An Information-Geometric View of the Platonic Representation Hypothesis

Alexander Lobashev LOBASHEVALEXANDER@GMAIL.COM
Glam AI, San Francisco, US

Editors: List of editors' names

Abstract

The Platonic Representation Hypothesis suggests that diverse, large-scale neural networks trained on similar data learn aligned internal representations. This work provides a theoretical justification for this phenomenon from an information-geometric and Bayesian perspective. We demonstrate that representation alignment is a direct consequence of posterior concentration in Bayesian learning. As the dataset size grows, the posterior distribution over model parameters concentrates on those that best approximate the true data distribution. For sufficiently expressive models (i.e., large capacity), this forces them to learn the same underlying function, resulting in aligned representations. We formalize this convergence and also prove a "disunion theorem", showing that models with different approximation capabilities will learn provably distinct representations, with a separation that grows exponentially with dataset size.

Keywords: Platonic Representation Hypothesis, Information Geometry, Bayesian Inference, Posterior Concentration, Representation Learning

1. Introduction

The Platonic Representation Hypothesis (Huh et al., 2024) refers to the empirically observed alignment of internal representations learned by different models with varying architectures or even trained on different data modalities. A key observation is that this representational alignment improves as model and dataset sizes grow. Conversely, when the connection between data and labels is random, the learned representations are misaligned (rokosbasilisk, 2024), suggesting that the existence of meaningful structure of the data distribution is needed.

While prior theoretical work has explored this hypothesis in simplified settings, such as linear networks, this paper examines it from a more general, information-geometric viewpoint. We frame machine learning models as parametric probability distributions and use the tools of Bayesian inference to explain why and when alignment should occur.

Our argument rests on a foundational result in Bayesian asymptotics: in the large data limit ($N \to \infty$), the posterior distribution over model parameters concentrates on the set of parameters that minimize the Kullback-Leibler (KL) divergence to the true data distribution (Berk, 1966).

2. Related Works

Kornblith et al. (2019) observed that the similarity between learned representations for different initializations grows with the model width and that later layers learn more similar representations. Later Imani et al. (2022) conducted more detailed study on alignment in deep CNN with VGG and ResNet architectures, further proving and strengthen the claim of Kornblith et al. (2019). Next, the Platonic Representation Hypothesis was proposed by Huh et al. (2024), who extended the prior idea of alignment to different modalities, providing empirical evidence for alignment between models trained separately on text and image datasets.

However, these studies considers mostly experimental results, which has led to research on a theoretical explanation for this phenomenon. For example, Ziyin and Chuang (2025) studies the convergence of representations in deep linear network without activation functions and prove existence of unique optimal representation, which could be reached during stochastic gradient descent optimization.

Contribution Our work takes a different approach by grounding the hypothesis in the principles of Bayesian statistics and information geometry (Amari and Nagaoka, 2000). The core mechanism we leverage – posterior concentration – is a classical result. Our main theoretical tool, Theorem 5, is a well-known result in Bayesian asymptotics, closely related to the seminal work of Berk (1966). Berk's theorem established that, even when a statistical model is misspecified (i.e., the true distribution is not in the model class), the posterior distribution still concentrates on the set of parameters that are closest to the true distribution in terms of KL divergence. Our analysis applies this powerful principle to the space of learned representations, providing a general explanation for alignment that depends not on a specific architecture, but on the model's capacity to approximate the true data distribution.

3. The Asymptotic Behavior of the Posterior

Our analysis begins with the theorem of Bayesian asymptotics, which describes how a posterior distribution on model weights evolves as we observe more data. The following theorem generalizes the proof of Theorem 3.1 by Lobashev et al. (2025) from exponential family distributions to general conditional probability measures $p(x|t)$, where t are the model parameters. It also resembles the results of Berk (1966) on the concentration of posterior distributions, adding exponential decay as a model distribution stays further from the real data. The complete proofs and proof sketches are given in the Appendix.

