection matrix are the *leverages*, which describe the influence each response value has on the fitted value for that same observation.

Therefore, the difference in equation 1.12 naturally splits into a determinant term and a quadratic form term:

$$\Delta_i = \underbrace{\frac{1}{2}\ln\left(\frac{|\hat{\Sigma}_{(-i)}|}{|\hat{\Sigma}_n|}\right)}_{\text{Determinant Difference}} + \underbrace{\frac{1}{2}\left(\mathbf{x}_i^T\hat{\Sigma}_{(-i)}^{-1}\mathbf{x}_i - \mathbf{x}_i^T\hat{\Sigma}_n^{-1}\mathbf{x}_i\right)}_{\text{Quadratic Difference}} \tag{2.6}$$

2.2.4 The Determinant Difference

We relate the LOO covariance matrix to the full covariance matrix:

$$\hat{\Sigma}_{(-i)} = \frac{1}{n-1}(n\hat{\Sigma}_n - \mathbf{x}_i\mathbf{x}_i^T) = \frac{n}{n-1}\left(\hat{\Sigma}_n - \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^T\right) \tag{2.7}$$

Since the (regularized) covariance matrix is invertible, we exploit the *matrix determinant lemma*, $|A + uv^T| = |A|(1 + v^T A^{-1}u)$ [16]:

$$|\hat{\Sigma}_{(-i)}| = \left(\frac{n}{n-1}\right)^d \left|\hat{\Sigma}_n - \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^T\right| \tag{2.8}$$

$$= \left(\frac{n}{n-1}\right)^d |\hat{\Sigma}_n|\left(1 - \frac{1}{n}\mathbf{x}_i^T\hat{\Sigma}_n^{-1}\mathbf{x}_i\right) \tag{2.9}$$

$$= \left(\frac{n}{n-1}\right)^d |\hat{\Sigma}_n|(1 - h_i) \tag{2.10}$$

Taking the logarithm of the ratio gives the determinant difference:

$$\frac{1}{2}\ln\left(\frac{|\hat{\Sigma}_{(-i)}|}{|\hat{\Sigma}_n|}\right) = \frac{1}{2}\left[d\ln\left(\frac{n}{n-1}\right) + \ln(1 - h_i)\right] \tag{2.11}$$

2.2.5 The Precision Matrix and Quadratic Form Difference

We apply the Sherman-Morrison identity [16] to the unnormalized precision matrix $S_{(-i)}^{-1}$.

Lemma 2.2.1 (Sherman-Morrison Identity)**.** *Suppose $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are column vectors. Then $A + \mathbf{u}\mathbf{v}^T$ is invertible if and only if $1 + \mathbf{v}^T A^{-1}\mathbf{u} \neq 0$. We have:*

$$(A + \mathbf{u}\mathbf{v}^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}}{1 + \mathbf{v}^T A^{-1}\mathbf{u}}. \tag{2.12}$$

So, in our case:

$$S_{(-i)}^{-1} = (S_n - \mathbf{x}_i\mathbf{x}_i^T)^{-1} = S_n^{-1} + \frac{S_n^{-1}\mathbf{x}_i\mathbf{x}_i^T S_n^{-1}}{1 - \mathbf{x}_i^T S_n^{-1}\mathbf{x}_i} = S_n^{-1} + \frac{S_n^{-1}\mathbf{x}_i\mathbf{x}_i^T S_n^{-1}}{1 - h_i} \tag{2.13}$$

Substituting the rigorous scalings $S_n^{-1} = \frac{1}{n}\hat{\Sigma}_n^{-1}$ and $S_{(-i)}^{-1} = \frac{1}{n-1}\hat{\Sigma}_{(-i)}^{-1}$:

$$\frac{1}{n-1}\hat{\Sigma}_{(-i)}^{-1} = \frac{1}{n}\hat{\Sigma}_n^{-1} + \frac{\frac{1}{n^2}\hat{\Sigma}_n^{-1}\mathbf{x}_i\mathbf{x}_i^T\hat{\Sigma}_n^{-1}}{1-h_i} \tag{2.14}$$

Multiplying by $n-1$ yields the LOO precision matrix:

$$\hat{\Sigma}_{(-i)}^{-1} = \frac{n-1}{n}\hat{\Sigma}_n^{-1} + \frac{n-1}{n^2}\frac{\hat{\Sigma}_n^{-1}\mathbf{x}_i\mathbf{x}_i^T\hat{\Sigma}_n^{-1}}{1-h_i} \tag{2.15}$$

We now calculate the quadratic form $\mathbf{x}_i^T\hat{\Sigma}_{(-i)}^{-1}\mathbf{x}_i$ by pre- and post-multiplying by $\mathbf{x}_i^T$ and $\mathbf{x}_i$. Recalling that $\mathbf{x}_i^T\hat{\Sigma}_n^{-1}\mathbf{x}_i = nh_i$:

$$\mathbf{x}_i^T\hat{\Sigma}_{(-i)}^{-1}\mathbf{x}_i = \frac{n-1}{n}(nh_i) + \frac{n-1}{n^2}\frac{(nh_i)^2}{1-h_i} \tag{2.16}$$

$$= (n-1)h_i + \frac{(n-1)h_i^2}{1-h_i} \tag{2.17}$$

$$= \frac{(n-1)h_i(1-h_i) + (n-1)h_i^2}{1-h_i} = \frac{(n-1)h_i}{1-h_i} \tag{2.18}$$

Now, we find the difference between the LOO quadratic form and the full quadratic form:

$$\frac{1}{2}\left(\mathbf{x}_i^T\hat{\Sigma}_{(-i)}^{-1}\mathbf{x}_i - \mathbf{x}_i^T\hat{\Sigma}_n^{-1}\mathbf{x}_i\right) = \frac{1}{2}\left[\frac{(n-1)h_i}{1-h_i} - nh_i\right] \tag{2.19}$$

$$= \frac{1}{2}\left[\frac{(n-1)h_i - nh_i(1-h_i)}{1-h_i}\right] \tag{2.20}$$

$$= \frac{1}{2}\left[\frac{nh_i - h_i - nh_i + nh_i^2}{1-h_i}\right] \tag{2.21}$$

$$= \frac{1}{2}\left[\frac{nh_i^2 - h_i}{1-h_i}\right] \tag{2.22}$$

2.2.6 The Final Influence Function

Combining the determinant difference and the quadratic difference, we arrive at the exact expression for the leave-one-out shift:

$$\Delta_i = \frac{1}{2}\left[d\ln\left(\frac{n}{n-1}\right) + \ln(1-h_i) + \frac{nh_i^2 - h_i}{1-h_i}\right] \tag{2.23}$$

2.3 Asymptotic Behaviour

We write the rescaled influence density, averaged on the n samples:

$$\frac{\overline{\Delta}}{d} = \frac{1}{2nd}\sum_{i=1}^{n}\left[\ln(1-h_i) + d\ln(\frac{n}{n-1}) + \frac{h_i(nh_i-1)}{1-h_i}\right] \tag{2.24}$$

In order to obtain a proper asymptotic version of the influence density, we must bridge the finite-dimensional geometric properties of the dataset with the asymptotic properties of large random matrices. The leverage score, as explicated above, is:

$$h_i = \frac{1}{n}\mathbf{x}_i^T \hat{\Sigma}_n^{-1} \mathbf{x}_i. \tag{2.25}$$

By the Sherman-Morrison identity, we can decouple the vector $\mathbf{x}_i$ from the matrix it is multiplied against by rewriting h_i in terms of the Leave-One-Out covariance matrix $\hat{\Sigma}_{(-i)}$, as an inverse expression of Equation 2.16:

$$h_i = \frac{\frac{1}{n}\mathbf{x}_i^T \hat{\Sigma}_{(-i)}^{-1} \mathbf{x}_i}{1 + \frac{1}{n}\mathbf{x}_i^T \hat{\Sigma}_{(-i)}^{-1} \mathbf{x}_i} \tag{2.26}$$

To establish the concentration of the quadratic form, we first recall the Hanson-Wright inequality [17], which bounds the tail probabilities for quadratic forms of sub-Gaussian random vectors.

Theorem 2.3.1 (Hanson-Wright Inequality)**.** *Let $\mathbf{v} \in \mathbb{R}^d$ be a random vector with independent, mean-zero, sub-Gaussian coordinates with unit variance ($\mathbb{E}[\mathbf{v}\mathbf{v}^T] = \mathbb{I}_d$), and let $A \in \mathbb{R}^{d\times d}$ be a deterministic matrix. Then, for any $t > 0$,*

$$\mathbb{P}\left(\left|\mathbf{v}^T A \mathbf{v} - \mathrm{Tr}(A)\right| > t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_{HS}^2}, \frac{t}{K^2\|A\|}\right)\right], \tag{2.27}$$

where $K = \|\mathbf{v}_i\|_{\psi_2}$ is the maximum sub-Gaussian norm of the coordinates, and $c > 0$ is an absolute constant.

The concentration on the trace for matrices in high dimensions follows:

Lemma 2.3.2 (Trace Lemma)**.** *For a random isotropic sub-Gaussian vector $\mathbf{v} \in \mathbb{R}^d$ and a deterministic matrix $A \in \mathbb{R}^{d\times d}$ with well-behaved norms, the quadratic form concentrates almost surely around its trace as the dimension grows:*

$$\mathbf{v}^T A \mathbf{v} \xrightarrow{a.s.} \mathrm{Tr}(A). \tag{2.28}$$

Because the observation $\mathbf{x}_i$ is an isotropic Gaussian vector ($\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T] = \mathbb{I}_d$) drawn independently from the remaining samples in $\hat{\Sigma}_{(-i)}$, it inherently satisfies the sub-Gaussian condition. Invoking the Lemma on concentration on the trace, the normalized quadratic form concentrates almost surely around its expected [2]:

$$\frac{1}{n}\mathbf{x}_i^T \hat{\Sigma}_{(-i)}^{-1} \mathbf{x}_i \xrightarrow{a.s.} \frac{1}{n}\,\mathrm{Tr}\left(\hat{\Sigma}_{(-i)}^{-1}\right). \tag{2.29}$$

[2]Here you can spot a subtlety. The covariance matrix is inherently random, while the lemma works for deterministic matrices. Actually, the observation $\mathbf{x}_i$ is drawn independently from the remaining samples, hence we can condition on $\hat{\Sigma}_{(-i)}$. Conditional on this matrix, $\mathbf{x}_i$ is an isotropic Gaussian vector ($\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T] = \mathbb{I}_d$) evaluated against a fixed matrix, allowing us to invoke concentration results such as the Hanson-Wright theorem [17].

Rank-1 Equivalence in the Thermodynamic Limit In the thermodynamic limit $(n, d \to \infty)$, the removal of a single rank-1 update $(\mathbf{x}_i \mathbf{x}_i^T)$ does not alter the macroscopic spectrum of the sample covariance matrix. Therefore, the trace of the LOO precision matrix is asymptotically equivalent to the full precision matrix:

$$\frac{1}{n}\mathrm{Tr}(\hat{\Sigma}_{(-i)}^{-1}) \simeq \frac{1}{n}\mathrm{Tr}(\hat{\Sigma}_n^{-1}) \tag{2.30}$$

Consequently, the individual leverages lose their dependence on the specific geometric realization of the data point $\mathbf{x}_i$. They self-average, collapsing to a uniform, deterministic constant $h_i \to h$ for all $i = 1, \ldots, n$.

Because h_i becomes uniform, the sum of all n leverage scores is simply nh. This sum corresponds exactly to the trace of the hat matrix H:

$$\sum_{i=1}^{n} h_i = nh = \mathrm{Tr}\left[\frac{1}{n} X \left(\frac{1}{n} X^T X + \sigma^2 \mathbb{I}_d\right)^{-1} X^T\right] \tag{2.31}$$

By the cyclic property of the trace, the non-zero eigenvalues of XAX^T are identical to the eigenvalues of $X^T XA$. Let $\hat{\lambda}_j$ denote the d empirical eigenvalues of the unregularized sample covariance matrix $\frac{1}{n}X^T X$. We can rewrite the trace as a sum over these eigenvalues:

$$nh = \sum_{j=1}^{d} \frac{\hat{\lambda}_j}{\hat{\lambda}_j + \sigma^2} \tag{2.32}$$

Dividing both sides by n, and simultaneously multiplying and dividing the right side by d, we introduce the aspect ratio $\gamma = d/n$:

$$h = \frac{d}{n}\frac{1}{d}\sum_{j=1}^{d} \frac{\hat{\lambda}_j}{\hat{\lambda}_j + \sigma^2} = \gamma \int \frac{\lambda}{\lambda + \sigma^2} d\hat{\mu}(\lambda) \tag{2.33}$$

Where $\hat{\mu}(\lambda)$ is the empirical spectral distribution (ESD) of the sample covariance matrix.

