APPROXIMATE SOLOMONOFF INDUCTION: ON THE ROBUSTNESS OF LEARNED UNIVERSAL PREDICTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Why large language models trained via log-loss minimization over massive and diverse datasets exhibit broad capabilities and generalizable skills is not yet fully understood. Grau-Moya et al. (ICML 2024) offer a conceptual explanation through the lens of meta-learning: if sequence-generating environments ν are drawn from a prior w_ν, minimizing population log-loss recovers the corresponding Bayesian mixture predictor $\xi = \sum_\nu w_\nu \nu$. In the idealized limit in which w is a universal prior, the optimal predictor corresponds to Solomonoff universal induction. Practical machine learning algorithms, however, generally produce only *approximately* optimal predictors. This raises a natural question: *Does the prediction performance remain robust under approximation*? In this paper, we first provide a tight bound on the worst-case effect of approximation on prediction error. We show any predictor ρ satisfying $D_{KL}(\xi\|\rho) \leq \eta$ retains bounded cumulative-expected-squared prediction error. However, while the Bayesian mixture predictor ξ has error $O(\log(1/w_\mu))$ in an environment μ, we establish a tight $\Theta(\eta/w_\mu)$ bound on the excess error of ρ relative to ξ. Thus, approximation can substantially degrade performance in environments assigned small prior probability. In sharp contrast, when environments are averaged according to prior w, we prove that the excess prediction error is only $O(\eta)$. Furthermore, we establish a distribution-shift theorem for dominated shifts from a training prior w to a test prior w', interpolating our average-case and worst-case guarantees. Finally, we extend our robustness analysis to a finite-sampling setting relevant to learning universal predictors with neural networks and Transformers, where $D_{KL}(\xi\|\rho)$ could be unbounded. Our theory extends the classical theory of Bayesian and Solomonoff Induction to approximate predictors and characterizes a broad spectrum of prediction behavior. It helps explain the sharp gap between strong empirical benchmark performance and poor worst-case performance in rare environments, while providing quantitative justification for fine tuning and synthetic data generation designed to protect important but underrepresented task families.

1 INTRODUCTION

Large language models (LLMs) trained for *sequential prediction* have acquired a remarkable range of capabilities from vast and diverse data. In recent years, scaling model capacity, expanding the volume and breadth of training data, increasing computational resources (Kaplan et al., 2020), and incorporating real-world human interaction (Ouyang et al., 2022) have driven substantial progress, transforming LLMs into increasingly general-purpose systems for problem solving and induction. Given the substantial and rapidly growing resources required to train frontier models, understanding the benefits and limitations of further scaling has become more important than ever, raising a range of fundamental research questions.

Classical learning frameworks, such as semi-supervised learning, PAC learning, and reinforcement learning, typically analyze learning relative to a specified task, data distribution, and hypothesis class. In contrast, a single LLM trained on heterogeneous data can perform a wide range of tasks—including translation, writing, coding, and mathematical reasoning—without task-specific retraining, guided only by instructions or a few demonstrations provided in context. Although pretraining is largely driven by minimizing next-token log loss, the resulting models exhibit capabilities involving generation, abstraction, reasoning, and planning, often without explicit supervision

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for each capability. Moreover, once trained, such a model can be deployed at a comparatively low marginal cost per inference (DeepSeek-AI, 2025). LLM training thus distills enormous amounts of data and computation into a reusable neural network whose fixed parameters support adaptive behavior across contexts and tasks. This phenomenon suggests that learning to predict sufficiently rich and heterogeneous sequences can give rise to broad knowledge and generalizable computational abilities. Precisely how such capabilities emerge from a seemingly simple prediction objective, however, remains incompletely understood, and is central to characterizing the fundamental power and limitations of the current LLM paradigm.

Motivating Prior Work. *Meta-learning*, or learning to learn, provides a conceptual framework for understanding this phenomenon (Genewein et al., 2026). Meta-learning treats the learning procedure itself as an object of optimization, using experience accumulated across tasks to improve the learning of new tasks (Hospedales et al., 2021). It involves two levels of learning, either explicitly or implicitly: an inner process learns a particular task from a specified dataset and objective, while an outer process optimizes aspects of the inner process–such as the model class, initialization, or optimization rule–with the aim of reducing the data or computation required to solve new tasks.

One approach is *memory-based meta-learning* (MBML) (Santoro et al., 2016), in which a memory-augmented model–an LLM whose context serves as memory, for example–is trained on data drawn from a distribution over tasks using a single optimization loop. The two levels of learning are realized through different mechanisms: the outer process updates the model's parameters across training tasks, whereas the inner process adapts to an individual task through changes in the model's memory state. At test time, the parameters remain fixed, and adaptation proceeds through memory dynamics without task-specific parameter updates. In-context learning thus emerges as a learned adaptation procedure implemented through the model's forward computation.

This MBML framework admits a precise Bayesian characterization (Ortega et al., 2019; Xie et al., 2022). For a fixed horizon n, suppose that each training sequence is generated by first sampling a task ν from a prior distribution w, then sampling data $x_{1:n}$ from ν. Suppose a sufficiently expressive model ρ_{θ^*} is trained to minimize the population log-loss objective

$$\mathcal{L}_n(\rho_\theta) = \mathbb{E}_{\nu\sim w}\mathbb{E}_{x\sim\nu}\left[-\sum_{t=1}^{n}\log_2 \rho_\theta(x_t \mid x_{<t})\right]$$

Then ρ_{θ^*} behaves as the Bayesian mixture predictor $\xi(x) = \sum_\nu w_\nu \cdot \nu(x)$. In particular,

$$\rho_{\theta^*}(x_t \mid x_{<t}) = \xi(x_t \mid x_{<t}) = \sum_\nu w_\nu(x_{<t})\nu(x_t \mid x_{<t}),$$

where $w_\nu(x_{<t}) \propto w_\nu \cdot \nu(x_{<t})$ is the posterior probability of ν. Thus, conditioned on an observed prefix $x_{<t}$, the Bayesian predictor combines the predictions of the candidate tasks according to their posterior probabilities, reflecting both their prior weights and their consistency with the observations. The appeal of the Bayesian perspective lies in its strong prediction guarantee (Hutter, 2005): at test time, ρ_{θ^*} can achieve efficient in-context adaptation to any task μ with positive prior weight $w_\mu > 0$:

$$S_n(\mu, \rho_{\theta^*}) := \sum_{t=1}^{n}\mathbb{E}_{x_{<t}\sim\mu}\sum_{x_t}\left(\mu(x_t \mid x_{<t}) - \rho_{\theta^*}(x_t \mid x_{<t})\right)^2 = O\left(\log\frac{1}{w_\mu}\right).$$

This provides a principled explanation of in-context learning as *amortized Bayesian prediction*: training amortizes learning across tasks by encoding an adaptive inference procedure into the forward computation of a neural network. In particular, consider the class of all computable environments[1] endowed with the universal prior $w_\nu = \Theta(2^{-K(\nu)})$, where $K(\nu)$ denotes the Kolmogorov complexity of ν. The resulting *single* Bayesian predictor becomes Solomonoff universal induction (Solomonoff, 1978; 1986), which achieves cumulative expected squared prediction error $O(K(\mu))$ simultaneously in *every* computable environment μ, independently of the prediction horizon. Building on this connection, Grau-Moya et al. (2024) proposed learning increasingly general predictors by meta-training neural networks on synthetic data generated by computationally bounded environments, represented by universal Turing machines with bounded program lengths and runtimes. Progressively relaxing these bounds while increasing model capacity and training data offers an

[1] We use the terms "task" and "environment" interchangeably.

principled approach toward universal prediction, provided that the corresponding prediction objectives can be optimized sufficiently well. Genewein et al. (2026) placed this framework within the broader context of *algorithmic compression*: minimizing sequential log loss can be understood as learning to compress observations by approximating and amortizing Bayesian prediction over their underlying generative mechanisms. Under a universal prior, the resulting predictor favors simpler computational explanations while averaging over all computable hypotheses according to both their simplicity and their consistency with the observations.

Our Main Contributions. The classical theory (Solomonoff, 1978; 1986; Hutter, 2005) focuses on the exact predictor that attains the minimum population log loss. In practice, however, limited model capacity and imperfect optimization–even when training data are abundant–mean that learned models generally only approximate the ideal Bayesian predictor. As emphasized by Genewein et al. (2026), extending the Bayesian view of meta-learning beyond exact realizability and perfect optimization remains an important open direction. Motivated by this line of work, we take a first step toward understanding the fundamental gap between ideal Bayesian prediction and practical learned predictors by addressing the following central question:

To what extent do the guarantees of the ideal Bayesian predictor continue to hold when practical learning architectures can only approximate it?

In this work, we develop a theory of Bayesian and universal sequential prediction under approximation, characterizing performance and robustness across a broad range of regimes:

Environment-Wise Worst-Case Robustness. We show in Theorem 1 that every approximate predictor ρ with excess population log loss at most η, equivalently $D_{\mathrm{KL}}(\xi\|\rho) \leq \eta$, retains a finite bound on its cumulative expected squared prediction error. Specifically, for any environment μ, the error of the ideal Bayesian predictor ξ is $O(\log(1/w_\mu))$, whereas that of ρ is bounded by $O\left(\log 1/w_\mu + \eta/w_\mu\right)$. We further show in Theorem 3 that the approximation-dependent term is tight. In particular, for a universal mixture, an error of order $\eta 2^{K(\mu)}$ is attained for infinitely many computable environments μ. We also extend the analysis to two other fundamental measures of prediction error: cumulative KL divergence and cumulative squared Hellinger distance.

