PRE-TRAINING AND IN-CONTEXT LEARNING IS BAYESIAN INFERENCE *a la* DE FINETTI

Naimeng Ye, Hanming Yang, Andrew Siah, Hongseok Namkoong
Columbia University
New York, 10027, USA
{ny2336, hy2781, andrew.siah, hn2369}@columbia.edu

ABSTRACT

In-context learning (ICL) has emerged as a powerful learning paradigm. Going back to De Finetti's work on Bayesian inference using observables—as opposed to priors on latent factors/parameters—we establish an *explicit* equivalence between ICL and Bayesian inference *a la* De Finetti. From this view, pre-training is precisely empirical Bayes: it optimizes the marginal likelihood of observed sequences; compared to fitting priors in conventional empirical Bayes, pre-training fits posterior predictives using transformers. Our observation highlights previously under-explored capabilities of ICL: statistical inference and uncertainty quantification. Our theory highlights the importance of predictive coherence and motivates a new regularizer for pre-training sequence models to be logically coherent Bayesians statisticians. Our preliminary empirical results demonstrate coherency regularization can substantially improve the inferential capabilities of ICL.

1 INTRODUCTION

In-context learning (ICL) has emerged as a powerful learning paradigm where autoregressive generation provides a versatile pattern recognition model without explicit training [6]. The ML folk wisdom is that ICL is akin to Bayesian reasoning. In this work, we provide an exactly characterization of in-context learning as *explicit* Bayesian inference. Our observation is tautological and does not rely on particular architectural structures or elaborate data generation models. Instead of the traditional Bayesian modeler who posits a prior and likelihood on latent factors that are fundamentally unobservable, we go back to De Finetti's celebrated work [9] which focused on modeling *observables* via posterior predictives.

Our main observation is that autoregressive generative models give rise to a Bayesian inference model *a la* De Finetti. We note that autoregressive loss (perplexity) is the *marginal likelihood* of an observed sequence prescribed by a Bayesian modeler. Thus, standard pre-training is empirical Bayes: instead of optimizing marginal likelihoods through priors, it optimizes it by modeling posterior predictives (a.k.a. autoregressive probabilities). Our perspective highlights the statistical capabilities of ICL: going beyond the predictive paradigms studied in prior works [14; 1], ICL provides natural statistical inference and uncertainty quantification. We expand the previously proposed downstream ICL tasks to include those that require uncertainty quantification.

Our theory suggests how to improve the coherence of the Bayesian reasoning capabilities of ICL. The formal validity of the Bayesian inference model on observables (which we denote "*a la* De Finetti") relies on the coherence of the pre-trained posterior predictive distributions. i.e., Does it follow the Bayes rule according to some prior? We focus on a particular coherence condition [10], and propose a regularizer that pre-trains models for valid Bayesian inference in addition to autoregressive predictive performance. Our preliminary experimental results demonstrate that our coherence regularizer significantly improves the quality of statistical inference.

Several authors recently explored formal models of ICL by using Hidden Markov Models (HMMs) [22], statistical learning with stability conditions [18], or gradient descent-based algorithms [8; 1]. We take a different approach by connecting several disparate literatures: Bayesian statistics [9; 2; 12; 11; 17; 4; 5; 10], meta learning [19; 20], and Bayesian deep learning [15]. By articulating the exact connection between in-context learning and Bayesian inference, we provide insights that ICL has the ability to pose as a Bayesian statistician, and not merely a predictive model.

2 BAYESIAN MODELING *a la* DE FINETTI

Consider i.i.d. sequences $Y^i_{1:T_i} = \{Y^i_1, \cdots, Y^i_{T_i}\}$ for $i = 1, \ldots, n$, where Y^i_t is a "token" that take on continuous or discrete values. Define one-step predictive probabilities

$$\widehat{p}_t(y) := \widehat{p}_t(y \mid Y_{1:t}) := \widehat{\mathbb{P}}\left(Y_{t+1} = y \mid Y_{1:t}\right). \tag{1}$$

Generative modeling fits a autoregressive model (e.g., decoder transformer) to optimize the log likelihood of the observed sequences

$$\text{Pre-training:} \quad \underset{\widehat{p}(\cdot)}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log \widehat{p}\left(Y^i_{1:T_i}\right) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T_i - 1} \log \widehat{p}_t\left(Y^i_{t+1}\right) \right\}. \tag{2}$$

A generative model $\widehat{p}$ can be used to tackle a range of different tasks by conditioning on any sequence/prompt $Y_{1:s}$ at inference time ("in-context learning"). Although our subsequent results and algorithms can consider covariates, we ignore them to simplify exposition.