When Physics Steps In As $n, d \to \infty$ with $d/n \to \gamma$, Random Matrix Theory from statistical mechanics guarantees that the empirical spectral measure $\hat{\mu}(\lambda)$ converges weakly, almost surely, to the deterministic Marchenko-Pastur probability measure $\mu_{MP}(\lambda)$[3]

Consequently, the discrete empirical integral converges to the theoretical expectation over the continuous bulk:

$$h \xrightarrow{n,d\to\infty} \gamma \mathbb{E}_{MP}\left[\frac{\lambda}{\lambda + \sigma^2}\right] \tag{2.34}$$

Setting $\mathcal{K} = \mathbb{E}_{MP}\left[\frac{\lambda}{\lambda+\sigma^2}\right]$, we arrive at the exact analytical substitution $h = \gamma\mathcal{K}$.

Substituting this concentration result into the summation in 2.24, we can evaluate the limit of each term inside the bracket individually.

[3]Appendix A for more on Random Matrix Theory

1. The First Logarithmic Term:

$$\frac{1}{nd}\sum_{i=1}^{n}\log(1-h_i) \xrightarrow{n,d\to\infty} \frac{1}{d}\log(1-\gamma\mathcal{K}) \to 0 \tag{2.35}$$

Since $\log(1-\gamma\mathcal{K})$ is $\mathcal{O}(1)$, dividing by $d \to \infty$ drives the term to zero.

2. The Second Logarithmic Term:

$$\frac{1}{nd}\sum_{i=1}^{n} d\log\left(\frac{n}{n-1}\right) = \log\left(1+\frac{1}{n-1}\right) \approx \frac{1}{n} \to 0 \tag{2.36}$$

3. The Dominant Rational Term: Expanding the numerator as $h_i(nh_i - 1) = nh_i^2 - h_i$, we get:

$$\frac{1}{nd}\sum_{i=1}^{n}\frac{nh_i^2-h_i}{1-h_i} \xrightarrow{n,d\to\infty} \frac{1}{d}\frac{n(\gamma\mathcal{K})^2-\gamma\mathcal{K}}{1-\gamma\mathcal{K}} = \frac{n}{d}\frac{\gamma^2\mathcal{K}^2}{1-\gamma\mathcal{K}} - \frac{1}{d}\frac{\gamma\mathcal{K}}{1-\gamma\mathcal{K}} \tag{2.37}$$

Since $1/d \to 0$, the second fraction vanishes. For the first fraction, we substitute the aspect ratio $n/d = 1/\gamma$:

$$\frac{1}{\gamma}\frac{\gamma^2\mathcal{K}^2}{1-\gamma\mathcal{K}} = \frac{\gamma\mathcal{K}^2}{1-\gamma\mathcal{K}} \tag{2.38}$$

Summing the limits of the three components and reintroducing the global $1/2$ factor, we recover the closed-form asymptotic expression, as shown in Figure 2.2:

$$\mathcal{I} = \frac{\overline{\Delta}}{d} \xrightarrow{n,d\to\infty} \frac{1}{2}\frac{\gamma\mathcal{K}^2}{1-\gamma\mathcal{K}} \tag{2.39}$$

2.3.1 The Unregularized Limit

In the unregularized regime ($\gamma < 1, \quad n > d$ and $\sigma^2 \to 0$), the sample covariance matrix is itself invertible. Taking the limit as $\sigma^2 \to 0$, the expected value of the inverse eigenvalues of the standard Marchenko-Pastur distribution yields $\mathcal{K} \to 1$.

Substituting this into our expression for $h = \gamma\mathcal{K}$:

$$h = \gamma \tag{2.40}$$

Which forces the unregularized Influence Density to diverge as the sample complexity approaches the interpolation threshold ($\gamma \to 1^-, \quad \alpha \to 1^+$):

$$\mathcal{I} = \frac{\bar{\Delta}}{d} = \frac{\gamma}{2(1-\gamma)} \tag{2.41}$$

This divergence mathematically translates the phenomenon of *memorization*: as the model exhausts its degrees of freedom, so $\gamma \to 1^-, \quad \alpha \to 1^+$, the out-of-sample stability shatters, and the influence of any single training point becomes catastrophically divergent, and is no longer captured by the asymptotic formula (Figure 2.3).

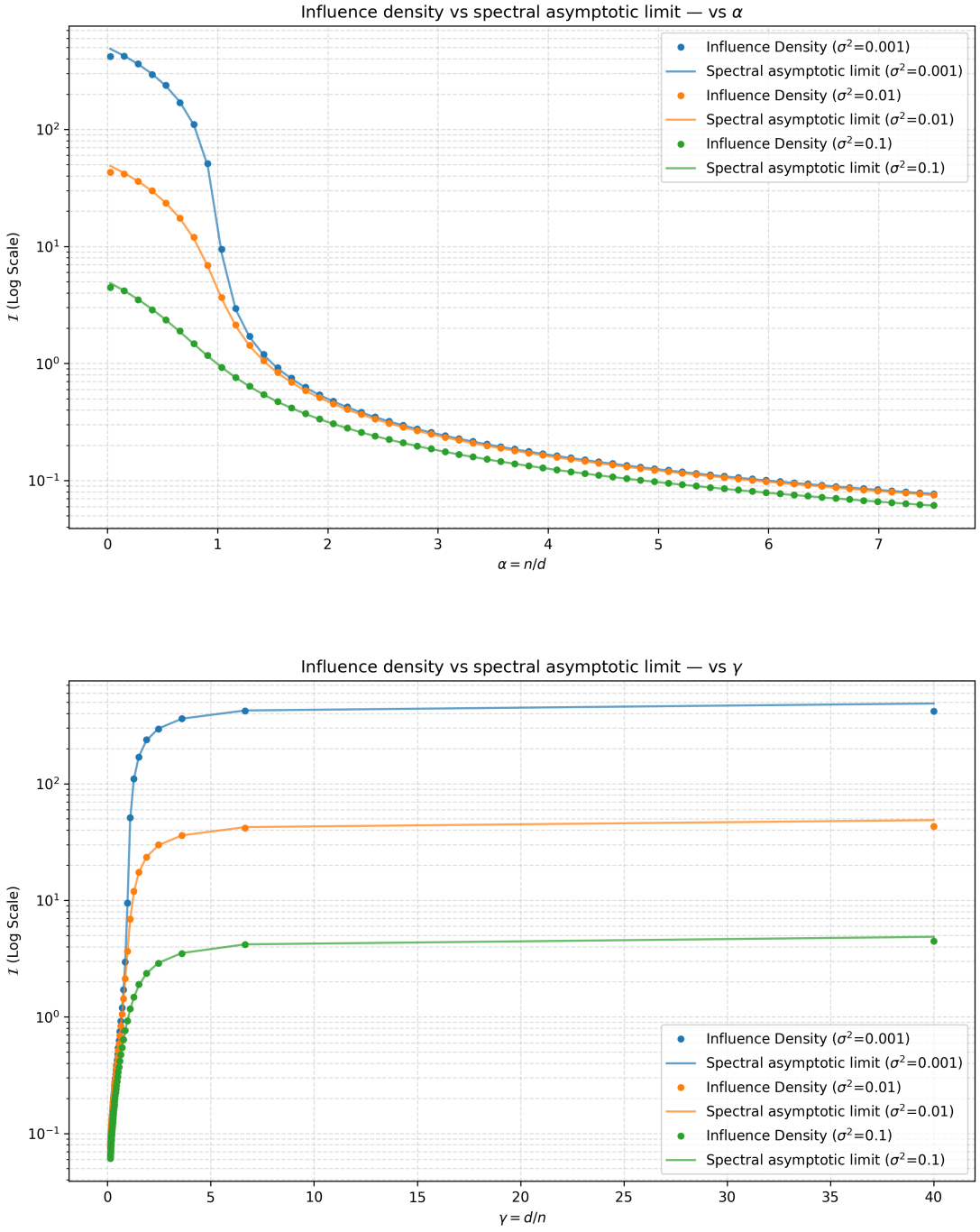


Figure 2.2: Influence function versus spectral expression for different values of the regularization coefficient, in function of α and γ. The asymptotic form was calculated in Python through numerical integration.

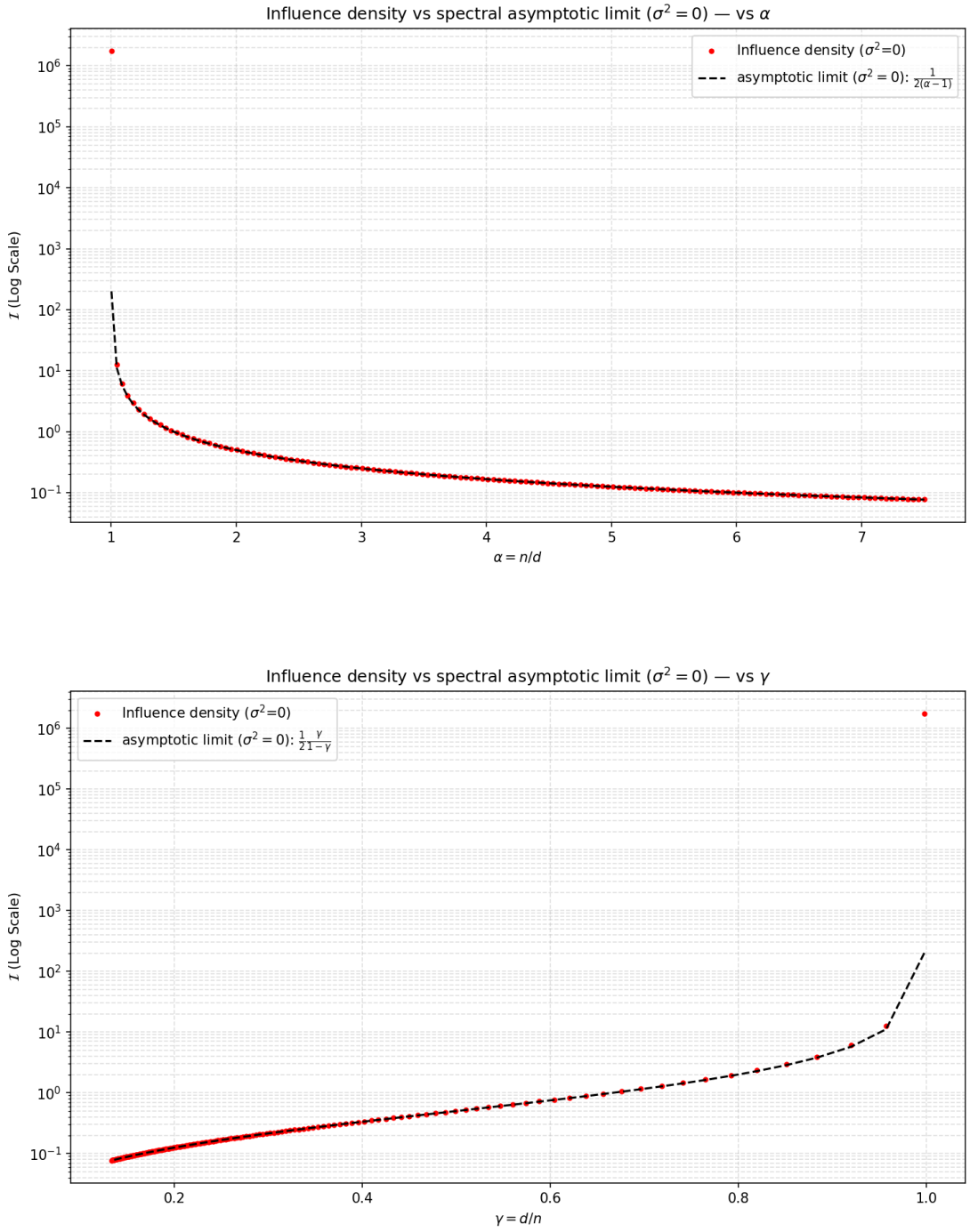


Figure 2.3: Influence function versus asymptotic exact expression in the thermo- dynamic limit, in the non-regularized regime, in function of α and γ.

Chapter 3

Bounds on $\mathcal{D}$

Another roof, another proof.
–Paul Erdős

3.1 Information-Theoretic Bounds on Generalization

To quantify how the empirical estimator $\hat{\Sigma}$ departs from structural stability, we analyze its behaviour through the lens of information theory. Specifically, we evaluate the Kullback-Leibler (KL) divergence between the empirical model and the theoretical baseline. Instead of calculating the forward divergence from the true underlying distribution to the empirical model, we calculate the **inverse KL divergence**: measuring the distance from the empirical estimator $\mathcal{N}(0, \hat{\Sigma})$ to the isotropic prior $\mathcal{N}(0, \Sigma^\star + \sigma^2 \mathbb{I}_d)$, shifted by a σ^2-unity to avoid problems in the trace [2] calculated afterwards in the Kullback-Leibler divergence. We explain the whole motivation right below.