Theorem 1 (Large-N Posterior Limit for General Models) *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measurable space. For each parameter t in a compact set $S \subset \mathbb{R}^n$, let $p(x \mid t)$ be a probability density. Assume the map $(t, x) \mapsto \log p(x \mid t)$ is continuous in t and uniformly continuous in x. Assume a continuous prior $p(t) > 0$ on S. Let $x_1, \ldots, x_N$ be i.i.d. samples from $p(\cdot \mid t')$ for some true parameter $t' \in S$. Then, almost surely,*

$$\lim_{N\to\infty} \left(p(t \mid x_1, \ldots, x_N)\right)^{1/N} = \exp\left(-D_{\mathrm{KL}}(p(\cdot \mid t') \| p(\cdot \mid t))\right). \tag{1}$$

This could be seen as an extension of a classic result in Bayesian asymptotics (Berk, 1966). We assume that true data distribution $p(x|t')$ lies in the same class of distributions as

our model and several different parameters t of the model could produce the same generated distribution $p(x|t)$. The later property holds for deep generative models, due to symmetries in their parameter spaces.

In simple terms, the theorem states that if the distribution $p(\cdot \mid t)$ is distinguishable from the true data distribution $p(\cdot \mid t')$, the posterior probability of such t becomes exponentially small as the number of data samples (N) grows. The rate of this decay is precisely the KL divergence. Consequently, the posterior mass concentrates entirely on the parameter t^* that minimizes this divergence. For the well-specified case (i.e. no symmetries in the model), it is $t^* = t'$. This forces the model to learn the true data.

4. Formalizing the Platonic Hypothesis

Using Theorem 5 as our foundation, we can now formalize the Platonic Representation Hypothesis. We first define our terms.

Definition 2 (Model and Representation)

1. *A **parametric model class** is a set of distributions $\mathcal{M} = \{p_t : t \in S\}$, where $S \subset \mathbb{R}^d$ is a compact parameter set.*
2. *A **representation map** is a continuous function $R : S \to \mathcal{R}$ that maps a parameter t to its corresponding representation $R(t)$ in a representation space $\mathcal{R}$. $R(t) = \{R(t)(x), x \in Dataset\}$ could be understood as a learned representations of the whole dataset, i. e. the set of all embedding vectors produced by a model with parameters t for all samples in the validation dataset.*
3. *The **approximation gap** of a parametric model class $\mathcal{M}$ is its minimal KL divergence to the true distribution $P^\star$: $\varepsilon_{\mathcal{M}} = \inf_{t \in S} D_{KL}(P^\star \| p_t)$. A model is **universally approximating** if $\varepsilon_{\mathcal{M}} = 0$.*

Theorem 3 (Platonic Representation Convergence) *Consider two model classes, $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, that are both universally approximating ($\varepsilon_{\mathcal{M}^{(1)}} = \varepsilon_{\mathcal{M}^{(2)}} = 0$). Assume that for both models, the optimal parameters (those that minimize KL divergence) map a dataset with N samples to the same unique "Platonic" representation $r_\star \in \mathcal{R}$. Let $\widetilde{\Pi}_N^{(j)}$ be the posterior distribution over the representation space $\mathcal{R}$ for model j. Then, for any neighborhood V of $r_\star$,*

$$\lim_{N\to\infty} \widetilde{\Pi}_N^{(1)}(V) = 1 \quad \textit{and} \quad \lim_{N\to\infty} \widetilde{\Pi}_N^{(2)}(V) = 1 \quad (a.s.).$$

In other words, the posterior distributions on the representation space will converge to the delta function at the $r_\star$.

Conversely, if models are unable to perfectly capture the true data distribution, they converge to different representations.

Theorem 4 (Exponential Separation of Models) *Consider two non-universally approximating model classes, $\mathcal{M}_A = \{p_t : t \in S_A\}$ and $\mathcal{M}_B = \{p_t : t \in S_B\}$, with corresponding representation maps R_A and R_B. Let $\mathcal{R}_A^\star$ and $\mathcal{R}_B^\star$ be the sets of representations*

produced by the parameters that best approximate the true data distribution $P^\star$ within each model class.