Typical Average-Case Robustness: In sharp contrast, we prove in Theorem 4 that, when performance is averaged over environments according to the prior w–the canonical average-case measure–the excess prediction error due to approximation is only $O(\eta)$. We show that an analogous average-case bound holds for cumulative KL divergence but not for cumulative squared Hellinger distance.

Distribution-Shift Robustness: Furthermore, we establish a distribution-shift theorem (Theorem 5) for dominated shifts from a training prior w to a test prior w', where $w'_\nu \leq C w_\nu$ for every environment ν and some $C \geq 1$. The resulting error bound is $O(\log C + C\eta)$, thereby interpolating between our average-case and worst-case guarantees.

Finite-Data Robustness: Finally, in Section 5, we study predictors learned from finitely many training sequences. We show exponential upper-tail bounds for population KL error, yet its expectation can remain infinite when unseen events receive zero probability. Nevertheless, by using a weaker alternative measure of approximation, we extend our robustness analysis to this setting. In particular, we bound the expected average excess prediction error by the representation and optimization errors plus a finite-sample term that decays exponentially with the number of training sequences.

Taken together, these results provide a broad characterization of robustness to approximation in sequential prediction. They explain the sharp contrast between strong aggregate benchmark performance and poor worst-case performance on rare environments. This sharp distinction between worst-case and average-case behavior helps explain why approximation may have a much milder effect in practice. It may also shed light on the strong empirical performance of learned universal predictors (Grau-Moya et al., 2024; Genewein et al., 2026) and, more broadly, of LLMs across a wide range of tasks and benchmarks. Our distribution-shift theorem has several implications beyond basic transfer principle for sequential prediction. It provides a quantitative rationale for targeted sampling and synthetic data generation designed to protect important but underrepresented families of tasks. In particular, it provides a theoretical rationale for expecting that augmenting real-world datasets with Universal-Turing-Machine-generated (UTM-generated) data may improve prediction performance on practically important task families, as suggested by Grau-Moya et al. (2024).

2 NOTATIONS AND PRELIMINARIES

We assume familiarity with the basic notions in probability theory and algorithmic information theory, including conditional probability and Turing machines. See, for example, (Li & Vitányi, 2008) for background. Throughout this paper, $\log$ denotes the base-2 logarithm, while $\ln$ denotes the natural logarithm. Entropy and KL divergence are measured in bits.

Sequence measures and mixtures. Let $\mathcal{X}$ be a finite alphabet with $|\mathcal{X}| \geq 2$. Suppose μ is a probability measures on $\mathcal{X}^\infty$ (representing a sequence generative model). We use $\mu(x_{1:n})$ to denote the probability that a sequence sampled from μ begins with $x_{1:n}$, and $\mu_{1:n}$ to denote the induced distribution on $\mathcal{X}^n$. We denote $x_{<t} := x_{1:t-1}$. Let $\mathcal{M} = \{\nu_1, \nu_2, \ldots\}$ be a countable class of measures. Let $\{w_i > 0\}$ be weights satisfying $\sum_i w_i = 1$. Then, define the mixture $\xi := \sum_{i=1} w_i \nu_i$. When $\mu = \nu_i$ is a selected measure from $\mathcal{M}$, we write $w_\mu := w_i$. Note that for every finite string x, $\xi(x) \geq w_\mu \mu(x)$.

The universal mixture. Solomonoff (1978; 1986) introduced a universal prior for *sequence prediction*. Unless stated otherwise, we fix a universal monotone Turing machine U throughout the paper. We say a measure ν is *estimable* if there exists a monotone Turing machine that takes an infinite sequence of independent, uniformly random bits and generates a sequence according to ν. Let $K(\nu)$ denote the length of the shortest program for ν on this machine.

Let $\mathcal{M}_U$ denote the class consisting of all estimable measures. Set

$$Z := \sum_{\nu \in \mathcal{M}_U} 2^{-K(\nu)}, \qquad w_\nu := \frac{2^{-K(\nu)}}{Z}, \qquad \xi_U := \sum_{\nu \in \mathcal{M}_U} w_\nu \nu.$$

The programs are described in a prefix-free fashion, so Kraft's inequality gives $0 < Z \leq 1$, and $w_\nu = \Theta(2^{-K(\nu)})$. The fair-coin measure belongs to $\mathcal{M}_U$, so $\xi_U(x) > 0$ for every finite string x. We also use $K(x)$ for the usual prefix-free Kolmogorov complexity of a finite string or integer x.

KL divergence. For two measures μ and ρ and $n \geq 1$, define

$$D^n_{\mathrm{KL}}(\mu \| \rho) := \sum_{x_{1:n} \in \mathcal{X}^n} \mu(x_{1:n}) \log \frac{\mu(x_{1:n})}{\rho(x_{1:n})}, \qquad D_{\mathrm{KL}}(\mu \| \rho) := \lim_{n \to \infty} D^n_{\mathrm{KL}}(\mu \| \rho).$$

We use the convention that $0 \log(0/q) = 0$ and $p \log(p/0) = +\infty$ for $p > 0$. The finite-horizon divergences are nondecreasing in n, so the supremum is also their limit. The chain rule gives that:

$$D^n_{\mathrm{KL}}(\mu \| \rho) = \sum_{t=1}^{n} \mathbb{E}_{x_{<t} \sim \mu} \left[D_{\mathrm{KL}}(\mu(\cdot \mid x_{<t}) \| \rho(\cdot \mid x_{<t})) \right]. \tag{1}$$

Cumulative prediction error. As in (Solomonoff, 1978), we use *cumulative squared prediction error* as our primary criterion for evaluating the prediction performance of ρ:

$$S_n(\mu, \rho) := \sum_{t=1}^{n} \mathbb{E}_{x_{<t} \sim \mu} \sum_{a \in \mathcal{X}} \left(\mu(a \mid x_{<t}) - \rho(a \mid x_{<t}) \right)^2, \qquad S(\mu, \rho) := \lim_{n \to \infty} S_n(\mu, \rho).$$

To examine the scope of our guarantees across different error criteria, we also consider *cumulative squared Hellinger error*:

$$\mathrm{Hel}_n(\mu, \rho) := \sum_{t=1}^{n} \mathbb{E}_{x_{<t} \sim \mu} \sum_{a \in \mathcal{X}} \left(\sqrt{\mu(a \mid x_{<t})} - \sqrt{\rho(a \mid x_{<t})} \right)^2, \qquad \mathrm{Hel}(\mu, \rho) := \lim_{n \to \infty} \mathrm{Hel}_n(\mu, \rho).$$

and *cumulative KL error*:

$$D_n(\mu, \rho) := \sum_{t=1}^{n} \mathbb{E}_{x_{<t} \sim \mu} \left[D_{\mathrm{KL}}(\mu(\cdot \mid x_{<t}) \| \rho(\cdot \mid x_{<t})) \right] = D^n_{\mathrm{KL}}(\mu \| \rho), \qquad D(\mu, \rho) := \lim_{n \to \infty} D_n(\mu, \rho).$$

The following inequalities relate these error criteria.

Lemma 1 (Theorem 3.2.3 in (Hutter et al., 2024))**.** *Let μ and ρ be probability measures on $\mathcal{X}^\infty$. For every $n \geq 1$,*

$$S_n(\mu, \rho) \leq (\ln 2)\, D_n(\mu \| \rho), \qquad \mathrm{Hel}_n(\mu, \rho) \leq (\ln 2)\, D_n(\mu \| \rho), \qquad S_n(\mu, \rho) \leq 2\, \mathrm{Hel}_n(\mu, \rho).$$

The same inequalities hold for the corresponding infinite-horizon errors S, Hel*, and D.*

3 Robustness under Approximation: Task-wise Guarantee

In this section, we establish a tight task-wise prediction guarantee for a trained model ρ that approximates the Bayesian predictor ξ with KL gap $D_{\mathrm{KL}}(\xi\|\rho) \le \eta$. The KL divergence provides the most natural measure of approximation quality because the population log-loss of ρ at horizon n admits the following decomposition, where H denotes the *Shannon entropy*:

$$\mathcal{L}_n(\rho) := \mathbb{E}_{x_{1:n}\sim\xi}\left[-\log \rho(x_{1:n})\right] = H(\xi_{1:n}) + D^n_{\mathrm{KL}}(\xi\|\rho),$$

Moreover, ξ minimizes the population loss over all probability measures on $\mathcal{X}^\infty$, attaining the minimum value $H(\xi_{1:n})$. Thus, $D_{\mathrm{KL}}(\xi\|\rho) \le \eta \implies D^n_{\mathrm{KL}}(\xi\|\rho) \le \eta \iff \mathcal{L}_n(\rho) \le \min_\pi \mathcal{L}_n(\pi) + \eta$.

3.1 Environment-Wise Guarantee: Upper Bound

We first extend the classical analysis of (Hutter, 2005) by showing that, under a finite KL-divergence, the cumulative squared prediction error remains finite for every environment in $\mathcal{M}$.

Theorem 1 (Environment-wise upper bound)**.** *If $D_{\mathrm{KL}}(\xi\|\rho) \le \eta$, then for every $\mu \in \mathcal{M}$,*

$$S(\mu,\rho) \le \frac{e\ln 2}{e-1}\left(\log\frac{1}{w_\mu} + \frac{\eta}{w_\mu}\right).$$

The proof is given in Appendix A, which bounds the cumulative KL error for every n by

$$D_n(\mu\|\rho) \le \frac{e}{e-1}\left(\log\frac{1}{w_\mu} + \frac{\eta}{w_\mu}\right).$$

Theorem 1 then follows from Lemma 1 that $S_n(\mu,\rho) \le (\ln 2)\, D_n(\mu\|\rho)$ and $\mathrm{Hel}_n(\mu,\rho) \le (\ln 2)\, D_n(\mu\|\rho)$. We thus also prove that, up to constant factors, this upper bound extends the cumulative squared Hellinger distance $\mathrm{Hel}(\mu,\rho)$.