3.1.1 Why an *excess* KL divergence?

To motivate this, we follow the argument of Maillard and Goldt [2].

Recall that our estimator is a *ridge*-regularized empirical covariance,

$$\hat{\Sigma} = \frac{1}{n} \sum_{\mu=1}^{n} \mathbf{x}_\mu \mathbf{x}_\mu^\top + \sigma^2 \mathbb{I}_d,$$

with a fixed regularization strength $\sigma^2 > 0$. As $n \to \infty$, the empirical term $\frac{1}{n} \sum_\mu \mathbf{x}_\mu \mathbf{x}_\mu^\top$ concentrates on the true covariance $\Sigma^\star$ by the law of large numbers—but $\hat{\Sigma}$ itself does not converge to $\Sigma^\star$: it converges to $\Sigma^\star + \sigma^2 I_d$. This residual gap is a deterministic bias (See Figure 3.1 on the left), fixed entirely by our choice of σ^2, that persists even with an infinite training set.

If we naively measured

$$d_{\mathrm{KL}}\big[\mathcal{N}(0, \hat{\Sigma}) \,\|\, \mathcal{N}(0, \Sigma^\star)\big],$$

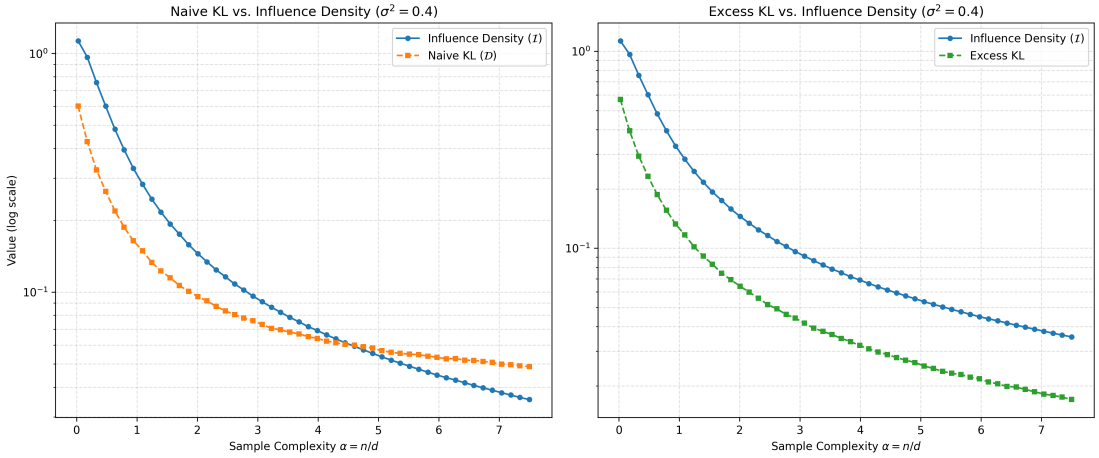


Figure 3.1: **Impact of Regularization Bias on KL Divergence.** Comparison between the Naive KL divergence (left) and the excess KL divergence (right) plotted against the influence density for a regularization strength of $\sigma^2 = 0.4$. On the left, measuring the Naive KL against the unregularized target $\mathbb{I}_d$ introduces a deterministic bias. As sample complexity $\alpha = n/d$ grows, the naive KL asymptotes to this constant offset, preventing fron causing an artificial crossing of the influence density bound. On the right, measuring the excess KL against the regularized target $\Sigma^\star + \sigma^2\mathbb{I}_d$ properly accounts for the shift, removing the bias. Consequently, the excess KL correctly vanishes and remains strictly upper-bounded by the influence density across all regimes.

this quantity would *never vanish*, no matter how much data we collect: it would asymptote to the fixed offset

$$d_0 = d_{\mathrm{KL}}\left[\mathcal{N}(0, \Sigma^\star + \sigma^2 I_d) \,\|\, \mathcal{N}(0, \Sigma^\star)\right],$$

a constant that reflects nothing about generalization – only about how strongly we chose to regularize. Any interesting, data-dependent signal would be swamped by this constant regularization offset, and *comparisons across different sample sizes n would be contaminated by a term that has nothing to do with n at all.*

The fix is to measure divergence against the regularized target $\Sigma^\star + \sigma^2 I_d$ instead – the value $\hat{\Sigma}$ would converge to *if* it were fitted on the true, infinite-data distribution with the same regularization (See Figure 3.1 on the right). This is precisely the quantity $\mathcal{D}$ defined above[1].

Using the spectral decomposition of the covariance matrix, we can express this rescaled excess KL divergence exclusively in terms of the Marchenko-Pastur bulk. As Maillard and Goldt show [2], the presence of a low-rank spike in $\Sigma^\star$ does not affect $\mathcal{D}$

[1]This choice is also motivated by a physical perspective: in the thermodynamic limit, if the model completely ignores the empirical data (infinite regularization, $\sigma^2 \to \infty$), the expected covariance matrix trivially collapses to the identity, and $\mathcal{D}$ correctly reflects that no information beyond the prior has been learned.

at leading order, so it suffices to work out the isotropic case $\Sigma^\star = I_d$ (indeed our framework) , for which the regularized target reduces to $(1+\sigma^2)I_d$. By defining the rescaled eigenvalue $\nu = \frac{\lambda+\sigma^2}{1+\sigma^2}$, the macroscopic limit simplifies elegantly to:

$$\mathcal{D} = -\frac{1}{2}\mathbb{E}_{MP}[\log \nu] \tag{3.1}$$

The full derivation is carried out in Appendix B.2 of Maillard and Goldt [2]; we do not reproduce it here and instead build directly on their result.

3.2 Bounding the Divergence

In this section, we demonstrate what we conjectured in 1 .

Theorem 3.2.1. *For every regularization level $\sigma^2 > 0$ and every sample complexity $\gamma = d/n$,*

$$\mathcal{D} \leq \mathcal{I}.$$

Proof. Let us recall the definition of the rescaled eigenvalue

$$\nu = \frac{\lambda+\sigma^2}{1+\sigma^2}.$$

We rely on the fundamental logarithmic inequality, which holds for any $\nu > 0$:

$$-\log \nu \leq \nu^{-1} - 1 \tag{3.2}$$

Taking the expectation with respect to the Marchenko-Pastur measure on both sides, the left-hand side perfectly recovers twice our rescaled excess KL divergence:

$$\mathbb{E}_{MP}[-\log \nu] = 2\mathcal{D} \tag{3.3}$$

For the right-hand side, we expand the expectation of the inverse variable:

$$\mathbb{E}_{MP}[\nu^{-1}] - 1 = \mathbb{E}_{MP}\left[\frac{1+\sigma^2}{\lambda+\sigma^2}\right] - 1 = (1+\sigma^2)\mathbb{E}_{MP}\left[\frac{1}{\lambda+\sigma^2}\right] - 1 \tag{3.4}$$

We can express this expected inverse moment formally through the Stieltjes transform of the Marchenko-Pastur distribution. Following the framework established in a recent high-dimensional generalization paper by Hastie et al. [18], the Stieltjes transform of the spectral measure μ_{MP} evaluated at $z \in \mathbb{C} \setminus \mathbb{R}^+$ is defined as:

$$s(z) = \mathbb{E}_{MP}\left[\frac{1}{\lambda - z}\right] \tag{3.5}$$

By evaluating the transform at the negative regularization parameter, $z = -\sigma^2$, we exactly recover our expected resolvent trace:

$$s(-\sigma^2) = \mathbb{E}_{MP}\left[\frac{1}{\lambda+\sigma^2}\right] \tag{3.6}$$

We can seamlessly link this Stieltjes transform to the macroscopic leverage integral we defined in the previous chapter. $\mathcal{K} = \mathbb{E}_{MP}\left[\frac{\lambda}{\lambda+\sigma^2}\right]$ through a simple algebraic decomposition of the numerator. By adding and subtracting σ^2, we obtain:

$$\begin{aligned}\mathcal{K} &= \mathbb{E}_{MP}\left[\frac{\lambda+\sigma^2-\sigma^2}{\lambda+\sigma^2}\right] \\ &= \mathbb{E}_{MP}\left[1-\frac{\sigma^2}{\lambda+\sigma^2}\right] \\ &= 1-\sigma^2 s(-\sigma^2)\end{aligned} \tag{3.7}$$

Rearranging this identity yields the exact expression for the expected resolvent trace strictly in terms of $\mathcal{K}$ and σ^2:

$$s(-\sigma^2) = \frac{1-\mathcal{K}}{\sigma^2} \tag{3.8}$$

At this stage, we invoke the self-consistent equation for the Stieltjes transform of the Marchenko-Pastur distribution for ridge-regularized estimators, that can be consulted in [18][2]:

$$s(-\sigma^2) = \frac{1}{1-\gamma+\sigma^2+\gamma\sigma^2 s(-\sigma^2)} \tag{3.9}$$

Substituting Equation 3.8 in the equation:

$$\frac{1-\mathcal{K}}{\sigma^2} = \frac{\mathcal{K}}{1-\gamma\mathcal{K}} \implies s = \frac{\mathcal{K}}{1-\gamma\mathcal{K}} = \mathbb{E}_{MP}\left[\frac{1}{\lambda+\sigma^2}\right] \tag{3.10}$$

Because $\sigma^2 > 0$ and $\mathcal{K} > 0$, the denominator forces the strict stability constraint $\gamma\mathcal{K} < 1$, making the asymptotic form of the influence density a well defined function. Substituting this into our inequality's right-hand side yields:

$$\begin{aligned}\mathbb{E}_{MP}[\nu^{-1}] - 1 &= (1+\sigma^2)\left(\frac{1-\mathcal{K}}{\sigma^2}\right) - 1 \\ &= \frac{1-\mathcal{K}}{\sigma^2} + (1-\mathcal{K}) - 1 \\ &= \frac{1-\mathcal{K}}{\sigma^2} - \mathcal{K}\end{aligned} \tag{3.11}$$

[2]Hastie, Montanari, Rosset and Tibshirani [18] state the Silverstein equation in terms of the *companion* Stieltjes transform $v \equiv v(-\sigma^2)$ of the $n \times n$ matrix $\frac{1}{n}X^\top X$, namely

$$\sigma^2 + \frac{\gamma}{1+v} = \frac{1}{v}.$$

The transform v is related to the direct Stieltjes transform $s \equiv s(-\sigma^2)$ of the $d \times d$ feature covariance $\frac{1}{n}XX^\top$ by the trace identity

$$v = \gamma\, s + \frac{1-\gamma}{\sigma^2},$$

which follows from the fact that the two matrices share the same non-zero eigenvalues and differ only by $|n-d|$ zero eigenvalues. Substituting this identity into the companion equation recovers the self-consistent equation for s written here.

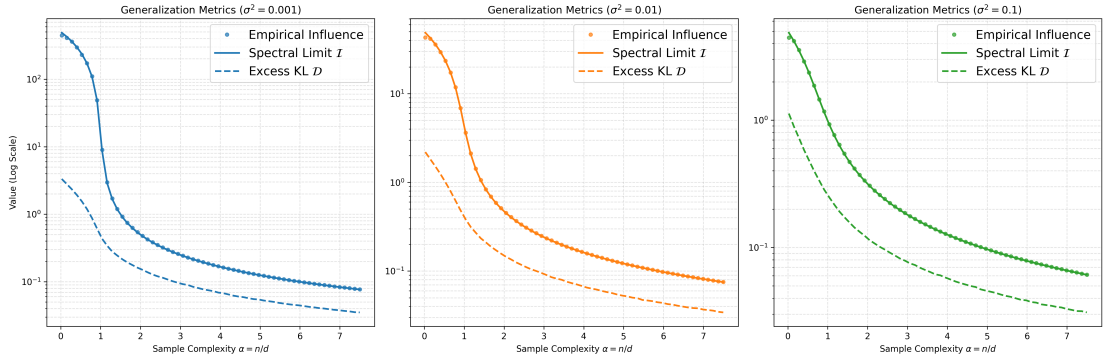


Figure 3.2: Bound right function and asymptotic limit comparison, for different values of regularization.