Assume that the best representations achievable by these two models are distinct, such that $\mathcal{R}_A^\star$ and $\mathcal{R}_B^\star$ are disjoint. This typically occurs when the models have different approximation capacities.

Let $\Pi_{N,A}$ be the posterior distribution over the representation space for model $\mathcal{M}_A$ given N data samples. For any neighborhood V_B that contains $\mathcal{R}_B^\star$ but is disjoint from $\mathcal{R}_A^\star$, the probability that model A generates a representation inside V_B vanishes exponentially:

$$\Pi_{N,A}(V_B) \leq Ce^{-N\gamma} \quad \textit{for some constants } C > 0 \textit{ and } \gamma > 0.$$

5. Explaining the Platonic Representation Hypothesis

These theorems provide theoretical insights for the empirical findings.

Why does alignment improve with model capacity? "Larger model capacity" corresponds to a smaller approximation gap ($\varepsilon_{\mathcal{M}}$). As capacity increases for two different model families, their approximation gaps $\varepsilon_{\mathcal{M}^{(1)}}$ and $\varepsilon_{\mathcal{M}^{(2)}}$ both approach zero. By Theorem 3, this forces both models to learn parameters that approximate the *same* true data distribution $P^\star$. If the function learned by the model uniquely determines its representation, then both models will converge to the same "Platonic" representation $r_\star$. Smaller models, with larger $\varepsilon_{\mathcal{M}}$, are governed by Theorem 4; they converge to their own best, but imperfect, approximations, which need not be the same, leading to misaligned representations.

Why does random data lead to misalignment? If the data and labels are unconnected (e.g., random labels), the true conditional distribution $P^\star(y|x)$ is independent of x. In this case, the optimization landscape (the expected log-likelihood) becomes flat or possesses many equally good global optima – the set of optimal representations $R(S_\star)$ is no longer a single point. Any representation is valid because none provides a better explanation of the random data than any other. It explains the observations of rokosbasilisk (2024).

Why does alignment improve with dataset size? The dataset size is represented by N in our theorems. The key insight is that the concentration of the posterior is **exponential** in N. As seen in Theorem 4, the posterior probability of a model producing a suboptimal representation vanishes at a rate of $e^{-N\gamma}$. This means that with a large dataset, even minor differences in how well parameters explain the data are massively amplified.

Why does alignment occur across datasets for different modalities? A crucial observation is that models trained on different, disjoint datasets also learn aligned representations. Let us take, for instance, two datasets of texts and images. This occurs because both datasets, if sampled from the same underlying reality, act as proxies for the same true data distribution $P^\star$.

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Robert H Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

Ehsan Imani, Wei Hu, and Martha White. Representation alignment in neural networks. *Transactions on Machine Learning Research*, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. Hessian geometry of latent space in generative models. *arXiv preprint arXiv:2506.10632*, 2025.

rokosbasilisk. Exploring the platonic representation hypothesis beyond in-distribution data. LessWrong, Oct 2024. URL https://www.lesswrong.com/posts/Su2pg7iwBM55yjQdt/exploring-the-platonic-representation-hypothesis-beyond-in. Accessed: 2025-09-05.

Liu Ziyin and Isaac Chuang. Proof of a perfect platonic representation hypothesis. *arXiv preprint arXiv:2507.01098*, 2025.