2.1 PRE-TRAINING IS EMPIRICAL BAYES ON POSTERIOR PREDICTIVES

We present a Bayesian modeling paradigm based on observables rather than latent parameters.

Classical Bayesians A classical Bayesian statistician posits a latent factor θ ("parameter"), a distribution over the parameter, P_θ ("prior"), and how θ governs data-generation, $\mathbb{P}(Y_{1:t} = \cdot \mid \theta)$ ("likelihood"). Given observed data $Y_{1:s}$, the main quantity of interest is the posterior distribution $\mathbb{P}(\theta = \cdot \mid Y_{1:s})$, which measures the epistemic uncertainty on the latent structure. De Finetti's characterization of an exchangeable sequence $Y_{1:\infty}$ provides the basis of modeling latent factors: there is θ such that $\mathbb{P}(Y_{1:\infty} = y_{1:\infty}) = \int \prod_{t=1}^{\infty} \mathbb{P}(Y_t = y_t \mid \theta) \mathbb{P}(d\theta)$. The main challenge with this modeling paradigm is the need to specify a model over latent factors and argue for its validity despite its fundamentally unobservable nature. A practical model validation metric is to check whether the posited model on unobservables explain observed data well [16]: for a posited model $\widehat{p}$, the *marginal likelihood* measures whether $\widehat{p}$ explains observed sequences $\mathcal{R}_n(\widehat{p}) := \frac{1}{n} \sum_{i=1}^{n} \log \int \widehat{p}\left(Y_{1:T_i} \mid \theta\right) \widehat{p}(\theta) d\theta$.

Bayesians *a la* De Finetti Instead of modeling unobservables, an alternative paradigm is to model the observable sequence $Y_{1:\infty}$. Unlike a fictitious latent factor θ that the modeler proposes but is never observed, the sequence $Y_{1:\infty}$ is observable in principle (the "future" sequence $Y_{s+1:\infty}$ is simply yet to be observed). De Finetti's celebrated works [7] focus on modeling the relationships between observable quantities (1); notably, we can now validate the modeler's claims by masking part of the observed data from the modeler.

De Finetti's original representation [9] goes beyond the previous representation result, but shows the latent factor θ in this representation result is entirely a function of observables $Y_{1:\infty}$. Indeed, for (almost) any realization of an exchangeable sequence $Y_{1:\infty}$, the strong law of large numbers dictates that there is a limiting probability $P_\infty(\cdot \mid y_{1:\infty})$ such that $\frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left\{Y_t(\omega) \leq y\right\} \to P_\infty(y \mid Y_{1:\infty}(\omega))$. The latent factor $\theta := P_\infty(\cdot \mid Y_{1:\infty})$ is entirely determined by the observations $Y_{1:\infty}$, and is no longer a fictitious never-observable quantity as in the classical regime.

To operationalize De Finetti's philosophy, we take the *posterior predictive probabilities* (1) as our modeling primitive to approximate the marginal likelihood $\mathbb{P}(Y_{1:T} = y_{1:T}) \approx \widehat{p}(y_{1:T}) = \prod_{t=0}^{T-1} \widehat{p}_t(y_{t+1})$. Instead of priors and likelihoods, the modeler specifies one-step probabilities (1) on *observables*. A long line of work in Bayesian statistics advocates for this approach to Bayesian modeling [2; 12; 11; 17; 4; 5; 10]. They propose simple parameterizations for one-step probabilities (e.g., copulas [17; 10]) and identify conditions under which one-step posterior predictive distributions implicitly characterize the prior and likelihood over the latent factor θ.

Since it is difficult to specify one-step probabilities over long sequences, we model them using sequence models (e.g., transformers) following previous works [15; 19; 20]. We adopt the empirical Bayes philosophy: when our one-step probabilities accurately model the data-generating distribution, masked observations will have high marginal likelihood $\widehat{p}(Y_{s+1:t} \mid Y_{1:s})$. Note that this is *precisely* the original pre-training problem (2)! We conclude pre-training and in-context learning is empirical Bayes on posterior predictive densities.

Machine Learning is a discipline of generalizing across learned representations. As showcased by the empirical success in language modeling, the empirical Bayes problem (2) will be effective if there is a wealth of previously observed sequences $\{Y^i_{1:T_i}, i = 1, \ldots, n\}$.

2.2 In-context learning as explicit Bayesian inference

One-step probabilities (1) may not necessarily correspond to posterior predictions consistent with a single prior. We discuss conditions on the one-step probabilities that guarantee a notion of predictive coherence, ensuring that they (roughly) follow the Bayes rule according to some prior. As we show, coherence provides valid statistical inference and guarantees good performance on downstream tasks that require uncertainty quantification.