Substituting this identity into Equation 3.11 produces a striking algebraic result:

$$\begin{aligned}\mathbb{E}_{MP}[\nu^{-1}] - 1 &= \frac{\mathcal{K}}{1-\gamma\mathcal{K}} - \mathcal{K} \\ &= \frac{\mathcal{K} - \mathcal{K}(1-\gamma\mathcal{K})}{1-\gamma\mathcal{K}} \\ &= \frac{\gamma\mathcal{K}^2}{1-\gamma\mathcal{K}}\end{aligned} \tag{3.12}$$

This final expression is, in fact, exactly twice the macroscopic influence density $\mathcal{I}$ we derived in the previous sections ($\mathcal{I} = \frac{\gamma\mathcal{K}^2}{2(1-\gamma\mathcal{K})}$). Therefore, the thermodynamic inequality closes perfectly[3]:

$$2\mathcal{D} \leq \frac{\gamma\mathcal{K}^2}{1-\gamma\mathcal{K}} = 2\mathcal{I} \implies \mathcal{D} \leq \mathcal{I} \tag{3.13}$$

The Unregularized Limit Divergence

This bound becomes particularly illuminating in the unregularized limit ($\sigma^2 \to 0$), when the leverage integral approaches unity ($\mathcal{K} \to 1$).

Applying this limit to our density equation, we observe the exact asymptotic behaviour of the divergence:

$$\mathcal{I} \xrightarrow{\sigma^2 \to 0} \frac{\gamma}{2(1-\gamma)} = \frac{1}{2(\alpha-1)} \tag{3.14}$$

Where $\alpha = n/d = 1/\gamma$ is the sample complexity. This confirms that as the system approaches the interpolation threshold ($\alpha \to 1^+$), the microscopic instability diverges.

[3]for fancier GIFs showing this empirically, try exploring `https://github.com/Bombaric/Thesis` :)

3.3 Sandwiching the Bounds

We now aim to find a strict lower bound for the rescaled excess KL divergence

$$\mathcal{D} = -\frac{1}{2}\mathbb{E}_{MP}(\log \nu)$$

using the asymptotic influence density

$$\mathcal{I} = \frac{\bar{\Delta}}{d} = \frac{\gamma \mathcal{K}^2}{2(1-\gamma\mathcal{K})} = \frac{1}{2}\mathbb{E}_{MP}(\nu^{-1} - 1)$$

The expectation is taken over the Marchenko-Pastur (MP) spectral measure, with the rescaled eigenvalue (random variable) ν.

3.3.1 The sign-flip problem

We initially tried to find a uniform constant C able to bound the actual functions pointwise across the entire support of the spectrum:

$$-\log \nu \geq C(\nu^{-1} - 1) \tag{3.15}$$

And then averaging both sides of the equation to reobtain an inequality of the form $C\,\mathcal{I} \leq \mathcal{D}$. Unfortunately, this pointwise inequality fails globally. Because the Marchenko-Pastur distribution is centered such that $\mathbb{E}[\nu] = 1$, the eigenvalues necessarily span across $\nu = 1$. When $\nu > 1$, the term $(\nu^{-1} - 1)$ becomes negative (Figure 3.3) Consequently, dividing by this term flips the direction of the inequality, making it impossible to establish a uniform constant C that holds for all ν. This condition is overly strict because we attempted to bound the arguments of the expectations directly. To resolve this, we must relax the pointwise constraint and evaluate the macroscopic expectations directly.

3.3.2 The zero-expectation trick

To bypass the sign-flip, we added the linear term $(\nu - 1)$ to both sides of the inequality. Working under the Marchenko-Pastur measure, we know that the expected value of the eigenvalues is strictly 1:

$$\mathbb{E}_{MP}[\nu - 1] = 0 \tag{3.16}$$

Adding this term alters the local geometric shape of the functions to guarantee non-negativity *without* changing their macroscopic expected values. Our new tilted pointwise inequality becomes:

$$-\log \nu + \nu - 1 \geq C(\nu + \nu^{-1} - 2) \tag{3.17}$$

So:

$$\frac{-\log \nu + \nu - 1}{\nu + \nu^{-1} - 2} \geq C \tag{3.18}$$

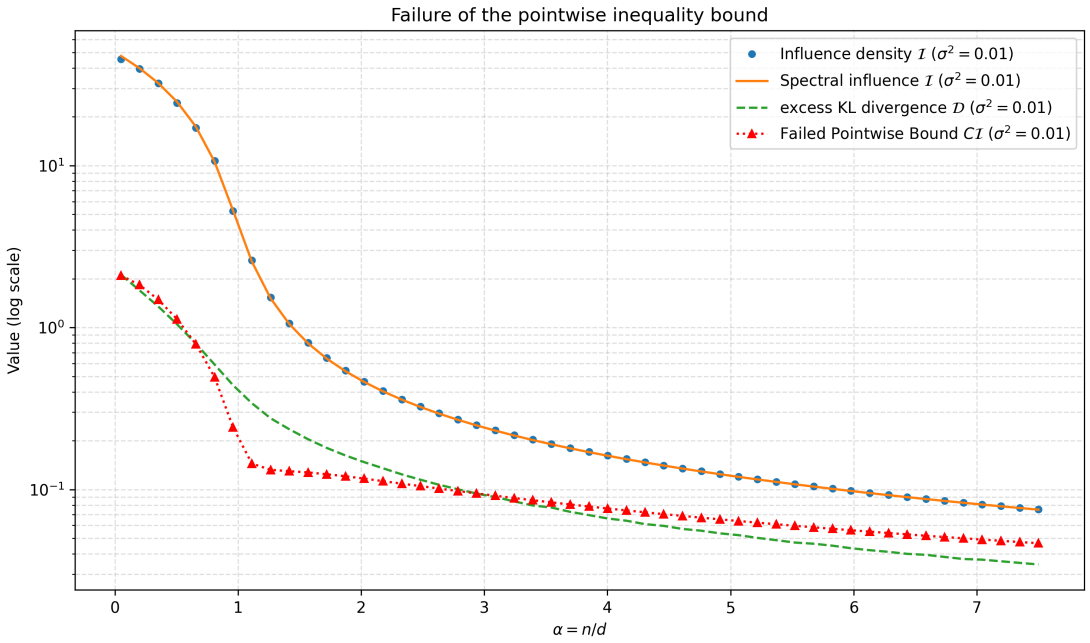


Figure 3.3: Failure of the inequality bound, working with the arguments of the expectations. The sign flip of the hypothesised lower bound goes above and below $\mathcal{D}$.

3.3.3 Monotonicity and the Constant $C(\sigma^2)$

Let us define our two functions:

$$f(\nu) = -\log \nu + \nu - 1 \qquad g(\nu) = \nu + \nu^{-1} - 2$$

By evaluating the ratio of these functions, we define $h(\nu)$:

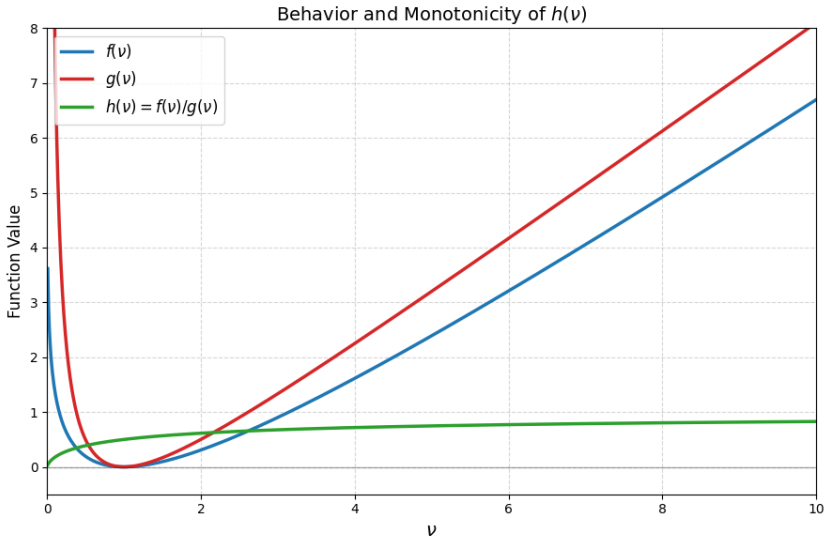
$$h(\nu) = \frac{f(\nu)}{g(\nu)} = \frac{\nu(\nu - 1 - \log \nu)}{(\nu - 1)^2}$$

Taking the derivative reveals that $h(\nu)$ is strictly monotonically increasing.

Therefore, its absolute minimum over the spectral support occurs exactly at the lower spectral edge, ν_{min}. By setting our constant $C \equiv h(\nu_{min})$, we guarantee that the pointwise inequality $f(\nu) \geq Cg(\nu)$ holds *almost surely* for all ν in the Marchenko-Pastur support.

To explicitly construct C, we track this minimum eigenvalue. The unregularized lower edge of the Marchenko-Pastur distribution is given by $\lambda_{min} = (1 - \sqrt{\gamma})^2$ if $\gamma < 1$, else 0. Under ridge regularization $\sigma^2 > 0$, our rescaled minimum eigenvalue becomes:

$$\nu_{min}(\gamma, \sigma^2) = \frac{\lambda_{min} + \sigma^2}{1 + \sigma^2}$$



In the over-parameterized regime ($\gamma \geq 1$), the unregularized edge hits exactly zero ($\lambda_{min} = 0$). However, thanks to the regularization term, our rescaled eigenvalue is strictly bounded away from zero:

$$\nu_{min}(\gamma \geq 1) = \frac{\sigma^2}{1+\sigma^2} > 0$$

It ensures the argument of the logarithm in our bound never diverges, providing a well-behaved, finite constant even when the system is heavily over-parameterized. Substituting ν_{min} into our ratio yields the exact theoretical constant:

$$C(\sigma^2) = \frac{\frac{\sigma^2}{1+\sigma^2}(\frac{\sigma^2}{1+\sigma^2} - 1 - \log(\frac{\sigma^2}{1+\sigma^2}))}{(\frac{\sigma^2}{1+\sigma^2} - 1)^2} = \sigma^2(1+\sigma^2)(\log(1+\frac{1}{\sigma^2}) - \frac{1}{1+\sigma^2}) \tag{3.19}$$

3.3.4 Applying the Expectation

We now invoke the **Monotonicity of the Lebesgue Integral**. Let $(\Omega, \mathcal{F}, P)$ be our spectral probability space. Since our functions $f(\nu)$ and $g(\nu)$ are integrable, and the inequality $f(\nu) \geq Cg(\nu)$ holds almost surely across the support, we can treat C purely as a deterministic constant dependent on σ^2 and safely factor it out of the expectation:

$$\mathbb{E}_{MP}[-\log\nu + \nu - 1] \geq C(\sigma^2)\mathbb{E}_{MP}[\nu + \nu^{-1} - 2] \tag{3.20}$$

Because the expectations of the $(\nu - 1)$ terms evaluate to zero by definition of the Marchenko-Pastur measure, they vanish entirely, leaving:

$$\mathbb{E}_{MP}[-\log\nu] \geq C(\sigma^2)\mathbb{E}_{MP}[\nu^{-1} - 1] \implies 2\mathcal{D} \geq 2C(\sigma^2)\mathcal{I} \tag{3.21}$$

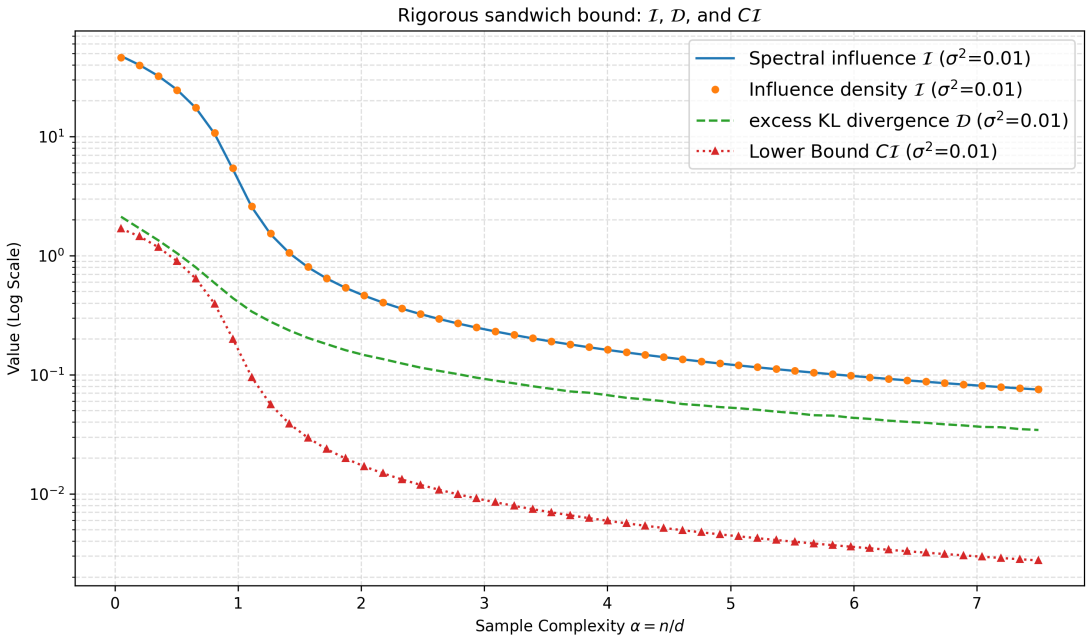


Figure 3.4: fixed lower bound using the average trick.