Appendix A. Proof of theoretical results

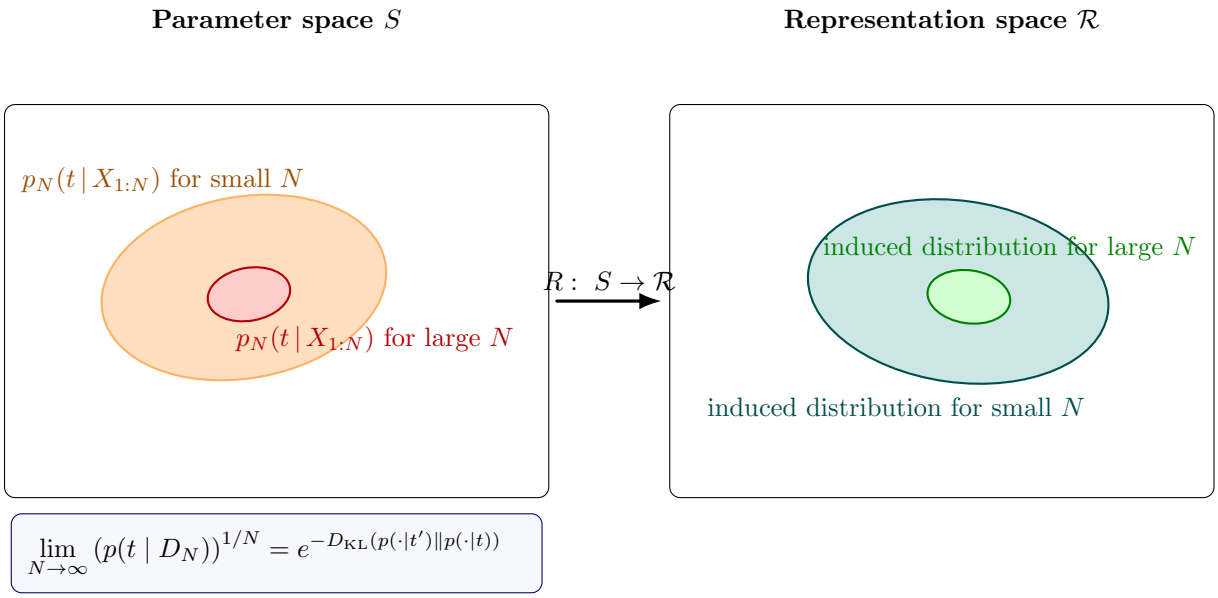

Figure 1: Illustration of Theorems 5 and 3. Posterior concentration in parameter space and its induced alignment in representation space. Concentric KL level sets around $t^\star$ visualize how the posterior sharpens with growing dataset size. The KL bound controls the large sample exponent. Mapping through R collapses the induced distribution in $\mathcal{R}$ toward a limiting representation $r^\star$.

Theorem 5 (Large-N Posterior Limit for General Models) *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measurable space. For each parameter t in a compact set $S \subset \mathbb{R}^n$, let $p(x \mid t)$ be a probability density. Assume the map $(t, x) \mapsto \log p(x \mid t)$ is continuous in t and uniformly continuous in x. Assume a continuous prior $p(t) > 0$ on S. Let $x_1, \ldots, x_N$ be i.i.d. samples from $p(\cdot \mid t')$ for some true parameter $t' \in S$. Then, almost surely,*

$$\lim_{N\to\infty} (p(t \mid x_1, \ldots, x_N))^{1/N} = \exp\bigl(-D_{\mathrm{KL}}(p(\cdot \mid t') \| p(\cdot \mid t))\bigr)\,. \tag{2}$$

Proof The posterior probability of the parameter t given the data D_N is given by Bayes' theorem:

$$p(t|D_N) = \frac{p(D_N|t)p(t)}{\int_S p(D_N|s)p(s)ds}. \tag{3}$$

Since the data samples are i.i.d. given the parameter t, the likelihood term is $p(D_N|t) = \prod_{i=1}^N p(x_i|t)$.