Specializing to the universal mixture ξ_U and using the relation $w_\mu = \Theta(2^{-K(\mu)})$, we obtain the following environment-wise prediction-error bound in terms of the Kolmogorov complexity $K(\mu)$.

Theorem 2 (Upper bound for universal prediction)**.** *If $D_{\mathrm{KL}}(\xi_U\|\rho) \le \eta$, then for every $\mu \in \mathcal{M}_U$,*

$$S(\mu,\rho) = O\left(K(\mu) + \eta\cdot 2^{K(\mu)}\right),$$

where the implicit constant depends only on the fixed universal Turing machine.

3.2 Environment-Wise Guarantee: Tight Worst-Case Lower Bound

Our next theorem shows that the approximation-dependent term η/w_μ in Theorem 1 is unavoidable and tight up to a constant factor. Remarkably, for the universal mixture, this tightness holds for infinitely many environments in $\mathcal{M}_U$. Throughout this subsection, we restrict attention to the binary alphabet $\mathcal{X} = \{0,1\}$.

Theorem 3 (Tight lower bound for universal prediction)**.** *For every $\eta > 0$, there are infinitely many estimable measures μ such that*

$$\sup_{\rho:\, D_{\mathrm{KL}}(\xi_U\|\rho)\le\eta} S(\mu,\rho) \ge \Omega\left(K(\mu) + \eta\, 2^{K(\mu)}\right),$$

where the implicit constant depends only on the fixed universal Turing machine.

Thus, even an arbitrarily small KL-approximation gap can coexist with a cumulative squared prediction error that grows exponentially with the Kolmogorov complexity of the environment.

We now describe the high-level idea behind the proof. The complete proof is given in Appendix C and formalizes the following rare-prefix construction.

Rare-prefix construction. The lower bound exploits prefixes that are unlikely under the Bayesian mixture ξ, but occur with probability one under a particular environment. For $z \in \{0,1\}^\infty$, let δ_z denote the deterministic environment that outputs z. Suppose $\mu = \delta_z$ and that a finite prefix y of z has mixture probability $q := \xi(y)$. We construct ρ by replacing the next N conditional predictions along z after the prefix y with the uniform distribution, while leaving all other predictions unchanged.

Each modified history has probability at most q under ξ, and replacing its next-symbol distribution with the uniform distribution incurs at most one bit of conditional KL divergence. Hence, the total KL cost is at most qN. Under μ, however, every modified history is visited with probability one, and each modified prediction contributes exactly $1/2$ to the squared prediction error. Taking $N = \lfloor \eta/q \rfloor$ yields the following lemma.

Lemma 2 (Lower bound from a rare prefix)**.** *Let $\mu = \delta_z \in \mathcal{M}$, and let $y \prec z$ be a finite prefix of z and $q := \xi(y)$. For every $\eta > 0$, there exists a probability measure ρ with $D_{\mathrm{KL}}(\xi\|\rho) \le \eta$ such that*

$$S(\mu,\rho) \ge S_{|y|}(\mu,\xi) + \frac{1}{2}\left\lfloor \frac{\eta}{q} \right\rfloor.$$

When $q = O(w_\mu)$, it yields an $\Omega(\eta/w_\mu)$ lower bound whenever η/w_μ is sufficiently large. The proof of Theorem 3 identifies infinitely many deterministic environments having prefixes that satisfy this condition. Since $w_\mu = \Theta(2^{-K(\mu)})$, the resulting approximation penalty is $\Omega(\eta\, 2^{K(\mu)})$.

Finally, we can show that Lemma 2 also extends to cumulative KL and squared Hellinger errors. Then by Lemma 1, we have, for every $n \ge 1$, $D_n(\mu\|\rho) \ge S_n(\mu,\rho)/\ln 2$ and $\mathrm{Hel}_n(\mu,\rho) \ge S_n(\mu,\rho)/2$. Thus, our construction can establish similar lower bounds as in Theorem 3 for cumulative KL and squared Hellinger errors.

4 Robustness under Approximation: Average-Case Analysis

In this section, we show that the environment-wise prediction guarantee under approximation can be sharply improved when performance is averaged across environments. We then establish a distribution-shift theorem for moving from the training prior w to a dominated test prior w'. The result interpolates between our average-case and environment-wise guarantees. Before stating the theorems and discussing their implications, we explain the motivation for developing an average-case theory of robustness under approximation. Section 3.2 shows that the approximation-dependent term η/w_μ in Theorem 1 is unavoidable in the worst case. For an environment with a small prior weight w_μ, this term can overwhelm the classical $O(\log(1/w_\mu))$ prediction guarantee for the ideal Bayesian predictor.

Such pessimistic worst-case guarantees, however, stand in contrast to the strong empirical performance of learned universal predictors (Grau-Moya et al., 2024; Genewein et al., 2026) and, more broadly, to the outstanding performance of LLMs across a wide range of tasks and benchmarks. This contrast raises the question of why approximation has a milder effect in practice. Empirical evaluations typically aggregate performance across many tasks or environments, whereas an environment-wise guarantee treats each environment individually. Accordingly, we first study prediction performance averaged over the same prior w that defines the training mixture ξ. We show that, in sharp contrast to the worst-case setting, approximation has a substantially milder effect under this average-case criterion. Our distribution-shift theorem can then be used to extend this guarantee to other test distributions.

4.1 Sharp Contrast between Typical and Task-Wise Prediction Guarantees

Theorem 4 (Approximation robustness: average-case upper bound)**.** *If $D_{\mathrm{KL}}(\xi\|\rho) \le \eta$, then*

$$\mathbb{E}_{\mu\sim w}\left[S_n(\mu,\rho) - S_n(\mu,\xi)\right] \le (\ln 2)\,\eta.$$

This theorem is a consequence of the following exact average-case decomposition:

Proposition 1 (Exact average-case decomposition)**.** *For every predictor ρ and every $n \ge 1$,*

$$\mathbb{E}_{\mu\sim w} S_n(\mu,\rho) = \mathbb{E}_{\mu\sim w} S_n(\mu,\xi) + S_n(\xi,\rho).$$

Theorem 4 matches the exact prediction-error baseline of the ideal mixture ξ: the average cost of replacing ξ with ρ is $S_n(\xi, \rho)$. Together with Theorem 3, this result shows that approximation can preserve aggregate performance while substantially degrading performance on individual tasks.

Remark 1. *Note that Theorem 4 cannot be obtained simply by averaging the bound in Theorem 1. Instead, the proof exploits the mixture structure to identify the average excess error exactly. To see this, consider the case when $\mathcal{M}$ is infinite, the following average diverges:*

$$\mathbb{E}_{\mu\sim w}\frac{1}{w_\mu} = \sum_{\mu\in\mathcal{M}} w_\mu \frac{1}{w_\mu} = \sum_{\mu\in\mathcal{M}} 1 = +\infty.$$

A fundamental differentiation: The role of approximation in Theorem 4 also highlights an interesting distinction between exact Bayesian prediction and its approximate counterpart. The classical theory applies to cumulative squared prediction error, KL error, and Hellinger error. As we show in Section 3.1, the task-wise guarantees under approximation extend to all three error criteria. By contrast, the average-case decomposition in Proposition 1 holds only for cumulative squared prediction error and cumulative KL error. For KL error, the chain rule immediately yields the analogous decomposition. $\mathbb{E}_{\mu\sim w} D_n(\mu\|\rho) = \mathbb{E}_{\mu\sim w} D_n(\mu\|\xi) + D_n(\xi\|\rho)$. Hence, if $D_{\mathrm{KL}}(\xi\|\rho) \le \eta$, then $\mathbb{E}_{\mu\sim w} D_n(\mu\|\rho) \le \mathbb{E}_{\mu\sim w} D_n(\mu\|\xi) + \eta$. For cumulative squared Hellinger error, neither the analogous decomposition nor a uniform additive $O(\eta)$ excess-error bound relative to ξ holds in general. The mixture predictor ξ need not minimize average squared Hellinger error; a counterexample is given in Appendix E.

Remark 2. *Prior work has shown that the average gains of knowledge distillation can coexist with degraded performance on underrepresented subgroups and tail classes (Lukasik et al., 2022; Wang et al., 2023). By viewing ξ as the teacher model and ρ as the student model, our results provide a complementary explanation in Bayesian sequential prediction: a small population KL distillation loss controls the student's average excess prediction error, but can coexist with substantial degradation relative to the teacher in rare environments.*

4.2 Approximation Robustness under Distribution Shift

The preceding analysis evaluates performance under the training prior w. We now consider a different test prior w'. Our next theorem shows that the average-case guarantee remains robust when the test prior is dominated by the training prior, in the sense that $w'_\nu \le C w_\nu$ for every environment ν and some $C \ge 1$. This result has several implications beyond establishing a basic transfer principle for sequential prediction and interpolating between the worst-case and average-case analyses of robustness under approximation. One potentially significant implication is that the theory provides basic support for the argument that augmenting real-world datasets with Universal-Turing-Machine-generated (UTM-generated) data can improve prediction performance on practically important families of tasks, as suggested by Grau-Moya et al. (2024).