3.4 Asymptotic behaviour and Convergence

We validated this bound across different sample complexities. We first observed the empirical pointwise behaviour of the eigenvalues, and then extended this uniformly across γ without calculating the empirical distribution of the eigenvalues. This is possible because C depends only on the regularization σ^2.

Chapter 4

Link to the order parameters

Mathematics compares the most diverse phenomena
and discovers the secret analogies that unite them.
– Jean Baptiste Joseph Fourier

4.1 Connecting the Influence Density $\mathcal{I}$ to the Overlap Factor q

We now have two central quantities, and it is time to link them. The first one is the convergence overlap q, introduced as *cosine similarity* by Kadkhodaie et al. [1] and then formalized by Goldt & Maillard [2] to measure how similarly two independently trained models behave. We introduced this quantity in Section 1.4.1. The second one is the LOO influence density $\mathcal{I}$, which quantifies how sensitive the estimated covariance is to the removal of a single training point, which we derived in the preceding chapters of this thesis. Both are expectations with respect to the Marchenko-Pastur distribution of the rescaled eigenvalue

$$\nu \;=\; \frac{\lambda+\sigma^2}{1+\sigma^2}, \qquad \mathbb{E}_{\mathrm{MP}}[\nu]=1,$$

and are defined as

$$q \;=\; \big[\mathbb{E}_{\mathrm{MP}}(\sqrt{\nu})\big]^2, \qquad \mathcal{I} \;=\; \tfrac{1}{2}\big[\mathbb{E}_{\mathrm{MP}}(\nu^{-1})-1\big]. \tag{4.1}$$

Remember that these formulas are defined in the thermodynamic regime. In particular, $\mathcal{I}$ was derived in this work, while q is defined in [2]. While $q \in [0,1]$ saturates naturally between the memorisation limit ($q \to 0$) and the generalisation limit ($q \to 1$), the influence $\mathcal{I} \in [0,\infty)$ does not. This asymmetry makes a direct comparison awkward. A natural fix is to look for a bounded, monotone function of $\mathcal{I}$ that shares the same $0 \to 1$

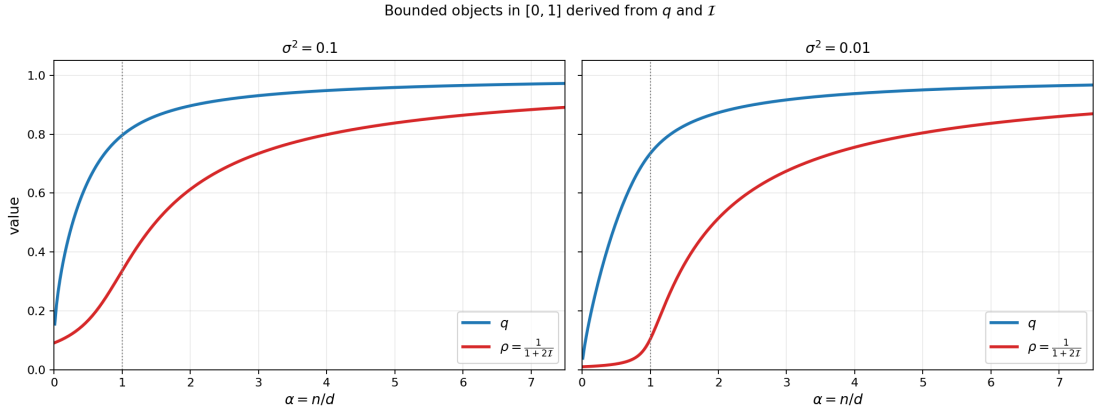


Figure 4.1: Comparison of q and $\rho = 1/(1 + 2\mathcal{I})$ as functions of α, for $\sigma^2 = 0.01$. Both live in $[0, 1]$ and saturate to 1 for large α, but they transition at different rates and decouple entirely for $\alpha < 1$.

saturation as q. Our ansatz[1] is

$$\rho \;:=\; \frac{1}{\mathbb{E}_{\mathrm{MP}}(\nu^{-1})} \;=\; \frac{1}{1 + 2\mathcal{I}} \;\in\; (0,\, 1], \tag{4.2}$$

which maps $[0, \infty)$ onto $(0, 1]$, with $\rho \to 0$ when $\mathcal{I}$ is large (few samples, high sensitivity) and $\rho \to 1$ when $\mathcal{I} \to 0$ (many samples, stable model) — the same direction as q (see Figure 4.1). The rest of this section studies how q and ρ are related, both through a general inequality and through asymptotic expansions in the two extreme regimes of $\alpha = n/d$. From now on, we write $\mathbb{E}$ instead of $\mathbb{E}_{\mathrm{MP}}$ throughout this chapter to lighten notation.

4.1.1 A general bound via Cauchy-Schwarz

Before entering the asymptotic regimes, one can establish a clean inequality between q and $\mathcal{I}$ that holds for all α and $\sigma^2 > 0$, which confirms once more what we found above. By the Cauchy-Schwarz inequality applied to $\nu^{1/4}$ and $\nu^{-1/4}$,

$$\mathbb{E}(\sqrt{\nu}) \cdot \mathbb{E}(\nu^{-1/2}) \;\geq\; \big[\mathbb{E}(1)\big]^2 = 1, \tag{4.3}$$

so $\mathbb{E}(\nu^{-1/2}) \geq 1/\sqrt{q}$. Since $x \mapsto x^2$ is convex, Jensen's inequality gives

$$\mathbb{E}(\nu^{-1}) \;\geq\; \big[\mathbb{E}(\nu^{-1/2})\big]^2 \;\geq\; \frac{1}{q}. \tag{4.4}$$

[1] This is an elegant way to bridge to the final result found rigorously as a bound. Other functions can work, like for example $\frac{1}{1+\mathcal{I}}$. For the sake of simplicity and theoretical symmetry, we include the factor 2 that naturally arises from the expectation $\mathbb{E}_{\mathrm{MP}}(\nu^{-1})$.

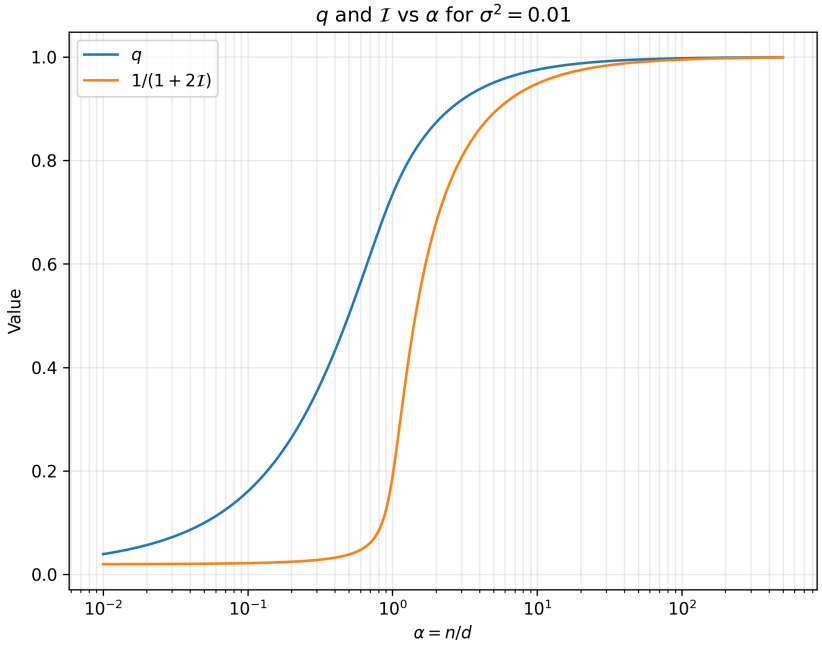


Figure 4.2: The Cauchy-Schwarz lower bound $1/(1+2\mathcal{I})$ compared to q, as functions of $\alpha = n/d$, for $\sigma^2 = 0.01$. The bound is non-trivial throughout and becomes tight as $\alpha \to 0$.

Recalling that $\mathbb{E}(\nu^{-1}) = 2\mathcal{I} + 1$, this becomes

$$q \;\geq\; \frac{1}{1+2\mathcal{I}}. \tag{4.5}$$

The bound is tight in the limit $\alpha \to 0$, as the small-α analysis below shows. Figure 4.2 plots q and its lower bound $1/(1+2\mathcal{I})$ as functions of α; the gap between them is visible throughout, and the two curves converge only for very small α (large γ).

4.1.2 The limit $\alpha \to +\infty$: Taylor expansion in $V = \mathrm{Var}(\nu)$

When $\alpha \to \infty$, the Marchenko-Pastur distribution concentrates around $\lambda = 1$, so $\nu \to 1$. So it is possible to write $\nu = 1 + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = V := \mathrm{Var}(\nu)$; the higher cumulants are $\mathcal{O}(V^{3/2})$ and can be neglected at this order.

Expanding $\sqrt{1+\varepsilon}$ to second order,

$$\mathbb{E}\big[\sqrt{1+\varepsilon}\big] = 1 - \frac{V}{8} + \mathcal{O}(V^{3/2}), \qquad q = \Big(1 - \frac{V}{8}\Big)^2 \approx 1 - \frac{V}{4}. \tag{4.6}$$

The same expansion for $(1+\varepsilon)^{-1}$ gives

$$\mathbb{E}\big[(1+\varepsilon)^{-1}\big] = 1 + V + \mathcal{O}(V^{3/2}), \qquad \mathcal{I} \approx \frac{V}{2}. \tag{4.7}$$

Putting the two together,

$$1 - q \;\approx\; \frac{V}{4}, \qquad 2\mathcal{I} \;\approx\; V, \tag{4.8}$$

so that our bounded influence ρ expands as:

$$\rho \;=\; \frac{1}{1+2\mathcal{I}} \;\approx\; \frac{1}{1+V} \;\approx\; 1 - V. \tag{4.9}$$

Consequently, we find:

$$1 - \rho \;\approx\; V \;\approx\; 4(1-q). \tag{4.10}$$

4.1.3 The limit $\alpha \to 0$: rigorous expansion of the Marchenko-Pastur measure

The $\alpha \to 0$ limit requires a subtler analysis, because the Marchenko-Pastur measure is no longer concentrated at a single point. For $\alpha < 1$, it consists of a point mass of weight $(1-\alpha)$ at $\lambda = 0$ (the zero modes) and a continuous bulk of weight α where eigenvalues scale as $\lambda \sim 1/\alpha \gg 1$, as explained in Appendix A.

To evaluate the integrals without the bulk domain escaping to infinity, we introduce the rescaled variable $x = \alpha\lambda$. The standard Marchenko-Pastur density $w(x)$ for this scaled variable is concentrated around $x = 1$. The expectation of any function $f(\nu)$ then splits exactly as:

$$\mathbb{E}_{\mathrm{MP}}[f(\nu)] \;=\; (1-\alpha) f(\nu(0)) + \alpha \int_{x_-}^{x_+} f\Big(\nu\Big(\frac{x}{\alpha}\Big)\Big) w(x)\, dx. \tag{4.11}$$

The influence density $\mathcal{I}$. For the inverse eigenvalue $\nu^{-1} = \frac{1+\sigma^2}{\lambda+\sigma^2}$, the exact zero-mode contribution is $(1-\alpha)\frac{1+\sigma^2}{\sigma^2}$. In the bulk integral, the integrand becomes $\frac{1+\sigma^2}{x/\alpha+\sigma^2} = \alpha\frac{1+\sigma^2}{x+\alpha\sigma^2}$. By Taylor-expanding the regular part for $\alpha \to 0$, we get:

$$\frac{1}{x+\alpha\sigma^2} \;=\; \frac{1}{x} - \alpha\frac{\sigma^2}{x^2} + \mathcal{O}(\alpha^2). \tag{4.12}$$

Integrating this against $w(x)$ requires the negative moments of the standard Marchenko-Pastur distribution, notably $\mathbb{E}_w[x^{-1}] = \frac{1}{1-\alpha} \approx 1+\alpha$. Multiplying by the outer weight α, the bulk contribution scales as $\alpha^2(1+\sigma^2)$. Adding this to the zero-mode term recovers the full second-order expansion:

$$\mathbb{E}_{\mathrm{MP}}[\nu^{-1}] \;\approx\; \frac{1+\sigma^2}{\sigma^2} - \alpha\frac{1+\sigma^2}{\sigma^2} + \alpha^2(1+\sigma^2). \tag{4.13}$$

Recalling $\mathcal{I} = \frac{1}{2}[\mathbb{E}_{\mathrm{MP}}(\nu^{-1}) - 1]$, we obtain:

$$\mathcal{I} \;\approx\; \frac{1}{2\sigma^2} - \alpha\frac{1+\sigma^2}{2\sigma^2} + \mathcal{O}(\alpha^2), \qquad \rho \;=\; \frac{1}{1+2\mathcal{I}} \;\approx\; \frac{\sigma^2}{1+\sigma^2}. \tag{4.14}$$

From this derivation we can infer that ρ is essentially constant in α in this regime — it is determined entirely by the regularisation σ^2, not by the amount of data.