Berti et al. [3] proposes predictive coherence condition that extends the familiar exchangeability.

Definition 1. *$Y_{1:\infty}$ is conditionally identically distributed (c.i.d.) if*

$$\mathbb{P}(Y_{t+2} = y \mid Y_{1:t}) = \mathbb{P}(Y_{t+1} = y \mid Y_{1:t}) =: p_t(y \mid Y_{1:t}) =: p_t(y) \quad \textit{for all } y \in \mathbb{R}. \tag{3}$$

The c.i.d. condition is a starting point for studying previously proposed formalizations of ICL [14]. For example, for a question answering task, we can generalize the condition to a subsequence level condition $\mathbb{P}(Y_{\tau_1:\tau_2} = y_{\tau_1:\tau_2} \mid Y_{1:\tau_1}) = \mathbb{P}(Y_{\tau_2:\tau_3} = y_{\tau_2:\tau_3} \mid Y_{1:\tau_1})$ where we denote each question-answer sequence as from τ_i to τ_{i+1}. However, we don't expect language to satisfy c.i.d. in general. Conditioning on an incomplete sentence, the distribution of the next token versus the last token should clearly not follow the same distribution. Relaxing this condition is a topic of future research.

A direct consequence of the c.i.d. condition is that one-step probabilities have a limit [3]. If $Y_{1:\infty}$ is c.i.d., $\mathbb{E}[p_t(y) \mid Y_{1:t-1}] = \mathbb{E}[\mathbf{1}\{Y_{t+1} = y\} \mid Y_{1:t-1}] = p_{t-1}(y)$ for any y. Consequently, equation 3 and the martingale convergence theorem yields

$$\exists \text{ random distribution } p_\infty(\cdot \mid Y_{1:\infty}) \text{ such that } p_t(y \mid Y_{1:t}) \to p_\infty(y \mid Y_{1:\infty}) \text{ almost surely.} \tag{4}$$

We interpret the random limit p_∞ as a "latent factor" entirely determined by the infinite observations.

Under the c.i.d. condition, autoregressive generation gives a natural Bayesian inference procedure based on the bootstrap [10]. Consider $\tau^\star = \int g(y) p_\infty(y) dy$, the mean of some function $g : \mathbb{R} \to \mathbb{R}$ under p_∞. Letting $y_{1:s}$ be the observed data, our goal is to generate a confidence/credible interval around $\tau^\star$. For T large enough, autoregressively generate $Y_{t+1} \sim p_t(\cdot \mid y_{1:s}, Y_{s+1:t})$ and compute $\widehat{\tau} = \tau(Y_T)$. Repeating this B times, we obtain $\widehat{\tau}^1, \ldots, \widehat{\tau}^B \overset{\text{iid}}{\sim} \mathbb{P}_T(\cdot | y_{1:s})$. Under the c.i.d. condition, Fong et al. [10] shows that this provides a valid inferential procedure as $T, B \to \infty$.

3 Downstream tasks that require uncertainty quantification

Under the c.i.d. condition, we show the pre-training objective (2) (perplexity) is the correct performance measure capturing Bayesian inferential capabilities.

Assumption A. *The true data-generating distribution and its pre-trained counterpart are c.i.d. (3).*

Our subsequent results rely on the limiting marginal likelihood / perplexity.

Theorem 1. *Under regularity conditions,* $\frac{1}{T}\sum_{t=1}^{T} \log \widehat{p}_t(Y_t) \to \int p_\infty(y) \log \widehat{p}_\infty(y) dy =: H(\widehat{p})$.

The true data-generating distribution is clearly the "best model": from Jensen's inequality, for any $\widehat{p}$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \log \widehat{p}_t(Y_t)\right] \le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \log p_t(Y_t)\right] \quad \text{and} \quad \int p_\infty(y) \log \widehat{p}_\infty(y) dy \le \int p_\infty(y) \log p_\infty(y) dy.$$

Pre-training guarantees valid inference under c.i.d. We articulate several frequently overlooked applications of in-context learning that require uncertainty quantification and Bayesian reasoning. First, we highlight how in-context learning can be used for statistical inference, going beyond the usual predictive applications. Second, we show Bayesian inference allows *length generalization* in sequence predictions: robustness to longer sequence prediction or shorter context length than seen during training.