Theorem 5 (Approximation robustness under distribution shift)**.** *Let w' be a probability distribution on $\mathcal{M}$ satisfying $w'_\nu \le C w_\nu$ for every $\nu \in \mathcal{M}$ and some $C \ge 1$. Define the corresponding mixture*

$$\xi' := \sum_{\nu\in\mathcal{M}} w'_\nu \nu.$$

If $D_{\mathrm{KL}}(\xi\|\rho) \le \eta$, then, for every $n \ge 1$,

$$\begin{aligned}\mathbb{E}_{\mu\sim w'} S_n(\mu,\rho) &\le \mathbb{E}_{\mu\sim w'} S_n(\mu,\xi') + (\ln 2)\left(\sqrt{\log C} + \sqrt{C\eta}\right)^2 \\ &\le \mathbb{E}_{\mu\sim w'} S_n(\mu,\xi) + (\ln 2)\left(\sqrt{\log C} + \sqrt{C\eta}\right)^2.\end{aligned} \tag{2}$$

In particular, the average excess error relative to ξ' is $O(\log C + C\eta)$.

Cumulative KL error satisfies a distribution-shift bound of the same order. Indeed, the KL decomposition and the domination $\xi' \le C\xi$, together with the KL transfer inequality (Lemma 3) used in the proof of Theorem 1, give

$$\mathbb{E}_{\mu\sim w'} D_n(\mu\|\rho) \le \mathbb{E}_{\mu\sim w'} D_n(\mu\|\xi') + \frac{e}{e-1}\left(\log C + C\eta\right).$$

Implication I: Rare task families and targeted sampling. Theorem 5 interpolates between average-case and environment-wise guarantees. Taking $w' = w$ and $C = 1$ recovers the average-case upper bound in Theorem 4. At the other extreme, if w' is concentrated on a single environment μ, we may take $C = 1/w_\mu$. Then $\xi' = \mu$, and

$$S_n(\mu, \rho) \leq (\ln 2) \left(\sqrt{\log \frac{1}{w_\mu}} + \sqrt{\frac{\eta}{w_\mu}} \right)^2 = O\left(\log \frac{1}{w_\mu} + \frac{\eta}{w_\mu} \right).$$

Thus, the theorem also recovers the order of the environment-wise bound in Theorem 1.

More generally, let $A \subseteq \mathcal{M}$ be a family of environments with prior mass $w(A) > 0$, and consider evaluation under $w' = w(\cdot \mid A)$. Since $w'_\nu \leq w_\nu / w(A)$, we may take $C = 1/w(A)$ and obtain

$$\mathbb{E}_{\mu \sim w'} S_n(\mu, \rho) \leq \mathbb{E}_{\mu \sim w'} S_n(\mu, \xi') + O\left(\log \frac{1}{w(A)} + \frac{\eta}{w(A)} \right). \tag{3}$$

For this conditional test distribution, the guarantee depends on the total prior mass of the task family.

This suggests a way to improve coverage of an important but underrepresented family. Let q be a probability distribution supported on that family, and define

$$\widetilde{w} := (1 - \lambda) w + \lambda q, \qquad \xi_q := \sum_{\nu \in \mathcal{M}} q_\nu \nu, \qquad \widetilde{\xi} := \sum_{\nu \in \mathcal{M}} \widetilde{w}_\nu \nu = (1 - \lambda) \xi + \lambda \xi_q,$$

where $\lambda \in (0, 1)$ is the fraction of training mass allocated to q. Since $q_\nu \leq \widetilde{w}_\nu / \lambda$, any predictor with $D_{\mathrm{KL}}(\widetilde{\xi} \| \rho) \leq \eta$ satisfies

$$\mathbb{E}_{\mu \sim q} S_n(\mu, \rho) \leq \mathbb{E}_{\mu \sim q} S_n(\mu, \xi_q) + O\left(\log \frac{1}{\lambda} + \frac{\eta}{\lambda} \right).$$

The approximation guarantee for this family therefore depends on its assigned training mass λ, independently of its rarity under the original prior w.

Within this model, the bound provides a quantitative rationale for increasing the training probability assigned to important but underrepresented task families, for example through targeted sampling or synthetic data generation. In particular, it supports augmenting UTM-generated training data with real-world datasets to improve prediction performance on practically important task families, as suggested by Grau-Moya et al. (2024).

Implication II: Fine-tuning with replay. Replaying pretraining data during adaptation can mitigate forgetting (Béthune et al., 2025; Ibrahim et al., 2024). Theorem 5 can be used to give a population-level analysis of this strategy. Interpret q as the target-task distribution and w as the original pretraining distribution. During fine-tuning, draw a fraction λ of sequences from ξ_q and a fraction $1 - \lambda$ from ξ. The resulting population training distribution is $\widetilde{\xi} = (1 - \lambda) \xi + \lambda \xi_q$.

Suppose the fine-tuned predictor ρ_{ft} approximately minimizes the population log-loss, in that

$$D_{\mathrm{KL}}(\widetilde{\xi} \| \rho_{\mathrm{ft}}) \leq \eta.$$

Both task distributions are dominated by the fine-tuning prior: $q_\nu \leq \frac{\widetilde{w}_\nu}{\lambda}$, $w_\nu \leq \frac{\widetilde{w}_\nu}{1 - \lambda}$. Applying Theorem 5 to each distribution yields the simultaneous guarantees

$$\mathbb{E}_{\mu \sim q} S_n(\mu, \rho_{\mathrm{ft}}) \leq \mathbb{E}_{\mu \sim q} S_n(\mu, \xi_q) + O\left(\log \frac{1}{\lambda} + \frac{\eta}{\lambda} \right),$$

$$\mathbb{E}_{\mu \sim w} S_n(\mu, \rho_{\mathrm{ft}}) \leq \mathbb{E}_{\mu \sim w} S_n(\mu, \xi) + O\left(\log \frac{1}{1 - \lambda} + \frac{\eta}{1 - \lambda} \right).$$

The first bound controls prediction error on the target distribution, while the second controls error on the original distribution. Since ξ minimizes average squared prediction error under w, the second bound also limits the possible increase in average error relative to the predictor before fine-tuning.

For fixed approximation accuracy η, increasing λ tightens the target-distribution bound while weakening the original-distribution bound. The mixing parameter λ therefore controls the trade-off between the performance guarantees on the target and original task distributions. Finally, we remark that these guarantees assume that the fine-tuned predictor achieves the stated population approximation accuracy η.

Implication III: Choice of reference universal Turing machine. Let U and U' be two universal Turing machines. By the invariance theorem, there is a constant $c_{U,U'} \geq 0$ such that

$$|K_U(\nu) - K_{U'}(\nu)| \leq c_{U,U'} \qquad \text{for every estimable } \nu.$$

Consider the universal priors $w_\nu = 2^{-K_U(\nu)}/Z_U$, $w'_\nu = 2^{-K_{U'}(\nu)}/Z_{U'}$, where $Z_V := \sum_{\nu \in \mathcal{M}} 2^{-K_V(\nu)}$. Then we have $w'_\nu/w_\nu = 2^{K_U(\nu) - K_{U'}(\nu)} \cdot Z_U/Z_{U'} \leq 4^{c_{U,U'}}$. If training uses the mixture ξ induced by U and $D_{\mathrm{KL}}(\xi\|\rho) \leq \eta$, then test under the prior induced by U' satisfies

$$\mathbb{E}_{\mu \sim w'} S_n(\mu, \rho) \leq \mathbb{E}_{\mu \sim w'} S_n(\mu, \xi') + O(c_{U,U'} + 4^{c_{U,U'}}\eta)\,.$$

Thus, changing the reference machine preserves the form of the average-case guarantee, with constants depending only on the pair of machines, independently of the environment and prediction horizon. These constants may nevertheless be large because their dependence on the compiler overhead is exponential.

5 Robustness of Learning and Prediction with Finite Data

We now examine whether our prediction guarantees remain robust when the predictor is learned from finitely many samples. Fix a prediction horizon n, and let ρ_m be learned from m independent training sequences $\mathcal{D}_m \sim (\xi_{1:n})^{\otimes m}$ by minimizing the empirical log-loss over the hypothesis class $\mathcal{H}$ up to an additive optimization error $\eta_{\mathrm{opt}} \geq 0$. In addition to this optimization error, the model class incurs a representation error $\eta_{\mathrm{rep}} := \min_{\rho \in \mathcal{H}} D^n_{\mathrm{KL}}(\xi\|\rho)$. We consider $\eta_{\mathrm{rep}} + \eta_{\mathrm{opt}} > 0$.