The overlap factor q. A similar rigorous expansion applies to $\sqrt{\nu}$. The zero modes yield a small contribution proportional to σ. For the bulk, the rescaled integrand isolates the singular dominant amplitude analytically, allowing a safe Taylor expansion of the remainder:

$$\sqrt{\nu\Big(\frac{x}{\alpha}\Big)} \;=\; \frac{1}{\sqrt{\alpha(1+\sigma^2)}}\sqrt{x+\alpha\sigma^2} \;\approx\; \frac{1}{\sqrt{\alpha(1+\sigma^2)}}\Big(\sqrt{x}+\alpha\frac{\sigma^2}{2\sqrt{x}}\Big). \tag{4.15}$$

Integrating this with respect to $w(x)$ uses the moments $\mathbb{E}_w[\sqrt{x}] \approx 1-\frac{\alpha}{8}$ and $\mathbb{E}_w[x^{-1/2}] \approx 1+\frac{3\alpha}{8}$. Combining this bulk integral (multiplied by the weight α) with the zero-mode trace yields the total expectation:

$$\mathbb{E}_{\mathrm{MP}}[\sqrt{\nu}] \;\approx\; \frac{1}{\sqrt{1+\sigma^2}}\Big[\sigma(1-\alpha)+\sqrt{\alpha}\Big(1+\alpha\Big(\frac{\sigma^2}{2}-\frac{1}{8}\Big)\Big)\Big]. \tag{4.16}$$

Squaring this expectation to compute q reveals a cross-term:

$$q \;\approx\; \frac{1}{1+\sigma^2}\Big[\sigma^2+2\sigma\sqrt{\alpha}+\alpha(1-2\sigma^2)+\mathcal{O}(\alpha^{3/2})\Big]. \tag{4.17}$$

This expansion shows that q does not merely vanish linearly. The $\sqrt{\alpha}$ term emerges directly from the interference between the zero-mode trace ($\sim \sigma$) and the strong bulk signal ($\sim \sqrt{\alpha}$). Thus, the two quantities completely decouple: q tracks this non-trivial interplay, while $\mathcal{I}$ remains strictly clamped by the zero modes and the regularisation floor. Figure 4.3 shows both limiting behaviours.

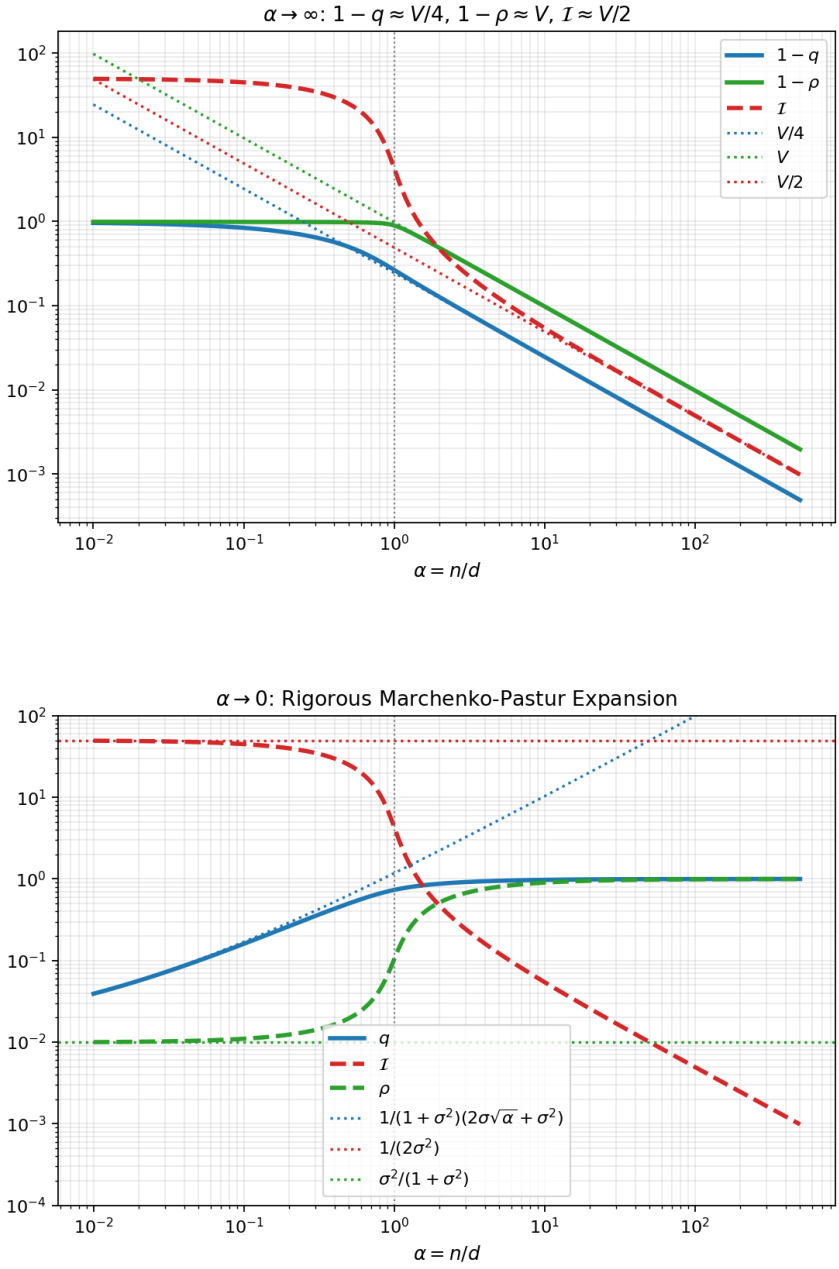


Figure 4.3: Asymptotic behaviour of q, $\mathcal{I}$, and their Taylor predictions as functions of α, for $\sigma^2 = 0.01$. *Up:* for large α, both $1-q$ and $\mathcal{I}$ scale as $V = \mathrm{Var}(\nu)$, leading to $1-\rho \approx 4(1-q)$. *Down*: for small α, q vanishes tracking the learned directions with a $\sqrt{\alpha}$ correction, while $\mathcal{I} \approx 1/(2\sigma^2)$ (set by the zero modes) — the two quantities become completely independent of each other.

Chapter 5

Conclusion

Don't cry because it ended, smile because it happened.

This thesis set out to answer a specific question: can the generalization gap of a diffusion model – measured through the excess Kullback-Leibler divergence $\mathcal{D}$ between the learnt and true distribution—be bounded using information available from a single trained model, rather than requiring the comparison of two independently trained models as in the KGSM experiment?

We answered this affirmatively within a tractable Gaussian framework, using tools from statistical learning and statistical mechanics. Starting from a Leave-One-Out influence function Δ_i, derived in closed form from the structure of the (regularized) sample covariance matrix via matrix algebraic identities, we showed in Chapter 2 that its thermodynamic limit $\mathcal{I}$ admits an explicit expression in terms of the Marchenko-Pastur spectral measure. In Chapter 3, we proved our our central result, $\mathcal{D} \leq \mathcal{I}$, for every value of regularization σ^2, and complemented it with a matching lower bound obtained via a pointwise inequality, establishing that $\mathcal{D}$ and $\mathcal{I}$ are of the same order across the whole range of sample complexities. Finally, in Chapter 4 we connected $\mathcal{I}$ to the convergence overlap q of Maillard and Goldt [2], deriving rigorous asymptotic expansions in both the $\alpha \to \infty$ and $\alpha \to 0$ regimes. These expansions revealed a qualitative distinction between the two quantities: for large sample complexity, $1 - q$ and $\mathcal{I}$ are both governed by the variance of the rescaled spectrum and scale proportionally; for small sample complexity, however, the two quantities decouple entirely—$\mathcal{I}$ saturates at a value fixed purely by the regularization strength σ^2, while q continues to track the geometric alignment between independently trained models through a non-trivial $\sqrt{\alpha}$ correction.

On the choice of framework

The Gaussian, isotropic, linear-denoiser setting adopted throughout this thesis was a voluntary simplification, not merely a matter of tractability. It is precisely this simplicity that allowed the central bound $\mathcal{D} \leq \mathcal{I}$ to be derived exactly, every step in Chapters 2 and

3 admits a closed form that can be checked, term by term, against numerical simulation. This strategy was indeed carried out by Maillard and Goldt [2], who used the same class of models[1] to give a rigorous foundation to the *memorization-to-generalization* transition that Kadkhodaie et al. [1] had observed empirically in a UNet trained on real face images. In both cases, the tractable model is a controlled setting in which a mechanism can first be isolated and proven, before being tested against more realistic architectures.

The specific motivation for studying an influence function follows the same logic. Computing q requires training two independent models to completion for every configuration of interest (dataset size, regularization, architecture)—a cost that scales linearly with the number of settings one wishes to examine, and that offers no signal until both models have finished training. On the other hand, the influence density $\mathcal{I}$ is computable from a single trained model, using only the empirical covariance the model has already fit. If $\mathcal{I}$ can be shown to track $\mathcal{D}$ as reliably as q does, it offers a substantially cheaper diagnostic—one that could, in principle, be monitored during training rather than only after the fact.

On the validity of $\mathcal{I}$ as a proxy

The bound $\mathcal{D} \leq \mathcal{I}$ was proven rigorously only for an isotropic target covariance $\Sigma^\star = \mathbb{I}_d$; Maillard and Goldt's original framework—and the richer phenomenology of memorization, convergence, and latent recovery it describes—involves a low-rank spike $\Sigma^\star = \mathbb{I}_d + \beta uu^\top$, and extending the present bound to that setting remains open, even though, having a very similar structure, we do not expect major changes.

In addiction, this derivation relies throughout on a linear denoiser, the simplest possible generative model, and on the thermodynamic limit $n, d \to \infty$ at fixed $\gamma = d/n$; the finite-size behaviour of $\mathcal{I}$, (Section 3.3), where the pointwise inequality underlying the lower bound could in principle be violated by fluctuating empirical eigenvalues, has not been rigorously controlled. Within these restrictions, however, the bound is exact and the numerical agreement with the theoretical predictions is very tight throughout every regime tested. Whether $\mathcal{I}$ remains a faithful proxy for generalization once these restrictions are lifted—spiked covariance, non-linear denoisers, finite n and d—is the natural next question, and one that would require empirical validation on models closer to those used in practice, beyond the analytical and simulation-based checks performed here.

Future directions

The most immediate extension is to repeat the derivation of Chapters 2 and 3 for the spiked covariance model of Maillard and Goldt, to determine whether $\mathcal{D} \leq \mathcal{I}$ survives the presence of a low-rank signal direction, and how the bound interacts with the third order parameter Q governing latent recovery. A second, more structural extension is to replace the linear denoiser with a **random features model**. Random features retain

[1] Apart from $\mathcal{O}(1)$ spikes, which are not relevant in the calculation of $\mathcal{D}$.

enough analytical structure to make a closed-form treatment plausible, while allowing the denoiser to implement a non-linear function of its input—and they introduce a notion of over-parametrization, cited in, Section 1.3, that is entirely absent from the present linear setting, where the number of effective parameters coincides with d by construction.

Finally, the comparison between models trained on disjoint data subsets—the core idea of the KGSM experiment—has a natural counterpart in classical statistical learning theory. Vapnik and Chervonenkis showed that a learning algorithm whose output changes little under the removal of a single training point enjoys a small generalization error [19]. The KGSM test can be read as a more extreme instance of the same principle: the entire training set is replaced by an independent draw, and one asks whether the learned function is left unchanged—that is, whether it is a function of the underlying data distribution rather than of the particular sample. The results of this thesis suggest that the more classical, and considerably cheaper, notion of leave-one-out stability may serve as a workable proxy for this stronger notion of distributional stability. Making this connection precise—relating $\mathcal{I}$ not only to q but directly to a stability-based generalization bound in the sense of Vapnik and Chervonenkis—is left as a direction for future work.

Bibliography

[1] Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stephane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations (ICLR)*, 2024.

[2] Antoine Maillard and Sebastian Goldt. Memorisation, convergence and generalisation in generative models. *arXiv preprint arXiv:2605.21402*, 2026.

[3] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, et al. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2023.