For simplicity, we again assume that we are interested in $\tau^\star = \int g(y_1, \ldots, y_k) p_\infty^k(y_1, \ldots, y_k) dy_1 \cdots dy_k$, the mean of a bounded function $g : \mathbb{R}^k \to \mathbb{R}$. Note that $\tau^\star$ has two sources of uncertainty: observables $Y_{1:\infty}$ that determine p_∞ (epistemic) and irreducible noise in the realizations of p_∞ (aleatoric). Given limited observables as context $Y_{1:s}$, ICL provides an effective way to perform inference on $\tau^\star$ via forward sampling observables $Y_{s:T}$ to a long horizon T. Under Assumption A, the pre-training objective (marginal log likelihood or negative log perplexity) offers pathwise control over $\widehat{\tau}_T - \tau^\star$.

$$\lim_{T\to\infty} \widehat{\tau}_T - \tau^\star \lesssim \|g\|_\infty \sqrt{k D_{\mathrm{kl}}\left(p_\infty \| \widehat{p}_\infty\right)} \propto H(\widehat{p}_\infty)^{\frac{1}{2}}.$$

Thus, pre-training problem (2) is the "right" objective to guarantee the quality of Bayesian inference when $\widehat{p}_t$ is c.i.d..

We show that pre-training objective (2) also governs the *sequential* prediction performance over long horizons. Given a "prompt" consisting of the sequence of tokens $Y_{1:s}$ at inference time, consider the canonical ICL prediction problem where we wish to generate $\widehat{Y}_{s+1}, \widehat{Y}_{s+2}, \cdots$ such that they closely match the unseen observations $Y_{s+1}, Y_{s+2}, \cdots$. We evaluate ourselves on the T-horizon squared loss $\widehat{R}_T := \frac{1}{T}\sum_{t=1}^T (\widehat{Y}_t - Y_t)^2$.

We are interested in *generalizing beyond the observed sequence length* seen during training; the model must learn to make accurate predictions for longer horizons. We study the limiting behavior of the model $\widehat{p}_t$ as $t \to \infty$ and show that the marginal log likelihood (2) directly controls length generalization capabilities of the fitted model.

$$\lim_{T\to\infty} \widehat{R}_T \le \mathrm{Var}\left(g(\widehat{Y}_\infty)\right) + \mathrm{Var}\left(g(Y_\infty)\right) + \|g\|_\infty D_{\mathrm{kl}}\left(p_\infty \| \widehat{p}_\infty\right) \propto \|g\|_\infty \left(1 + H(\widehat{p})\right).$$

c.i.d. regularization Our discussion highlights the importance of predictive coherence guaranteed by the c.i.d. condition (3). Since the pre-training objective (2) does not guarantee coherence, we propose a regularizer that measure violations to c.i.d.

$$\sum_{t=0}^{T_i-2} D_{\mathrm{kl}}\left(\widehat{p}_t(\cdot) \| \, \widehat{p}(\widehat{Y}^i_{t+2} = \cdot \mid X^i_{1:t}, Y^i_{1:t}, X^i_{t+1})\right).$$

The second distribution is not directly obtainable from an autoregressive model (e.g., transformer). Recalling $D_{\mathrm{kl}}\left(\mathcal{N}(\mu_1, \sigma_1^2) \| \mathcal{N}(\mu_2, \sigma_2^2)\right) = \frac{1}{2}\left(\sigma_2^{-1}\sigma_1 - 1 + \sigma_2^{-1}(\mu_2 - \mu_1)^2 + \ln\frac{\sigma_2}{\sigma_1}\right)$, we estimate its mean and variance using Monte Carlo samples $\zeta \sim \widehat{p}_{t+1}$.

4 Experiments

We present preliminary experimental results that highlight the importance of predictive coherence in ICL. We consider the meta-learning linear regression setting in Garg et al. [14]. Consider a class of linear functions $\mathcal{F} = \{f \mid f(x) = w^T x, x \in \mathbb{R}^d\}$ where each environment/sequence w^i generates observations $Y^i_t = w^{i\top} X^i_t + \epsilon^i_t$ for $X^i_t \overset{\text{iid}}{\sim} N(0, I_d)$ and $\epsilon^i_t \sim N(0, 0.1)$. Instead of point predictions, we modify the architecture of the sequence model to output a probability $\widehat{p}^i_t = \mathcal{N}(\mu_t(X^i_{1:t+1}, Y^i_{1:t}), \sigma_t(X^i_{1:t+1}, Y^i_{1:t}))$. We compare the performance of two types of generative models: one trained to minimize the usual pre-training objective (2), and one also optimized for predictive coherence using c.i.d. regularizer. Specifically, we train two 1D models on context length 5 and 10, and test the models' performances on inference and prediction tasks, given context length 5.