We show an exponential upper-tail bound on the population KL error: for every $\gamma > 0$, there exist constants $c_\gamma, C_\gamma > 0$ such that[2], for all $m \geq 1$,

$$\Pr_{\mathcal{D}_m}\left\{D^n_{\mathrm{KL}}(\xi\|\rho_m) > \eta_{\mathrm{rep}} + \eta_{\mathrm{opt}} + \gamma\right\} \leq C_\gamma e^{-c_\gamma m}. \tag{4}$$

Nevertheless, this high-probability guarantee need not imply a finite expected KL error. A learned predictor may assign zero probability to an event that has positive population probability but is absent from the training data. Precisely, it is possible that

$$\mathbb{E}_{\mathcal{D}_m} D^n_{\mathrm{KL}}(\xi\|\rho_m) = +\infty \qquad \text{for every finite } m. \tag{5}$$

Despite this difficulty, we show that our squared prediction error guarantees remain meaningful in expectation. It follows from two observations. First, if we characterize the approximation quality with a weaker condition $S_n(\xi, \rho_m) \leq \eta$ instead of $D^n_{\mathrm{KL}}(\xi\|\rho_m) \leq \eta$, then there exist constants c and C such that for any $m \geq 1$,

$$\mathbb{E}_{\mathcal{D}_m} S_n(\xi, \rho_m) \leq \eta_{\mathrm{rep}} + \eta_{\mathrm{opt}} + C \cdot n e^{-c \cdot m}, \tag{6}$$

which means that the finite-sample contribution to the approximation error decays exponentially with the training sample size. Moreover, one can check that all of our theorems extend from the KL approximation condition to the weaker condition $S_n(\xi, \rho) \leq \eta$, with corresponding changes to the constants. In particular, Theorem 4 and equation 6 together imply: for any $m \geq 1$,

$$\mathbb{E}_{\mathcal{D}_m}\left[\mathbb{E}_{\mu \sim w} S_n(\mu, \rho_m) - \mathbb{E}_{\mu \sim w} S_n(\mu, \xi)\right] \leq \eta_{\mathrm{rep}} + \eta_{\mathrm{opt}} + C \cdot n e^{-c \cdot m}. \tag{7}$$

6 Conclusion

We characterize how approximation affects Bayesian prediction and Solomonoff induction. Our results show that approximation can preserve average performance while substantially degrading performance on individual tasks. The analysis also extends to dominated shifts from the training prior to the test prior and is robust in the finite-sample setting.

An important direction for future work is to develop mitigation strategies that provide stronger guarantees across all tasks simultaneously. We explore initial approaches in Appendices H and I.

Another open direction is length generalization. Our finite-sample analysis considers training and test sequences of the same length; extending these guarantees to longer test sequences remains a challenge. One possibility is to combine pretrained predictors with continual learning, using minimum description length principle (Grünwald, 2007; Poland & Hutter, 2005) to guide adaptation as new observations become available.

[2]The constants C, c in equation 4, equation 6 and equation 7 can depend on $\xi_{1:n}$ and n.

7 STATEMENTS

AI use statement In this work, we used generative AI tools for language polishing, literature search, and developing and writing all the proofs. We have reviewed all AI-assisted content and checked all the proofs. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement The work is theoretical and does not involve human subjects, private data, potentially harmful insights, methodologies and applications, or other activities that would require special ethical safeguards.

Reproducibility statement The paper does not include experiments; the main contributions are theoretical results and proofs. All assumptions are stated in the definitions and theorem statements. Complete proofs of the claims are also included in the paper.

REFERENCES

Louis Béthune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. Scaling laws for forgetting during finetuning with pretraining data injection. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/bethune25a.html.

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Tim Genewein, Jordi Grau-Moya, Li Kevin Wenliang, Laurent Orseau, and Marcus Hutter. Algorithmic compression via pretrained neural networks. *Entropy*, 28(6):596, 2026.

Jordi Grau-Moya, Tim Genewein, Marcus Hutter, Laurent Orseau, Grégoire Delétang, Elliot Catt, Anian Ruoss, Li Kevin Wenliang, Christopher Mattern, Matthew Aitchison, and Joel Veness. Learning universal predictors. In *International Conference on Machine Learning*, 2024.

Peter D Grünwald. *The minimum description length principle*. The MIT press, 2007.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169, 2021.

Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2005.

Marcus Hutter, David Quarel, and Elliot Catt. *An introduction to universal artificial intelligence*. CRC Press, 2024.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=DimPeeCxKO.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer, New York, NY, 3 edition, 2008. ISBN 978-0-387-49820-1. doi: 10.1007/978-0-387-49820-1. URL https://doi.org/10.1007/978-0-387-49820-1.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: understanding and mitigating biases in distillation. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=ph3AYXpwEb.

Pedro A. Ortega, Jane X. Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alexander Pritzel, Pablo Sprechmann, Siddhant M. Jayakumar, Tom McGrath, Kevin J. Miller, Mohammad Gheshlaghi Azar, Ian Osband, Neil C. Rabinowitz, András György, Silvia Chiappa, Simon Osindero, Yee Whye Teh, Hado van Hasselt, Nando de Freitas, Matthew M. Botvinick, and Shane Legg. Meta-learning of sequential strategies. *CoRR*, abs/1905.03030, 2019. URL http://arxiv.org/abs/1905.03030.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Jan Poland and Marcus Hutter. Asymptotics of discrete mdl for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1842–1850. JMLR.org, 2016.

Ray Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory*, 24(4):422–432, 1978.

Ray Solomonoff. The application of algorithmic probability to problems in artificial intelligence. In *Machine Intelligence and Pattern Recognition*, volume 4, pp. 473–491. 1986.

Serena Wang, Harikrishna Narasimhan, Yichen Zhou, Sara Hooker, Michal Lukasik, and Aditya Krishna Menon. Robust distillation for worst-class performance: on the interplay between teacher and student objectives. In Robin J. Evans and Ilya Shpitser (eds.), *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2237–2247. PMLR, 2023. URL https://proceedings.mlr.press/v216/wang23e.html.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

A Proof of Theorem 1

Lemma 3 (KL transfer under domination)**.** *Let P, Q, R be probability distributions on the same finite space Ω. Suppose that $Q \geq w \cdot P$ for some $w \in (0, 1]$. Then*

$$D_{\mathrm{KL}}(P\|R) \leq \frac{e}{e-1}\left(\log\frac{1}{w} + \frac{1}{w}\cdot D_{\mathrm{KL}}(Q\|R)\right).$$

Proof. The claim is immediate when $w = 1$, because then $Q = P$. We may therefore assume $0 < w < 1$. The conclusion is also immediate if $D_{\mathrm{KL}}(Q\|R) = \infty$.

Write $Q = wP + (1-w)V$ for a probability distribution V, and define

$$\alpha(x) := \frac{wP(x)}{Q(x)}$$

on the support of Q, with $\alpha(x) = 0$ elsewhere. Then $0 \leq \alpha \leq 1$ and $\mathbb{E}_Q\alpha = w$. Introduce an auxiliary binary variable Z. Under $\widetilde{Q}$, let $X \sim Q$ and

$$Z \mid X = x \sim \mathrm{Bern}(\alpha(x)).$$

Thus $\widetilde{Q}(X \in \cdot \mid Z = 1) = P$. Define $\widetilde{R}$ using the same conditional law of Z given X, but with $X \sim R$, and put

$$r := \mathbb{E}_R \alpha = \Pr_{\widetilde{R}}[Z = 1].$$

Finiteness of $D_{\mathrm{KL}}(Q\|R)$ implies $0 < r < 1$. Let R_1 denote the conditional law of X under $\widetilde{R}$ given $Z = 1$, and define R_0 analogously. One can check that

$$R_1(x) = \frac{\alpha(x)}{r} \cdot R(x), \quad R_0(x) = \frac{1 - \alpha(x)}{1 - r} \cdot R(x), \quad \text{and } R = rR_1 + (1 - r)R_0.$$

The conditional kernel from X to Z is the same under the two joint distributions. Hence

$$D_{\mathrm{KL}}(\widetilde{Q}\|\widetilde{R}) = D_{\mathrm{KL}}(Q\|R).$$

Applying the chain rule in the other direction gives

$$\begin{aligned} D_{\mathrm{KL}}(Q\|R) &= d_{\mathrm{B}}(w\|r) + wD_{\mathrm{KL}}(P\|R_1) + (1 - w)D_{\mathrm{KL}}(V\|R_0) \\ &\geq d_{\mathrm{B}}(w\|r) + wD_{\mathrm{KL}}(P\|R_1), \end{aligned} \tag{8}$$

where

$$d_{\mathrm{B}}(u\|v) := u \log \frac{u}{v} + (1 - u) \log \frac{1 - u}{1 - v}$$

is binary KL divergence. Moreover,

$$\frac{R_1(x)}{R(x)} = \frac{\alpha(x)}{r},$$

and therefore

$$\begin{aligned} D_{\mathrm{KL}}(P\|R) &= D_{\mathrm{KL}}(P\|R_1) + \mathbb{E}_P[\log \alpha] - \log r \\ &\leq D_{\mathrm{KL}}(P\|R_1) - \log r. \end{aligned}$$

Together with equation 8, this yields

$$D_{\mathrm{KL}}(P\|R) \leq \frac{D_{\mathrm{KL}}(Q\|R) - d_{\mathrm{B}}(w\|r)}{w} - \log r. \tag{9}$$

We next use the inequality

$$-\log r \leq \frac{e}{e - 1} \left(\log \frac{1}{w} + \frac{1}{w} d_{\mathrm{B}}(w\|r) \right), \qquad 0 < w, r < 1. \tag{10}$$

To verify it, first observe that

$$\log \frac{1}{w} + \frac{1}{w} d_{\mathrm{B}}(w\|r) = -\log r + \frac{1 - w}{w} \log \frac{1 - w}{1 - r}.$$

If $r \geq w$, the second term is nonnegative, and equation 10 follows. If $r < w$, the desired inequality is equivalent to

$$e(1 - w) \log \frac{1 - r}{1 - w} \leq -w \log r.$$

For fixed w, let

$$F_w(r) := -w \log r - e(1 - w) \log \frac{1 - r}{1 - w}.$$

This function is strictly convex on $(0, w)$, and its stationary point is

$$r_* := \frac{w}{e - (e - 1)w}.$$

Writing $a := 1 - w$, direct substitution gives

$$F_w(r_*) = G(a) := [1 + (e - 1)a] \log[1 + (e - 1)a] - (1 - a) \log(1 - a) - ea \log e.$$

Now $G(0) = G(1) = 0$ and

$$G'(a) = (e - 1) \log[1 + (e - 1)a] + \log(1 - a),$$

while

$$G''(a) = \frac{1}{\ln 2}\left(\frac{(e-1)^2}{1+(e-1)a} - \frac{1}{1-a}\right).$$

The numerator of G'' after taking a common denominator is affine and strictly decreasing, positive at $a = 0$ and negative at $a = 1$. Thus G' first increases and then decreases. Since $G'(0) = 0$ and $G'(a) \to -\infty$ as $a \uparrow 1$, it is initially positive and then crosses zero exactly once. Hence G first increases and then decreases to $G(1) = 0$. It follows that $G(a) \geq 0$, proving equation 10.