We are interested in the limit of $(p(t|D_N))^{1/N}$. Let's first analyze the likelihood part of the numerator. We can write it as:

$$(p(D_N|t))^{1/N} = \left(\prod_{i=1}^N p(x_i|t)\right)^{1/N} = \exp\left(\frac{1}{N}\sum_{i=1}^N \log p(x_i|t)\right). \tag{4}$$

Since the samples x_i are drawn from the true distribution $p(x|t')$, the Strong Law of Large Numbers states that the average of $\log p(x_i|t)$ converges almost surely to its expectation under $p(x|t')$:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \log p(x_i|t) \overset{\text{a.s.}}{=} \mathbb{E}_{x\sim p(x|t')}[\log p(x|t)]. \tag{5}$$

The prior term $(p(t))^{1/N}$ converges to 1 as $N \to \infty$. Therefore, the numerator of $(p(t|D_N))^{1/N}$ converges almost surely to:

$$\lim_{N\to\infty} (p(D_N|t)p(t))^{1/N} \overset{\text{a.s.}}{=} \exp\left(\mathbb{E}_{x\sim p(x|t')}[\log p(x|t)]\right). \tag{6}$$

The denominator term, also known as the evidence or marginal likelihood, is given by:

$$Z_N = \int_S p(D_N|s)p(s)ds = \int_S \exp\left(N \cdot \frac{1}{N}\sum_{i=1}^{N} \log p(x_i|s)\right) p(s)ds. \tag{7}$$

Let $g_N(s) = \frac{1}{N}\sum_{i=1}^{N} \log p(x_i|s)$. By the Law of Large Numbers, $g_N(s)$ converges almost surely to $g(s) = \mathbb{E}_{x\sim p(x|t')}[\log p(x|s)]$. The integral thus takes the form $\int_S e^{N\cdot g(s)}p(s)ds$.

For large N, such an integral can be approximated by Laplace's method. The value of the integral is dominated by the maximum value of the function $g(s)$ over the domain S. The maximizer s^* is found by:

$$s^* = \arg\max_{s\in S} g(s) = \arg\max_{s\in S} \mathbb{E}_{x\sim p(x|t')}[\log p(x|s)].$$

We can relate this maximization to the KL divergence:

$$D_{\text{KL}}(p(x|t') \,\|\, p(x|s)) = \mathbb{E}_{x\sim p(x|t')}[\log p(x|t')] - \mathbb{E}_{x\sim p(x|t')}[\log p(x|s)].$$

Maximizing $\mathbb{E}_{x\sim p(x|t')}[\log p(x|s)]$ is equivalent to minimizing the KL divergence. By Gibbs' inequality, $D_{\text{KL}} \geq 0$, with equality holding if and only if $p(x|s) = p(x|t')$. Thus, the unique maximizer is $s^* = t'$.

The maximum value of the exponent is $g(t') = \mathbb{E}_{x\sim p(x|t')}[\log p(x|t')]$. By Laplace's method, the asymptotic limit of the N-th root of the evidence is:

$$\lim_{N\to\infty} (Z_N)^{1/N} \overset{\text{a.s.}}{=} \exp\left(\max_{s\in S} g(s)\right) = \exp\left(\mathbb{E}_{x\sim p(x|t')}[\log p(x|t')]\right). \tag{8}$$

By combining the limits for the numerator (6) and the denominator (8), we obtain the final result:

$$\begin{aligned}
\lim_{N\to\infty} (p(t|D_N))^{1/N} &\overset{\text{a.s.}}{=} \frac{\exp\left(\mathbb{E}_{x\sim p(x|t')}[\log p(x|t)]\right)}{\exp\left(\mathbb{E}_{x\sim p(x|t')}[\log p(x|t')]\right)} \\
&\overset{\text{a.s.}}{=} \exp\left(\mathbb{E}_{x\sim p(x|t')}[\log p(x|t)] - \mathbb{E}_{x\sim p(x|t')}[\log p(x|t')]\right) \\
&\overset{\text{a.s.}}{=} \exp\left(-\mathbb{E}_{x\sim p(x|t')}\left[\log \frac{p(x|t')}{p(x|t)}\right]\right) \\
&\overset{\text{a.s.}}{=} e^{-D_{\text{KL}}(p(x|t') \,\|\, p(x|t))}.
\end{aligned}$$