[4] Preetum Nakkiran, Arwen Bradley, Hattie Zhou, and Madhu Advani. Step-by-step diffusion: An elementary tutorial, 2024.

[5] https://cvpr2022-tutorial-diffusion-models.github.io/.

[6] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

[7] Pierre Marion and Yu-Han Wu. Understanding diffusion models requires rethinking (again) generalization. arXiv preprint arXiv:2605.06077, 2026.

[8] Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

[9] Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. *arXiv preprint arXiv:2505.17638*, 2025. NeurIPS 2025.

[10] Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models. *arXiv preprint arXiv:2505.16959*, 2025.

[11] Yu-Han Wu, Pierre Marion, Gérard Biau, and Claire Boyer. Taking a big step: Large learning rates in denoising score matching prevent memorization. *arXiv preprint arXiv:2502.03435*, 2025.

[12] Moritz Hardt and Benjamin Recht. *Patterns, Predictions, and Actions: Foundations of Machine Learning*. Princeton University Press, 2022.

[13] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1986.

[14] Richard von Mises. On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18(3):309–348, 1947.

[15] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning, with Applications in R*. Springer, 2nd edition, 2023. Corrected printing, June 2023.

[16] Kaare Brandt Petersen and Michael Syskind Pedersen. *The Matrix Cookbook*. Technical University of Denmark, version 2012-11-15 edition, 2012.

[17] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *arXiv preprint arXiv:1306.2872*, 2013.

[18] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

[19] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[20] G. Livan, M. Novaes, , and P. Vivo. *Introduction to Random Matrices: Theory and Practice*. Springer, 2018.

[21] Oriol Bohigas and Marie-Joya Giannoni. Chaotic motion and random matrix theories. 1984.

[22] Hugh L. Montgomery. The pair correlation of zeros of the zeta function. In *Analytic number theory*, volume XXIV of *Proc. Sympos. Pure Math.* American Mathematical Society, 1973.

[23] Emre Telatar. Capacity of multi-antenna gaussian channels. *European Transactions on Telecommunications*.

[24] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* ***1****(4)*, 1967.

[25] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, New York, NY, 2nd edition, 2010.

Appendix A

Random Matrix Theory

A.1 Random Matrices in Physics: some History

The story of random matrix theory (RMT) begins in *Nuclear Physics*. In the early 1950s, Eugene Wigner found that the Hamiltonians of heavy nuclei such as uranium are $N \times N$ matrices with N on the order of hundreds or thousands, their entries determined by complicated quantum-mechanical interactions that no one could compute from first principles. Wigner thought it was time to leave the idea of knowing the matrix and instead model it by a random matrix drawn from a suitable ensemble. Under this ansatz, observable statistical properties of nuclear energy levels – eigenvalues of the Hamiltonian, for example– could be predicted without ever knowing the underlying matrix.

The success was striking and new spectral properties were studied. Wigner's *semicircle law* (Figure A.1) states that, for a large real symmetric matrix H of size $N \times N$ whose off-diagonal entries are independent with mean zero and variance $1/N$, the empirical spectral distribution converges in probability to the semicircle distribution

$$\rho_{\mathrm{SC}}(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \qquad x \in [-2,\, 2].$$

It is important to define some typical examples of random matrices, which are classified based on their dimension and the type of random entries.

- **Gaussian Ensemble**: The three main examples are the Gaussian orthogonal (GOE), unitary (GUE), and symplectic (GSE) ensembles. These are classified by the Dyson index β, which takes values 1, 2, and 4 respectively.
- **Wishart-Laguerre Ensemble**: in our case, the Wishart ensemble will be useful, which will be covered later in the next section to analyse the covariance matrix.

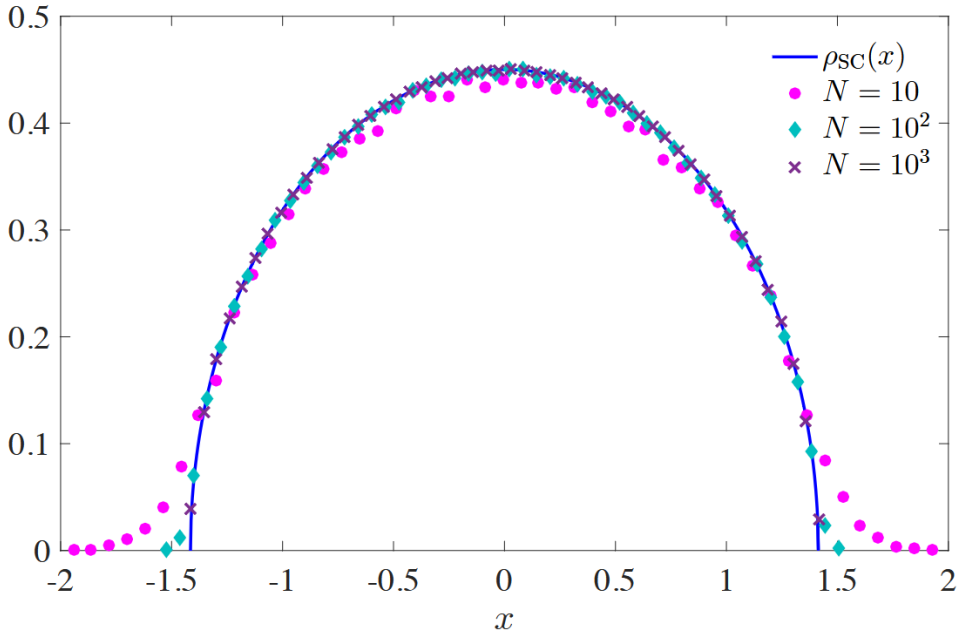


Figure A.1: Wigner's semicircle: numerical proof that increasing N brings to convergence towards the semicircle. [20]

Beyond nuclear physics, random matrices subsequently appeared in:

- **Quantum chaos [21]**: the eigenvalue statistics of classically chaotic quantum systems coincide with those of the Gaussian Orthogonal Ensemble.
- **Number theory [22]**: the pair-correlation of the zeros of the Riemann zeta function matches the GOE eigenvalue statistics, a connection that remains mysterious to this day.
- **Wireless communications and information theory [23]** : the capacity of certain types of channels is determined by the singular values of large random matrices.
- **Finance and data science**: the spectrum of empirical correlation matrices of asset returns is governed by the Marchenko–Pastur law, separating signal from noise.

The present chapter will focus on the **Marchenko–Pastur (MP) law**, the analogue of the semicircle law for *rectangular* random matrices and sample covariance matrices. Proved by the Ukrainian mathematicians Volodymyr Marchenko and Leonid Pastur in 1967 [24], it has become a staple argument in statistics, machine learning, and signal processing wherever one encounters a large data matrix whose dimensions are of comparable size.

A.2 Random Samples and the Sample Covariance Matrix

Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ be independent random vectors drawn from a population with mean 0 and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The *sample covariance matrix* is

$$S_n \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\top} \;=\; \frac{1}{n} X X^{\top},$$

where $X = [\mathbf{x}_1 \mid \cdots \mid \mathbf{x}_n] \in \mathbb{R}^{d\times n}$. When $\Sigma = \sigma^2 \mathbb{I}_d$, the columns of X are i.i.d. isotropic Gaussian vectors (after appropriate scaling), and S_n belongs to the **Wishart–Laguerre** ensemble.

Definition A.2.1 (Wishart matrix). Let $X \in \mathbb{R}^{d\times n}$ have i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. The matrix $W = XX^{\top}$ is called a *real Wishart matrix* with parameters (d, n, σ^2), written $W \sim \mathcal{W}_d(n, \sigma^2 \mathbb{I})$.

It is interesting to study how the eigenvalues of W are distributed. By classical results the joint density[1] of the ordered eigenvalues $0 \le \lambda_1 \le \cdots \le \lambda_d$ of $\frac{1}{\sigma^2} W$ is proportional to

$$\prod_{i=1}^{d} \lambda_i^{\alpha/2}\, e^{-\lambda_i/2} \cdot \prod_{i<j} |\lambda_j - \lambda_i|^{\beta},$$

where $\alpha = \beta(n - d + 1) - 2$ encodes the shape parameter and $\beta = 1$ for real matrices [20]. Here comes the question we wonder to answer: *What does the empirical spectral distribution*

$$\hat{\mu}_n \;=\; \frac{1}{d}\sum_{i=1}^{d} \delta_{\lambda_i(S_n)}$$

look like as $d, n \to \infty$*?*

[1] The Coulomb-gas analogy—interpreting eigenvalues as charged particles in a confining potential on the positive half-line—provides the physical intuition: the $\prod |\lambda_j - \lambda_i|$ repulsion term prevents eigenvalue clustering, and together with the exponential confining potential and the hard wall at the origin (eigenvalues must remain non-negative), it shapes the limiting spectrum. [20]

A.3 The Marchenko–Pastur Law

Theorem A.3.1 (Marchenko–Pastur, 1967 [24])**.** *Let $X \in \mathbb{R}^{d\times n}$ have i.i.d. entries with mean 0 and variance σ^2. Assume $d, n \to \infty$ with*

$$\gamma \;:=\; \frac{d}{n} \;\longrightarrow\; \gamma_\infty \in (0,\infty).$$

Then the empirical spectral distribution of $S_n = \frac{1}{n} XX^\top$ converges almost surely (in the weak sense) to the Marchenko–Pastur distribution μ_{MP} *with density*

$$f_\gamma(x) \;=\; \frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{2\pi\,\sigma^2\,\gamma\,x}\;\cdot\;\mathbf{1}_{[\lambda_-,\,\lambda_+]}(x) \;+\; \Big(1-\frac{1}{\gamma}\Big)^{+}\delta_0, \tag{A.1}$$

where $(\cdot)^+ = \max(\cdot, 0)$ and the edge points are

$$\lambda_\pm \;=\; \sigma^2\Big(1 \pm \sqrt{\gamma}\Big)^2. \tag{A.2}$$

The point mass at zero. When $\gamma > 1$ (i.e., $d > n$), the matrix S_n has rank at most $n < d$, so at least $d - n$ eigenvalues are exactly zero. In the limit this manifests as a Dirac mass of weight $1 - 1/\gamma$ at the origin.

Support width. The bulk spectrum is confined to the compact interval $[\lambda_-, \lambda_+]$ whose width is $\lambda_+ - \lambda_- = 4\sigma^2\sqrt{\gamma}$ and grows with $\sqrt{\gamma}$ (see Figure A.3). As $\gamma \to 0$, both edges converge to σ^2 and the distribution narrows to a Dirac mass. As $\gamma \to 1$ (square matrix), $\lambda_- \to 0$ and the density diverges as $x^{-1/2}$ near the origin, producing a hard-edge singularity at zero. Figure A.2 shows these shapes for several values of γ.

Moments. The k-th moment of the MP distribution (for $\gamma \le 1$, so there is no atom at zero) is the *Narayana polynomial* evaluated at γ [24]:

$$m_k \;=\; \int x^k \,\mathrm{d}\mu_{\mathrm{MP}}(x) \;=\; \sum_{j=1}^{k} \frac{1}{k}\binom{k}{j}\binom{k}{j-1}\gamma^j,$$

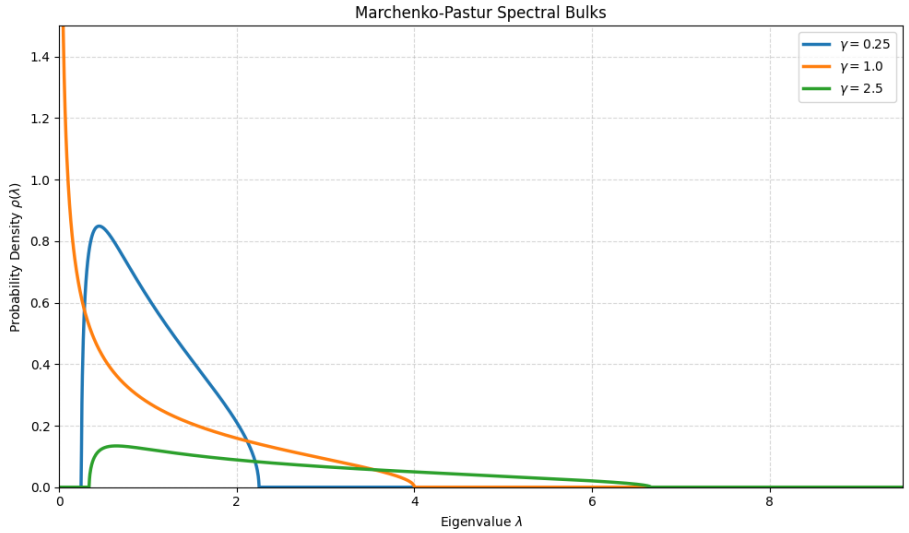


Figure A.2: The Marchenko–Pastur density f_γ for three values of the aspect ratio $\gamma = d/n$, with $\sigma^2 = 1$. As $\gamma \to 0$ the distribution concentrates near $\sigma^2 = 1$. For $\gamma = 1$ (orange solid curve) the density acquires a $x^{-1/2}$ singularity at the origin.