Let $c := e/(e-1)$. Substituting equation 10 into equation 9, and using $d_{\mathrm{B}}(w\|r) \leq D_{\mathrm{KL}}(Q\|R)$, gives

$$\begin{aligned} D_{\mathrm{KL}}(P\|R) &\leq \frac{D_{\mathrm{KL}}(Q\|R)}{w} + (c-1)\frac{d_{\mathrm{B}}(w\|r)}{w} + c\log\frac{1}{w} \\ &\leq c\left(\log\frac{1}{w} + \frac{1}{w}D_{\mathrm{KL}}(Q\|R)\right). \end{aligned}$$

□

Proof of Theorem 1. Fix n, and apply Lemma 3 to

$$P = \mu_{1:n}, \qquad Q = \xi_{1:n}, \qquad R = \rho_{1:n}.$$

Since $\xi \geq w_\mu \mu$,

$$D^n_{\mathrm{KL}}(\mu\|\rho) \leq \frac{e}{e-1}\left(\log\frac{1}{w_\mu} + \frac{D^n_{\mathrm{KL}}(\xi\|\rho)}{w_\mu}\right) \leq \frac{e}{e-1}\left(\log\frac{1}{w_\mu} + \frac{\eta}{w_\mu}\right).$$

Applying Lemma 1 and taking the supremum over n gives the conclusion. □

B Proof of Lemma 2

Let $m := |y|$ and set $N := \lfloor \eta/q \rfloor$. Note that

$$q = \xi(y) \geq w_\mu \mu(y) = w_\mu > 0.$$

For $0 \leq \ell < N$, write

$$h_\ell := z_{1:m+\ell}.$$

We construct a measure ρ on $\{0,1\}^\infty$ as follows: set

$$\rho(\cdot \mid h_\ell) := \mathrm{Bern}(1/2), \qquad 0 \leq \ell < N,$$

and let ρ agree with ξ at every other history of positive ξ-probability. At histories of zero ξ-probability, choose an arbitrary conditional distribution.

For any distribution p on $\{0,1\}$,

$$D_{\mathrm{KL}}(p\|\,\mathrm{Bern}(1/2)) = 1 - H(p) \leq 1,$$

where $H(p)$ is the binary entropy. Hence, for every n, the chain rule gives

$$D^n_{\mathrm{KL}}(\xi\|\rho) \leq \sum_{0\leq\ell<N} \xi(h_\ell) D_{\mathrm{KL}}\left(\xi(\cdot \mid h_\ell)\|\,\mathrm{Bern}(1/2)\right) \leq \sum_{0\leq\ell<N} \xi(h_\ell) \leq Nq.$$

Taking the supremum over n yields

$$D_{\mathrm{KL}}(\xi\|\rho) \leq \eta.$$

The modified conditionals occur only after the prefix y. Therefore, the first m predictions along z are unchanged, and

$$S_m(\mu,\rho) = S_m(\mu,\xi).$$

At each history h_ℓ, the next-symbol distribution under $\mu = \delta_z$ is a point mass, whereas $\rho(\cdot \mid h_\ell) = \mathrm{Bern}(1/2)$. Thus

$$(\mu(1 \mid h_\ell) - \rho(1 \mid h_\ell))^2 + (\mu(0 \mid h_\ell) - \rho(0 \mid h_\ell))^2 = \frac{1}{2}.$$

Consequently,

$$S(\mu,\rho) \geq S_{m+N}(\mu,\rho) = S_m(\mu,\xi) + \frac{N}{2}.$$

This finishes the proof.

C Proof of Theorem 3

We remark Lemma 2 already gives a nearly exponential lower bound in the universal setting. Specifically, for every n, since $\sum_{h\in\{0,1\}^n} \xi_U(h) = 1$, there is a string $h_n \in \{0,1\}^n$ such that $\xi_U(h_n) \le 2^{-n}$. Let $\mu_n := \delta_{h_n 0^\infty}$. We have

$$n - O(1) \le K(\mu_n) \le n + O(\log n).$$

The first inequality is by combining (i) $\xi_U(h_n) \le 2^{-n}$ and (ii) $\xi_U(h_n) \ge w_{\mu_n} = \Theta(2^{-K(\mu_n)})$. Lemma 2 therefore yields, for all sufficiently large n,

$$\sup_{\rho: D_{\mathrm{KL}}(\xi_U\|\rho)\le\eta} S(\mu_n,\rho) = \Omega\left(\eta \frac{2^{K(\mu_n)}}{K(\mu_n)^{O(1)}}\right).$$

This matches the upper bound in Theorem 1 up to a polynomial factor. We next show that this factor can be removed.

We will make use of the following coding theorem.

Lemma 4 (Corollary 4.3.1 in (Li & Vitányi, 2008))**.** *If P is a lower semi-computable semi-measure on $\{0,1\}^*$ with $\sum_{y\in\{0,1\}^*} P(y) \le 1$, then there is a constant $c_P = K(P) + O(1)$ such that $K(x) \le -\log P(x) + c_P$.*

Proof of Theorem 3. For $m \ge 1$, define

$$\mathcal{C}_m := \{1^m 0 s : s \in \{0,1\}^m\}, \qquad \mathcal{C} := \bigcup_{m\ge 1} \mathcal{C}_m.$$

Note that the set $\mathcal{C}$ is prefix-free. Since $|\mathcal{C}_m| = 2^m$, a counting argument gives a string $y_m \in \mathcal{C}_m$ satisfying $K(y_m) \ge m$. Here

$$K(y) := \min\{|p| : U(p) = y\}$$

is the prefix Kolmogorov complexity of the string y. Denote $\mu_m := \delta_{y_m 0^\infty}$ the deterministic measure outputting $y_m 0^\infty$. Note that μ_m is estimable, as it can hard-code the self-delimiting program generating y_m and then output zeros forever. In addition, we have $K(\mu_m) \le K(y_m) + O(1)$.

Let M_U be the Solomonoff prior

$$M_U(y) := \sum_{p: U(p)=y*} 2^{-|p|},$$

which is semicomputable. Since $\mathcal{C}$ is prefix-free, $\sum_{y\in C} M_U(y) \le 1$. By letting $P(y) = M_U(y)\cdot \mathbf{1}_{y\in C}$, Lemma 4 therefore gives

$$M_U(y) = O\left(2^{-K(y)}\right) \qquad \text{for every } y \in \mathcal{C},$$

Furthermore, a standard fact in algorithmic information theory (see e.g. Section 1.2.5 in (Hutter, 2005)) says that $\xi_U(y) = O(M_U(y))$. Thus,

$$\xi_U(y) = O\left(2^{-K(y)}\right) \qquad \text{for every } y \in \mathcal{C},$$

In particular,

$$\xi_U(y_m) = O\left(2^{-K(y_m)}\right) = O\left(2^{-K(\mu_m)}\right).$$

Fix $\eta > 0$. Applying Lemma 2 with $\mu = \mu_m$ and $y = y_m$ gives: there exists a universal constant $C_U > 0$, such that

$$\sup_{\rho:\, D_{\mathrm{KL}}(\xi_U\|\rho)\le\eta} S(\mu_m,\rho) \ge \frac{1}{2}\left\lfloor \frac{\eta 2^{K(\mu_m)}}{C_U}\right\rfloor \ge \frac{1}{4C_U}\left(K(\mu_m) + \eta 2^{K(\mu_m)}\right),$$

for infinitely many m. □

D Proof of Proposition 1

For a history h with $\xi(h) > 0$, let $w(\mu \mid h) := w_\mu \mu(h)/\xi(h)$ denote the posterior weight of μ. For any symbol $a \in \mathcal{X}$,

$$\xi(a \mid h) = \sum_{\mu \in \mathcal{M}} w(\mu \mid h)\mu(a \mid h).$$

The variance decomposition gives

$$\sum_{\mu \in \mathcal{M}} w(\mu \mid h)\big(\mu(a \mid h) - \rho(a \mid h)\big)^2 = \sum_{\mu \in \mathcal{M}} w(\mu \mid h)\big(\mu(a \mid h) - \xi(a \mid h)\big)^2 + \big(\xi(a \mid h) - \rho(a \mid h)\big)^2.$$

Multiplying by $\xi(h)$ and using $\xi(h)w(\mu \mid h) = w_\mu \mu(h)$ yields

$$\sum_{\mu \in \mathcal{M}} w_\mu \mu(h)\big(\mu(a \mid h) - \rho(a \mid h)\big)^2 = \sum_{\mu \in \mathcal{M}} w_\mu \mu(h)\big(\mu(a \mid h) - \xi(a \mid h)\big)^2 + \xi(h)\big(\xi(a \mid h) - \rho(a \mid h)\big)^2.$$

Summing over all histories $h \in \mathcal{X}^{t-1}$, all symbols $a \in \mathcal{X}$, and all $1 \le t \le n$ gives

$$\mathbb{E}_{\mu \sim w} S_n(\mu, \rho) = \mathbb{E}_{\mu \sim w} S_n(\mu, \xi) + S_n(\xi, \rho).$$

E A counterexample for squared Hellinger error

Let $\mu_1 = \delta_{1^\infty}$ and $\mu_0 = \delta_{0^\infty}$, with prior weights $w_1 = 3/4$ and $w_0 = 1/4$, respectively. Write $q_r = r\mu_1 + (1-r)\mu_0$, so that $\xi = q_{3/4}$. After the first symbol x_1, every q_r with $r \in (0,1)$ predicts the remaining sequence perfectly. Thus its averaged cumulative squared Hellinger error is

$$R(r) := \mathbb{E}_{\mu \sim w}\mathrm{Hel}(\mu, q_r) = 2 - 2\big[w_1\sqrt{r} + w_0\sqrt{1-r}\big].$$

This is minimized at

$$r_* = \frac{w_1^2}{w_1^2 + w_0^2} = \frac{9}{10} \neq \frac{3}{4}.$$

Hence ξ is not a minimizer, ruling out the analogous decomposition into its average error and a nonnegative approximation term.