Statistical inference We are interested in generating a confidence interval on w, where we observe $(X_{1:s}, Y_{1:s})$ at inference/ICL time. Instead of performing inference over w through a prior and likelihood, we autoregressively generate $Y^i_{t+1} \sim p^i_t(\cdot)$ and compute $\widehat{w}^i_1, \ldots, \widehat{w}^i_B$ via the Bayesian bootstrap procedure described at the end of Section 2.2. In Figure 1, we observe that c.i.d regularization provides a coherent forward sampled trajectories, compared to a naive generative model that degrades with more forward samples. Quantitatively, on models trained with context length 5, we observe that the parameter estimated from KL-regularized sequences were 71.57% closer to the ground truth parameter in terms of absolute distance (regularized 0.1710, unregularized 0.6014). On models trained with context length 10, KL-regularization also improved absolute distance from the ground truth parameter by 88.10% (regularized 0.0639, unregularized 0.5369).

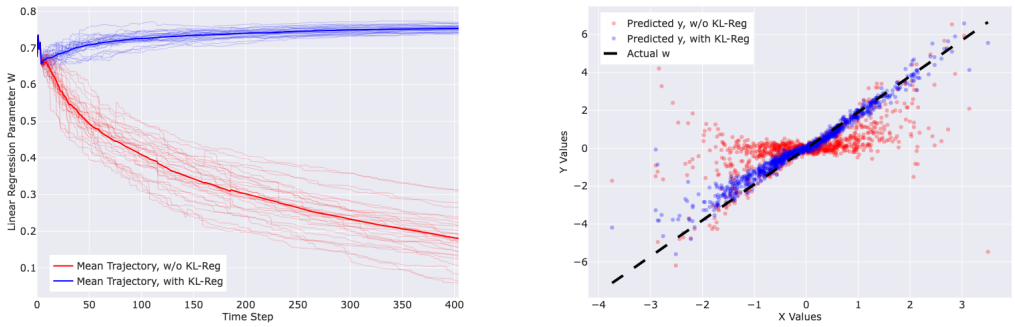

Figure 1. **Plot**(a):$\widehat{w}$ trajectories across independently sampled forward predictions, with context $\{x_{1:5}, y_{1:5}\}$. As we recursively forward sample, the least squared $\widehat{w}$ on each trajectory converges for c.i.d. regularized model. In contrast, the non-regularized model generates sample paths that degrade as i increases. **Plot** (b):Forward prediction up to horizon $T = 1000$, with context $\{x_{1:5}, y_{1:5}\}$, and covariates $x_{5:T}$. We then plot the generated $\{\widehat{y}_{5:T}\}$ against the given coviates $x_{5:T}$ as the above graph.

Sequential prediction We also study the model's *length generalization* ability, predicting beyond the sequence length seen during pre-training. Figure 1 shows that regularized model gives better T-horizon predictions. We calculate T-horizon squared loss with context length 5 and $T = 1000$. On models trained with context length 5, we observe a 92.39% increase (regularized 0.2075, unregularized 2.7249). On models trained with context length 10, we observe a 94.99% increase (regularized 0.0597, unregularized 1.1924).

REFERENCES

[1] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems 36*, 2023.

[2] Patrizia Berti, Eugenio Regazzini, and Pietro Rigo. Well calibrated, coherent forecasting systems. *Theory of Probability & Its Applications*, 42(1):82–102, 1998.

[3] Patrizia Berti, Luca Pratelli, and Pietro Rigo. Limit theorems for a class of identically distributed random variables. *Annals of Probability*, 32(3):2029 – 2052, 2004.

[4] Patrizia Berti, Emanuela Dreassi, Luca Pratelli, and Pietro Rigo. A class of models for bayesian predictive inference. *Bernoulli*, 27(1):702 – 726, 2021.

[5] Patrizia Berti, Emanuela Dreassi, Fabrizio Leisen, Luca Pratelli, and Pietro Rigo. Bayesian predictive inference without a prior. *Statistica Sinica*, 34(1), 2022.

[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners. *arXiv:2005.14165 [cs.CL]*, 2020.

[7] Donato Michele Cifarelli and Eugenio Regazzini. De finetti's contribution to probability and statistics. *Statistical Science*, 11(4):253–282, 1996.

[8] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019. Association for Computational Linguistics, 2023.

[9] B de Finetti. Classi di numeri aleatori equivalenti. la legge dei grandi numeri nel caso dei numeri aleatori equivalenti. sulla legge di distribuzione dei valori in una successione di numeri aleatori equivalenti. *R. Accad. Naz. Lincei, Rf S 6a*, 18:107–110, 1933.