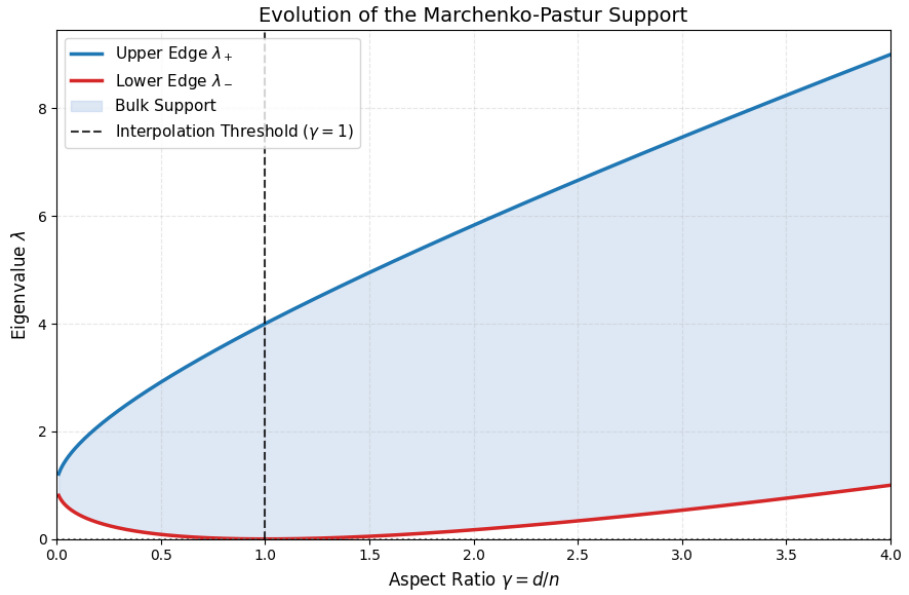


Figure A.3: The edge points $\lambda_\pm = \sigma^2(1 \pm \sqrt{\gamma})^2$ as functions of the aspect ratio γ, with $\sigma^2 = 1$. The shaded region is the support of the MP distribution. For $\gamma > 1$ the lower edge λ_- moves away from zero again.

A.4 Derivation via the Stieltjes Transform

We sketch the elegant *resolvent* (or Stieltjes transform) approach, as presented in [20] and [25].

A.4.1 The Stieltjes Transform in RMT

In the context of Random Matrix Theory (RMT), the Stieltjes transform is directly linked to the resolvent of a matrix, making it a powerful tool for analyzing spectral properties. Let A be an $n \times n$ Hermitian matrix and F_n be its empirical spectral distribution (ESD). The Stieltjes transform of F_n is given by

$$s_n(z) = \int \frac{1}{x - z} \, \mathrm{d}F_n(x) = \frac{1}{n} \operatorname{Tr}(A - zI)^{-1}.$$

A key advantage of this formulation is that it allows the use of algebraic matrix identities to recursively analyze the spectrum. Specifically, the Stieltjes transform can be expressed as:

$$s_n(z) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{a_{kk} - z - \alpha_k^*(A_k - zI)^{-1}\alpha_k}$$

where A_k is the $(n-1) \times (n-1)$ matrix obtained from A with the k-th row and column removed, and α_k is the k-th column vector of A with the k-th element removed.

This decomposition isolates the contribution of the k-th row and column from the rest of the matrix. It forms the foundational algebraic step we will briefly show in the derivation of the Marchenko-Pastur law.

A.4.2 Convergence to the Marchenko-Pastur Law: Proof Overview

Following this approach, covered in [25], the proof that the ESD of the sample covariance matrix converges to the Marchenko-Pastur (MP) law relies on proving the convergence of their respective Stieltjes transforms.

Let $s_n(z)$ be the Stieltjes transform of the ESD of the sample covariance matrix S_n. The proof proceeds in this way:

Deriving the Stieltjes transform of the MP law. First of all, the exact analytic form of the Stieltjes transform of the theoretical Marchenko-Pastur distribution, denoted as $s(z)$, must be established. Using complex analysis and residues analysis, it is shown that $s(z)$ satisfies a specific quadratic equation. For $\gamma < 1$, the explicit formula is:

$$s(z) = \frac{\sigma^2(1 - \gamma) - z + \sqrt{(z - \sigma^2 - \gamma\sigma^2)^2 - 4\gamma\sigma^4}}{2\gamma z\sigma^2}.$$

Almost sure convergence of the random part. The proof then evaluates the empirical Stieltjes transform $s_n(z)$ by first showing that it concentrates around its expected value. It must be proven that for any fixed $z \in \mathbb{C}^+$, the difference $s_n(z) - \mathbb{E}[s_n(z)] \to 0$ almost surely. This is accomplished by expressing the difference as a sum of bounded martingale differences, applying bounds to these differences and utilizing the Borel-Cantelli lemma.

Mean convergence to the MP transform. The final step is proving that the expected Stieltjes transform, $\mathbb{E}[s_n(z)]$, converges to the theoretical Stieltjes transform $s(z)$ derived in the first step. THe expected value is defined algebraically, and the solution brings to only one solution with positive imaginary part, which perfectly matches the Stieltjes transform of the MP law.

Matching the last two steps together This way, it is established that $s_n(z) \to s(z)$ almost surely for every $z \in \mathbb{C}^+$. The final considerations deal with the type of convergence: the pointwise convergence of Stieltjes transforms in the upper-half complex plane implies the weak convergence of their corresponding probability measures, this rigorously demonstrates that the spectral density of the sample covariance matrix converges to the Marchenko-Pastur law.

A.5 Empirical Verification

Figure A.4 compares the theoretical MP density (red curve) against empirical histograms obtained by diagonalising 3 000 independent Wishart matrices of size $N = 200$. For both $\gamma = 0.25$ and $\gamma = 0.5$ the agreement is visually perfect even at this moderate matrix size. The vertical dashed lines mark the predicted edge points $\lambda_\pm$, and no empirical eigenvalue escapes the support.

A.6 Low Rank Perturbations: the BBP transition

Up to now, we have only dealt with very simple covariance matrices. But what if the structure changes, with specific linear information via a low-rank perturbation? What happens to the bulk of the eigenvalues? The BBP phase transition addresses this in the *spiked covariance model*, where the population covariance matrix is $\Sigma = \mathbb{I}_d + \beta \mathbf{u}\mathbf{u}^T$.

Here, $\beta > 0$ represents the *signal-to-noise ratio* (SNR): the ratio of the signal variance (along the spike direction u) to the noise variance in all other directions. When observing the eigenvalue histogram of the sample covariance matrix S_n, a single eigenvalue can pop out from the bulk support of the Marčenko–Pastur distribution (See Figure A.5).

Whether this separation occurs depends on how β compares to the dimensionality ratio $\gamma = d/n$. It can be solved for the leading eigenvalue of the rank-one perturbation and taking the large-d limit. Evaluating this at $\lambda = \lambda_+$ gives the critical value

$$\beta_c = \sqrt{\gamma} = \sqrt{d/n}$$

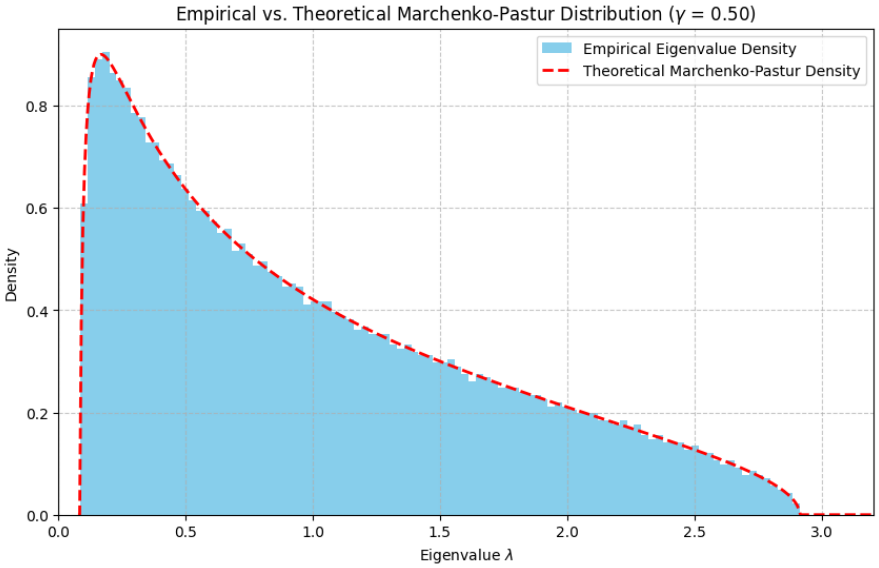


Figure A.4: Empirical spectral histograms (blue) of 3000 Wishart matrices of size $N = 200$ versus the Marchenko–Pastur density (red), for $\gamma = 0.25$ (left) and $\gamma = 0.50$ (right). Dashed lines indicate the theoretical edge points $\lambda_{\pm}$. The agreement confirms the convergence predicted by Theorem A.3.1.

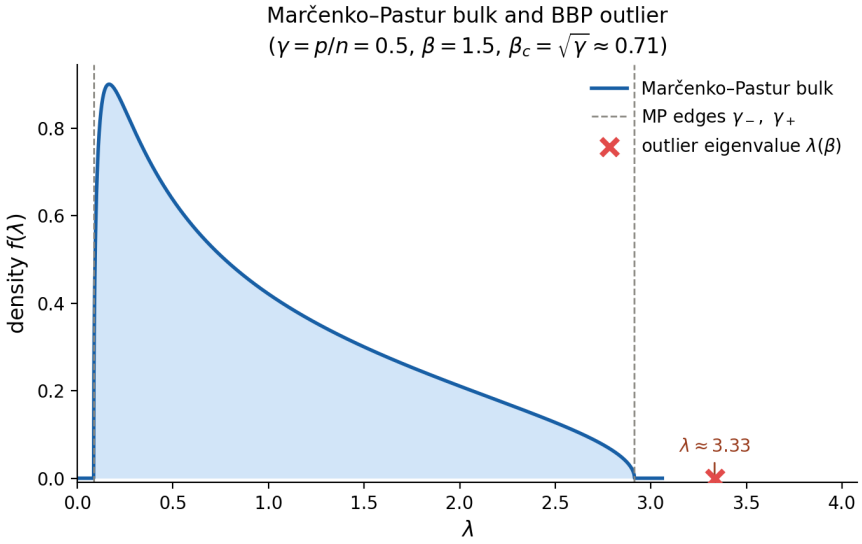


Figure A.5: An outlier is visible in this density. An Idea could be that of thinking how this is visible changing some parameters, like for example β.

Table A.1: Parameters of the Marchenko–Pastur distribution MP(γ, σ^2).

Parameter	**Expression**
Aspect ratio	$\gamma = d/n$
Support	$[\lambda_-, \lambda_+]$, where $\lambda_\pm = \sigma^2(1 \pm \sqrt{\gamma})^2$
Atom at 0	weight $\max(0,\, 1 - 1/\gamma)$ when $\gamma > 1$
Mean	σ^2
Variance	$\sigma^4\gamma$
Skewness	$\sqrt{\gamma}$
Excess kurtosis	$\gamma + 2$
R-transform	$R(z) = \sigma^2\gamma/(1 - \sigma^2 z)$
Stieltjes transform	solution of (A.1)
BBP critical SNR	$\beta_c = \sigma^2\sqrt{\gamma}$

Therefore, when $\beta \leq \sqrt{d/n}$, the spike remains hidden inside the bulk, and the largest eigenvalue of S_n simply converges to the M-P upper edge, $(1 + \sqrt{\gamma})^2$. However, once $\beta > \sqrt{d/n}$, the top eigenvalue separates from the bulk and converges instead to $(\beta + 1)(1 + \gamma/\beta)$.

Conditions such as a low SNR, a small sample size n, and high dimensionality d all raise the critical threshold β_c, making the spike harder to detect.

A.7 Open Problems

Table A.1 collects the key parameters of the MP distribution for reference.

The Marchenko–Pastur law is essentially the sample-covariance analogue of the Central Limit Theorem (CLT): where the CLT governs sums of independent scalars, MP governs the limiting spectrum of sums of independent rank-one matrices.

There are still open questions. The finite-n fluctuations at the edge (Tracy–Widom) are well understood in the Gaussian case but less so otherwise. Universality — whether the same limiting behaviour holds for non-Gaussian entries — is an active area. So are extensions to correlated data: Wigner-type matrices, and spiked covariance models with many spikes instead of just a few.