Moreover, let $\rho_u = q_{3/4-u}$ for $u > 0$ sufficiently small. Since ρ_u and ξ differ only in their first-step predictions,

$$\eta_u := D_{\mathrm{KL}}(\xi \| \rho_u) = w_1 \log \frac{w_1}{w_1 - u} + w_0 \log \frac{w_0}{w_0 + u} = \Theta(u^2).$$

However,

$$R(3/4 - u) - R(3/4) = \frac{\sqrt{3} - 1}{2}\, u + O(u^2) = \Theta(\sqrt{\eta_u}).$$

Thus, no finite constant C can make $R(3/4 - u) \le R(3/4) + C\eta_u$ hold for all sufficiently small u.

F Proof of Theorem 5

By Proposition 1, we have

$$\mathbb{E}_{\mu \sim w'} S_n(\mu, \rho) = \mathbb{E}_{\mu \sim w'} S_n(\mu, \xi') + S_n(\xi', \rho), \tag{11}$$

and

$$\mathbb{E}_{\mu \sim w'} S_n(\mu, \xi) = \mathbb{E}_{\mu \sim w'} S_n(\mu, \xi') + S_n(\xi', \xi) \ge \mathbb{E}_{\mu \sim w'} S_n(\mu, \xi').$$

In the following, we will show that

$$S_n(\xi', \rho) \le (\ln 2)\left(\sqrt{\log C} + \sqrt{C\eta}\right)^2, \tag{12}$$

which concludes the proof.

The assumption $w'_\nu \le C w_\nu$ implies that, for every finite history h,

$$\xi'(h) = \sum_{\nu \in \mathcal{M}} w'_\nu \nu(h) \le C \sum_{\nu \in \mathcal{M}} w_\nu \nu(h) = C\xi(h).$$

In particular, $D_{\mathrm{KL}}^n(\xi'\|\xi) \le \log C$. Hence, by Lemma 1, we have

$$S_n(\xi', \xi) \le (\ln 2) \cdot D_{\mathrm{KL}}^n(\xi'\|\xi) \le (\ln 2) \cdot \log C. \tag{13}$$

Next, define

$$B_n := \sum_{t=1}^{n} \sum_{h \in \mathcal{X}^{t-1}} \xi'(h) \sum_{a \in \mathcal{X}} \left(\xi(a \mid h) - \rho(a \mid h)\right)^2.$$

Using $\xi'(h) \le C\xi(h)$ and Lemma 1, we obtain

$$B_n \le C \sum_{t=1}^{n} \sum_{h \in \mathcal{X}^{t-1}} \xi(h) \sum_{a \in \mathcal{X}} \left(\xi(a \mid h) - \rho(a \mid h)\right)^2 = C \cdot S_n(\xi, \rho) \le (C \ln 2) \cdot D_{\mathrm{KL}}^n(\xi\|\rho) \le (\ln 2) \cdot C\eta. \tag{14}$$

Expanding the square, we obtain

$$S_n(\xi', \rho) = S_n(\xi', \xi) + B_n + 2T_n,$$

where

$$T_n := \sum_{t=1}^{n} \sum_{h \in \mathcal{X}^{t-1}} \xi'(h) \sum_{a \in \mathcal{X}} \big(\xi'(a \mid h) - \xi(a \mid h)\big)\big(\xi(a \mid h) - \rho(a \mid h)\big).$$

By the Cauchy–Schwarz inequality,

$$|T_n| \le \sqrt{S_n(\xi', \xi) B_n}.$$

It follows that

$$\begin{aligned} S_n(\xi', \rho) &\le S_n(\xi', \xi) + B_n + 2\sqrt{S_n(\xi', \xi) B_n} \\ &= \left(\sqrt{S_n(\xi', \xi)} + \sqrt{B_n}\right)^2. \end{aligned}$$

Together with equation 13 and equation 14, this yields equation 12.

G Proofs for Section 5

Fix a horizon n. Write $P := \xi_{1:n}$, $H := H(P)$, and $a := \eta_{\mathrm{rep}} + \eta_{\mathrm{opt}}$. Let $\rho^* \in \mathcal{H}$ denote the predictor achieving $D_{\mathrm{KL}}^n(\xi\|\rho^*) = \eta_{\mathrm{rep}}$. Write $\widehat{R}_m$ for the empirical log-loss, so that $\widehat{R}_m(\rho_m) \le \widehat{R}_m(\rho^*) + \eta_{\mathrm{opt}}$.

Proof of equation 4. Let $\widehat{P}_m$ be the empirical distribution of the training sequences. For $0 < \tau < 1$, define

$$E_{m,\tau} := \left\{\exists x \in \mathrm{supp}(P) : |\widehat{P}_m(x) - P(x)| > \tau P(x)\right\}.$$

Write $d := |\mathrm{supp}(P)|$ and

$$p_{\min} := \min_{x \in \mathrm{supp}(P)} P(x) > 0,$$

Hoeffding's inequality and a union bound give

$$\Pr[E_{m,\tau}] \le 2d \exp(-2m\tau^2 p_{\min}^2).$$

Conditioned on $\overline{E}_{m,\tau}$, the nonnegativity of log-loss and approximate empirical optimality imply

$$(1-\tau)\big[H + D_{\mathrm{KL}}^n(\xi\|\rho_m)\big] \le \widehat{R}_m(\rho_m) \le (1+\tau)(H + \eta_{\mathrm{rep}}) + \eta_{\mathrm{opt}}.$$

Thus,

$$D_{\mathrm{KL}}^n(\xi\|\rho_m) \le a + \frac{\tau}{1-\tau}\big[2(H + \eta_{\mathrm{rep}}) + \eta_{\mathrm{opt}}\big].$$

For any $\gamma > 0$, choose $\tau > 0$ small enough that the second term is at most γ. The claim follows with $C_\gamma = 2d$ and $c_\gamma = 2\tau^2 p_{\min}^2$.

Example establishing equation 5. Take $n = 1$, $P = \text{Bernoulli}(1/2)$, and $\mathcal{H}$ consisting of $\rho_0 := \delta_0$ and $\rho_1 := \text{Bernoulli}(1/4)$. On the event that all m training observations are zero, which has probability 2^{-m}, exact empirical log-loss minimization selects $\rho_m = \rho_0$. Since $D_{\mathrm{KL}}(P\|\rho_0) = +\infty$, the expected KL error is infinite for every finite m. Padding deterministic zeros extends the example to any fixed n.

Proof of equation 6. Recall $a = \eta_{\text{rep}} + \eta_{\text{opt}}$, let $\gamma_a := (1/\ln 2 - 1)a$. By $0 \le S_n \le 2n$ and $S_n(\xi, \rho_m) \le (\ln 2) D^n_{\mathrm{KL}}(\xi\|\rho_m)$, we have

$$\mathbb{E}_{\mathcal{D}_m}\left[(S_n(\xi,\rho_m) - a)_+\right] \le 2n \Pr\{D^n_{\mathrm{KL}}(\xi\|\rho_m) > a + \gamma_a\} \le 2nC_{\gamma_a} e^{-c_{\gamma_a} m}.$$

Theorems 4, 5, and 1 under squared-error approximation. Suppose $S_n(\xi, \rho) \le \eta$. By Proposition 1, we have

$$\mathbb{E}_{\mu\sim w} S_n(\mu, \rho) = \mathbb{E}_{\mu\sim w} S_n(\mu, \xi) + S_n(\xi, \rho) \le \mathbb{E}_{\mu\sim w} S_n(\mu, \xi) + \eta.$$

In the proof of Theorem 5, by almost the same argument except adjusting equation 14 to $B_n \le C\eta$, we will get

$$\mathbb{E}_{\mu\sim w'} S_n(\mu, \rho) \le \mathbb{E}_{\mu\sim w'} S_n(\mu, \xi') + \left(\sqrt{(\ln 2)\log C} + \sqrt{C\eta}\right)^2.$$

Letting w' concentrate on a single environment μ, we recover Theorem 1, up to constants:

$$S_n(\mu, \rho) \le \left(\sqrt{(\ln 2)\log\frac{1}{w_\mu}} + \sqrt{\frac{\eta}{w_\mu}}\right)^2 = O\left(\log\frac{1}{w_\mu} + \frac{\eta}{w_\mu}\right).$$

Theorem 3 under squared-error approximation. By Lemma 1,

$$\{\rho : D_{\mathrm{KL}}(\xi_U\|\rho) \le \eta/\ln 2\} \subseteq \{\rho : S(\xi_U, \rho) \le \eta\}.$$

Thus Theorem 3, applied with budget $\eta/\ln 2$, implies the same infinite-horizon lower bound under squared-error approximation, up to constants.