[10] Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society, Series B*, 2023.

[11] Sandra Fortini and Sonia Petrone. Predictive distribution (de f inetti's view). *Wiley StatsRef: Statistics Reference Online*, pp. 1–9, 2014.

[12] Sandra Fortini, Lucia Ladelli, and Eugenio Regazzini. Exchangeability, predictive distributions and parametric models. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 86–109, 2000.

[13] Ankit Garg, Tengyu Ma, and Huy L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems 27*, 2014.

[14] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

[15] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2018.

[16] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis*. Chapman & Hall, third edition, 2013.

[17] P Richard Hahn, Ryan Martin, and Stephen G Walker. On recursive bayesian predictive distributions. *Journal of the American Statistical Association*, 113(523):1085–1093, 2018.

[18] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023.

[19] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *Proceedings of the Tenth International Conference on Learning Representations*, 2022x.

[20] Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[21] Henry Teicher. Strong laws for martingale differences and independent random variables. *Journal of Theoretical Probability*, 11:979–995, 1998.

[22] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv:2111.02080 [cs.CL]*, 2021.

A PROOF OF THEOREM 1

We first restate the theorem with formal assumptions.

A.1 NOTATION

In previous notation, we assume the predictive distributions are all from measure space $(\Omega^\infty, \mathcal{A}^\infty)$, so we always condition on $y_\infty = Y_{1:\infty}(\omega^\infty)$. Under this notation, for any $t < \infty$, we re-interpret $p_t(\cdot \mid y_\infty)$ and $\widehat{p}_t(\cdot \mid \widehat{y}_\infty)$ as the one-step *random* posterior predictive densities at time $t+1$, given the history up to time t, i.e. ignoring the present and the future realizations.

$$p_t(\cdot \mid \cdot)(\omega^\infty) = \mathbb{P}(\cdot \mid Y^\omega_{1:t}) \qquad \hat{p}_t(\cdot \mid \cdot)(\omega^\infty) = \widehat{\mathbb{P}}(\cdot \mid \widehat{Y}^\omega_{1:t})$$

For notational simplicity again, we drop the realization $y_{1:\infty}$ in the conditioning in following proof.

We also impose the following coupling assumption so that we can exchangeably condition on random variables generated by true p or $\widehat{p}$: for any $A \in \mathcal{A}^\infty$

$$\mathbb{E}[\mathbb{1}_A(\widehat{Y}_{t+1}) \mid Y_{1:t}] = \mathbb{E}[\mathbb{1}_A(\widehat{Y}_{t+1}) \mid \widehat{Y}_{1:t}]$$

This condition allows the following equation to hold

$$\begin{aligned}\mathbb{E}[\widehat{p}_t(y) \mid Y_{1:t-1}] = \mathbb{E}[\mathbb{1}(\widehat{Y}_{t+1} = y \mid \widehat{Y}_{1:t}) \mid Y_{1:t-1}] &= \mathbb{E}[\mathbb{1}(\widehat{Y}_{t+1} = y) \mid Y_{1:t-1}] \\ &= \mathbb{E}[\mathbb{1}(\widehat{Y}_{t+1}) \mid \widehat{Y}_{1:t-1}] = \widehat{p}_{t-1}(y)\end{aligned}$$

where the second equality is due to the tower law and the second to last equality is due to the coupling assumption. Therefore, we have that the martingale property for $\widehat{p}$ holds even if the observations had come from the true distribution p instead of iteself.

Theorem 2. *Under the following assumptions:*

1. *The true data-generating distribution $Y_{t+1} \sim p_t(\cdot)$ is c.i.d. (Definition 1). Moreover, assume that the one-step posterior predictive probabilities converges uniformly, a.s. i.e.*
$$\sup_y |p_t(y \mid y_\infty) - p_\infty(y \mid y_\infty)| \to 0$$

2. *The pre-trained one-step posterior predictive probabilities $\widehat{Y}_{t+1} \sim \widehat{p}_t$ also obey the c.i.d identity (3), i.e., $\widehat{\mathbb{P}}(\widehat{Y}_{t+2} = \cdot \mid y_{1:t}) = \widehat{p}_t(\cdot \mid y_{1:t})$ for all $t \in \mathbb{N}$ and $y_{1:\infty}$. Moreover, assume that the one-step posterior predictive probabilities converges uniformly, a.s. i.e.*
$$\sup_y |\log \widehat{p}_t(y \mid y_\infty) - \log \widehat{p}_\infty(y \mid y_\infty)| \to 0$$