This completes the proof. ∎

Theorem 6 (Platonic Representation Convergence) *Consider two model classes, $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, that are both universally approximating ($\varepsilon_{\mathcal{M}^{(1)}} = \varepsilon_{\mathcal{M}^{(2)}} = 0$). Assume that for both models, the optimal parameters (those that minimize KL divergence) map a dataset with N samples to the same unique "Platonic" representation $r_\star \in \mathcal{R}$. Let $\widetilde{\Pi}_N^{(j)}$ be the posterior distribution over the representation space $\mathcal{R}$ for model j. Then, for any neighborhood V of $r_\star$,*

$$\lim_{N\to\infty} \widetilde{\Pi}_N^{(1)}(V) = 1 \quad \textit{and} \quad \lim_{N\to\infty} \widetilde{\Pi}_N^{(2)}(V) = 1 \quad (\textit{a.s.}).$$

In other words, the posterior distributions on the representation space will converge to the delta function at the $r_\star$.

Proof [Proof Sketch] By Theorem 5, the posterior mass for each model $\mathcal{M}^{(j)}$ concentrates on the set of parameters $S_\star^{(j)}$ that minimize the KL divergence. Since both models are universally approximating, this means they concentrate on parameters that perfectly describe the true distribution $P^\star$. By assumption, the continuous representation map $R^{(j)}$ maps all parameters in $S_\star^{(j)}$ to the same representation $r_\star$. Therefore, the induced posterior over representations must concentrate on $r_\star$ for both models. ■

Theorem 7 (Exponential Separation of Models) *Consider two non-universally approximating model classes, $\mathcal{M}_A = \{p_t : t \in S_A\}$ and $\mathcal{M}_B = \{p_t : t \in S_B\}$, with corresponding representation maps R_A and R_B. Let $\mathcal{R}_A^\star$ and $\mathcal{R}_B^\star$ be the sets of representations produced by the parameters that best approximate the true data distribution $P^\star$ within each model class.*

Assume that the best representations achievable by these two models are distinct, such that $\mathcal{R}_A^\star$ and $\mathcal{R}_B^\star$ are disjoint. This typically occurs when the models have different approximation capacities.

Let $\Pi_{N,A}$ be the posterior distribution over the representation space for model $\mathcal{M}_A$ given N data samples. For any neighborhood V_B that contains $\mathcal{R}_B^\star$ but is disjoint from $\mathcal{R}_A^\star$, the probability that model A generates a representation inside V_B vanishes exponentially:

$$\Pi_{N,A}(V_B) \le Ce^{-N\gamma} \quad \textit{for some constants } C > 0 \textit{ and } \gamma > 0.$$

Proof [Proof Sketch] The proof is a direct consequence of the exponential concentration rate established in Theorem 5. The key insight is the existence of a "KL gap."

Let $\varepsilon_A = \inf_{t\in S_A} D_{\mathrm{KL}}(P^\star \| p_t)$ be the minimum possible KL divergence (the approximation gap) for model class $\mathcal{M}_A$. The posterior for model A will concentrate around the set of parameters that achieve this minimum KL.

By assumption, the set of optimal representations $\mathcal{R}_A^\star$ is separate from the neighborhood V_B. Therefore, any parameter $t \in S_A$ that produces a representation $R_A(t) \in V_B$ is necessarily sub-optimal for model A. Its KL divergence must be strictly greater than the minimum possible value, i.e., $D_{\mathrm{KL}}(P^\star \| p_t) > \varepsilon_A$.

This creates a strictly positive gap, $\Delta_{KL} = D_{\mathrm{KL}}(P^\star \| p_t) - \varepsilon_A > 0$. According to Theorem 5, the posterior mass on any such sub-optimal parameter t is suppressed by a factor proportional to $e^{-N\Delta_{KL}}$. The constant γ in the theorem corresponds to the smallest such KL

gap for parameters producing representations in V_B. Integrating over all such parameters that map into V_B preserves this exponential decay, yielding the result. ■