3. *Martingale with difference sequence $\{D_t = \log \widehat{p}_t(Y_t \mid Y_{1:t-1}) - \int p_t(y) \log \widehat{p}_t(y) dy\}$ being L_2, and that*
$$\sum_{n=1}^{\infty} P\left(|D_n| > \frac{n}{\log\log n} \mid \mathcal{F}_{n-1}\right) < \infty \quad a.s.$$
$$\sqrt{\sum_{i}^{n} \mathbb{E}\left[D_j^2 \mid \mathcal{F}_{j-1}\right] \cdot n \log\log n} \to 0 \quad a.s.$$

4. *$\int \log \widehat{p}_\infty(y) dy$ is bounded.*

we have the following convergence almost surely

$$\frac{1}{T}\sum_{t=1}^{T} \log \widehat{p}_t(Y_t) \to \int p_\infty(y) \log \widehat{p}_\infty(y) dy =: H(\widehat{p}). \tag{5}$$

Proof. Again for notational convenience, we drop the realization $Y_{1:\infty}$ in the conditioning in following proof.

$$\frac{1}{T}\sum_{t=1}^{T}\log\widehat{p}_t(Y_t) - \int p_\infty(y)\log\widehat{p}_\infty(y)dy = \frac{1}{T}\sum_{t=1}^{T}(\log\widehat{p}_t(Y_t) \qquad - \int p_t(y)\log\widehat{p}_t(y)dy) \tag{6}$$

$$+ \frac{1}{T}\sum_{t=1}^{T}(\int p_t(y)\log\widehat{p}_t(y)dy \;\; - \int p_t(y)\log\widehat{p}_\infty(y)dy) \tag{7}$$

$$+ \frac{1}{T}\sum_{t=1}^{T}(\int p_t(y)\log\widehat{p}_\infty(y)dy - \int p_\infty(y)\log\widehat{p}_\infty(y)dy) \tag{8}$$

We then prove that each of the above term converges almost surely to 0.
For the first term, note that

$$\log\widehat{p}_t(Y_t \mid Y_{1:t-1}) \in \mathcal{F}_t$$
$$\int p_t(y)\log\widehat{p}_t(y)dy = \mathbb{E}_{Y\sim p(\cdot|Y_{1:t-1})}[\log\widehat{p}_t(Y_t \mid Y_{1:t-1})] \in \mathcal{F}_{t-1}$$

Hence,

$$\begin{aligned}\mathbb{E}[\log\widehat{p}_t(Y_t \mid Y_{1:t-1}) &- \mathbb{E}_{Y\sim p(\cdot|Y_{1:t-1})}[\log\widehat{p}_t(Y_t \mid Y_{1:t-1})] \mid \mathcal{F}_{t-1}] \\ &= \mathbb{E}\left[\log\widehat{p}_t(Y_t \mid Y_{1:t-1}) \mid \mathcal{F}_{t-1}\right] - \mathbb{E}_{Y\sim p(\cdot|Y_{1:t-1})}[\log\widehat{p}_t(Y_t \mid Y_{1:t-1})] \\ &= 0\end{aligned}$$

Therefore, the sequence $D_t = \log\widehat{p}_t(Y_t \mid Y_{1:t-1}) - \int p_t(y)\log\widehat{p}_t(y)dy = \log\widehat{p}_t(Y_t \mid Y_{1:t-1}) - \mathbb{E}[\log\widehat{p}_t(Y_t) \mid Y_{1:t-1}]$ is clearly a martingale difference sequence. By the SLLN for martingale differences, under same conditons around the variance, we have $D_t \to 0$ almost surely by corollary 2 of [21]. Note that furthermore, we can always use azuma-bernstein type of concentration results, here to characterize deviation, though this requires a much stronger assumption on the moment of the difference sequence.
For the second term, we have that each individual term converges to 0 almost surely.

$$\begin{aligned}\int p_t(y)\log\widehat{p}_t(y)dy - \int p_t(y)\log\widehat{p}_\infty(y)dy &= \int p_t(y)(\log\widehat{p}_t(y) - \log\widehat{p}_\infty(y))dy \\ &\leq \sup_y |\log\widehat{p}_t(y) - \log\widehat{p}_\infty(y)| \int p_t(y)dy \\ &\to 0\end{aligned}$$

Then clearly the averaged summation also converges to the 0 almost surely.
Similarly for the last term, we have

$$\begin{aligned}\int p_t(y)\log\widehat{p}_\infty(y)dy - \int p_\infty(y)\log\widehat{p}_\infty(y)dy &\leq \sup_y |p_t(y) - p_\infty(y)| \int \log\widehat{p}_\infty(y)dy \\ &\to 0\end{aligned}$$

and hence the averaged summation also converges to 0 almost surely. This concludes the proof. □

B Experimental Setting

We train the GPT2 model as specified in [13] with 4 heads, 128 input embeddings, and 12 linear layers on four A100 Nvidia GPUs. We use the Adam optimizer with a learning rate of 0.0001 and the KL regularized autoregressive loss. KL regularization term is calculated with $M = 120$ MC samples. Our model is exposed to $640,000$ data points, for 1 epoch, on batch size 32.

C Results

On context length 5 and 10, we train two models each with $\lambda = 0$ and $\lambda = 100$. We then test the models' performance on inference and prediction based tasks.

C.1 Sequential Prediction

We calculate T-horizon squared loss with context length 5 and $T = 1000$. On models trained with context length 5, we observe a 92.39% increase, and on models trained with context length 10, we observe a 94.99% increase. Refer to the following table for specific numbers 1, and figure 2 for a visualizaiton of the mean squared error across time-steps.

	Train context length 5		Train context length 10	
Next-step Prediction	$\lambda = 0$	$\lambda = 100$	$\lambda = 0$	$\lambda = 100$
	0.016047357	0.07333319	0.004384955	0.008139687
T-horizon Prediction	$\lambda = 0$	$\lambda = 100$	$\lambda = 0$	$\lambda = 100$
	2.7249305	0.20747349	1.1924113	0.059748333

Table 1. Mean squared error of next-step and T-horizon predictions given context $\{x_{1:5}, y_{1:5}\}$. With regularization, the model performs better at long horizon sequential prediction, either with same context length as training data, or with shorter context length than training data. However, we do see that regularization slighly undermines the models' ability to predict the next-step. We suspect that this is due to the fact that regularization forces the model to be more consistent in forward predictions, which may lead to a slight increase in immediate next step prediction error.

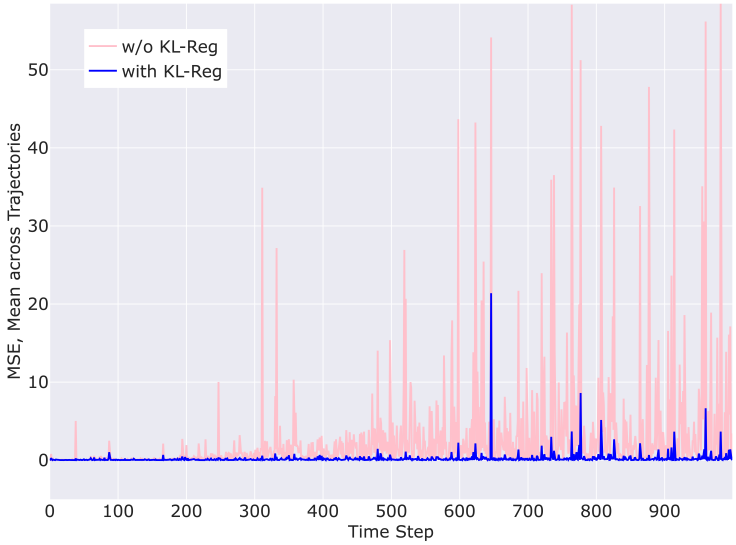

Figure 2. MSE, averaged across trajectories at each forward sampling time-step, given context $\{x_{1:5}, y_{1:5}\}$. With regularization, the model is able to forward predict with better consistency and accuracy.

C.2 Statistical Inference

We calculate the absolute distance of the estimated parameter from the ground truth parameter. On models trained with context length 5, we observe that the parameter estimated from KL-regularized sequences were 71.57% closer to the ground truth parameter in terms of absolute distance. On models trained with context length 10, KL-regularization also improved absolute distance from the ground truth parameter by 88.10%. Refer to the following table for specific numbers 2, and figure 1 for a visualizaiton of the parameter trajectories.

	Train context length 5		Train context length 10	
Distance to true w	$\lambda = 0$	$\lambda = 100$	$\lambda = 0$	$\lambda = 100$
	0.6014	0.1710	0.5369	0.0639
Distance to Least Square solution by context w	$\lambda = 0$	$\lambda = 100$	$\lambda = 0$	$\lambda = 100$
	0.6119	0.1595	0.5234	0.0749

Table 2. Absolute distance of the estimated regression parameter from both the ground truth parameter, and the least square parameter with regards to context $\{x_{1:5}, y_{1:5}\}$. With regularization, the model is able to estimate the parameter with lower error.