H Mitigation With Category Information

The worst-case bounds of order $\Theta(\eta/w_\mu)$ in Theorem 1 arise when approximation is controlled only through a single aggregate KL bound $D_{\mathrm{KL}}(\xi\|\rho) \le \eta$. When available, training-time task category information enables separate control of approximation errors within each task category, potentially improving category-wise prediction performance. Notably, the learned predictor does not require category tags at test time.

Let $\mathcal{T} = \{T_1, \ldots, T_g\}$ be a finite set of categories, with $\text{tag} : \mathcal{M} \to \mathcal{T}$ defining the partition $\mathcal{M}_j := \{\mu \in \mathcal{M} : \text{tag}(\mu) = T_j\}$. Write

$$p_j := \sum_{\mu\in\mathcal{M}_j} w_\mu, \qquad \xi_j := \sum_{\mu\in\mathcal{M}_j} \frac{w_\mu}{p_j}\mu.$$

Thus $\xi = \sum_{j=1}^g p_j\xi_j$. Recall that, by Theorem 1, an aggregate approximation guarantee $D_{\mathrm{KL}}(\xi\|\rho) \le \eta$ gives

$$S_n(\xi_j, \rho) = O\left(\log\frac{1}{p_j} + \frac{\eta}{p_j}\right).$$

Direct control of the category-wise KL error avoids this inverse-category-mass amplification. The following guarantee follows immediately from Lemma 1 and Proposition 1.

Proposition 2 (Category-wise approximation)**.** *If $D^n_{\mathrm{KL}}(\xi_j\|\rho) \le \eta_j$, then*

$$\mathbb{E}_{\mu\sim w(\cdot|j)} S_n(\mu, \rho) - \mathbb{E}_{\mu\sim w(\cdot|j)} S_n(\mu, \xi_j) = S_n(\xi_j, \rho) \le (\ln 2)\eta_j.$$

In particular, if the tag identifies the generating task and $D^n_{\mathrm{KL}}(\xi_\mu\|\rho) \le \eta_\mu$, then

$$S_n(\mu, \rho) = O(\eta_\mu).$$

To control the category-wise approximation errors $D^n_{\mathrm{KL}}(\xi_j\|\rho)$, one may minimize the following worst-category excess log-loss, or a suitable surrogate:

$$\min_\rho \max_{1\le j\le g} D^n_{\mathrm{KL}}(\xi_j\|\rho).$$

I HIGHER MOMENTS OF SEQUENCE LOG-LOSS

One may instead minimize the following sample-only training objective:

$$\mathcal{L}_p(\rho) := \mathbb{E}_{x\sim\xi}\left[\big(-\log\rho(x)\big)^p\right], \quad p \geq 2.$$

Fix a horizon n. Let q be the minimizer of $\mathcal{L}_p$ among all probability distributions on $\mathcal{X}^n$. Suppose the predictor ρ satisfies

$$\mathcal{L}_p(\rho_{1:n}) \leq \mathcal{L}_p(q) + \eta.$$

Lemma 5. *For every $\mu \in \mathcal{M}$,*

$$S_n(\mu,\rho) = O_{p,|\mathcal{X}|}\left(\log\frac{1}{w_\mu} + \log(n+1) + \left(\frac{\eta}{w_\mu}\right)^{1/p}\right).$$

Here, the implicit constant in $O_{p,|\mathcal{X}|}(\cdot)$ may depend on p and $|\mathcal{X}|$, but not on any other parameters.

Proof. Let $\Omega_n := \{x \in \mathcal{X}^n : \xi(x) > 0\}$ be the support of $\xi_{1:n}$. Since $\xi \geq w_\mu\mu$, the distribution $\mu_{1:n}$ is also supported on Ω_n. The minimizer q assigns all its mass to Ω_n, since otherwise, renormalizing its restriction to Ω_n would strictly decrease the objective. If $|\Omega_n| = 1$, then the result follows immediately from

$$\big(-\log\rho(x)\big)^p \leq \eta$$

on the unique point $x \in \Omega_n$. We henceforth assume $|\Omega_n| \geq 2$. In this case, $q(x) \in (0,1)$ for every $x \in \Omega_n$.

Write $\ell(x) := -\log q(x)$. Consider the Lagrangian

$$\mathscr{L}(q,\lambda) := \sum_{x\in\Omega_n}\xi(x)\big(-\log_2 q(x)\big)^p + \lambda\left(\sum_{x\in\Omega_n} q(x) - 1\right).$$

Since the minimizer is an interior point, the first-order condition gives

$$\frac{\partial\mathscr{L}}{\partial q(x)} = -\frac{p\,\xi(x)\ell(x)^{p-1}}{q(x)\ln 2} + \lambda = 0, \qquad x \in \Omega_n.$$

Thus, setting $c_n := \lambda\ln 2/p$, we obtain

$$\xi(x)\ell(x)^{p-1} = c_n q(x), \qquad x \in \Omega_n. \tag{15}$$

Since $\xi(x) > 0$, $\ell(x) > 0$, and $q(x) > 0$, we also have $c_n > 0$. Summing equation 15 over x gives

$$c_n = \mathbb{E}_{x\sim\xi}\left[\ell(x)^{p-1}\right].$$

Since $u \mapsto u^{(p-1)/p}$ is concave, Jensen's inequality gives

$$c_n \leq \left(\mathbb{E}_{x\sim\xi}\left[\ell(x)^p\right]\right)^{(p-1)/p} = \mathcal{L}_p(q)^{(p-1)/p} \leq \mathcal{L}_p\big(\mathrm{Unif}(\Omega_n)\big)^{(p-1)/p} \leq \big(n\log|\mathcal{X}|\big)^{p-1}.$$

We next show that q dominates ξ. If $\ell(x) \geq 1$, then equation 15 gives

$$\frac{q(x)}{\xi(x)} = \frac{\ell(x)^{p-1}}{c_n} \geq \frac{1}{\big(n\log|\mathcal{X}|\big)^{p-1}}.$$

If $\ell(x) < 1$, then $q(x) > 1/2$, and hence

$$\frac{q(x)}{\xi(x)} \geq q(x) > \frac{1}{2},$$

The two cases imply

$$q(x) \geq \frac{\xi(x)}{C_{n,p}}, \qquad C_{n,p} := 2\big(n\log|\mathcal{X}|\big)^{p-1}. \tag{16}$$

We now use the approximate optimality of ρ. Define $z(x) := \log(q(x)/\rho(x))$. Since $\mathcal{L}_p(\rho_{1:n}) < \infty$, it has that $\rho(x) > 0$ on Ω_n, and thus

$$-\log\rho(x) = \ell(x) + z(x) \geq 0.$$

For $p \geq 2$ and $u, v \geq 0$, we have

$$v^p \geq u^p + pu^{p-1}(v-u) + |v-u|^p. \tag{17}$$

Indeed, when $v = u + s$ with $s \geq 0$, subtracting the right-hand side from the left-hand side gives a function $F_+(s)$ satisfying $F_+(0) = 0$ and

$$F'_+(s) = p\left((u+s)^{p-1} - u^{p-1} - s^{p-1}\right) \geq 0.$$

When $v = u - s$ with $0 \leq s \leq u$, the corresponding function satisfies $F_-(0) = 0$ and

$$F'_-(s) = p\left(u^{p-1} - (u-s)^{p-1} - s^{p-1}\right) \geq 0.$$

Applying equation 17 with

$$u = \ell(x), \qquad v = -\log \rho(x) = \ell(x) + z(x),$$

we obtain

$$\mathcal{L}_p(\rho) - \mathcal{L}_p(q) = \mathbb{E}_{x\sim\xi}\left[(\ell(x)+z(x))^p - \ell(x)^p\right] \geq \mathbb{E}_{x\sim\xi}\left[|z(x)|^p\right] + p\mathbb{E}_{x\sim\xi}\left[\ell(x)^{p-1}z(x)\right].$$

By equation 15,

$$\mathbb{E}_{x\sim\xi}\left[\ell(x)^{p-1}z(x)\right] = c_n \sum_{x\in\Omega_n} q(x)\log\frac{q(x)}{\rho(x)} = c_n D_{\mathrm{KL}}(q\|\rho_{1:n}) \geq 0.$$

It follows from the approximation guarantee that

$$\mathbb{E}_{x\sim\xi}\left[\left|\log\frac{q(x)}{\rho(x)}\right|^p\right] \leq \eta. \tag{18}$$

Since $\xi(x) \geq w_\mu \mu(x)$, equation 16 and equation 18 imply

$$q(x) \geq \frac{w_\mu}{C_{n,p}}\mu(x)$$

and

$$\mathbb{E}_{x\sim\mu}\left[|z(x)|^p\right] \leq \frac{\eta}{w_\mu}.$$

Therefore, by Holder's inequality,

$$\begin{aligned} D^n_{\mathrm{KL}}(\mu\|\rho) &= \mathbb{E}_{x\sim\mu}\left[\log\frac{\mu(x)}{q(x)} + z(x)\right] \leq \log\frac{C_{n,p}}{w_\mu} + \mathbb{E}_{x\sim\mu}|z(x)| \leq \log\frac{C_{n,p}}{w_\mu} + \left(\frac{\eta}{w_\mu}\right)^{1/p} \\ &= O_{p,|\mathcal{X}|}\left(\log\frac{1}{w_\mu} + \log(n+1) + \left(\frac{\eta}{w_\mu}\right)^{1/p}\right). \end{aligned}$$

By applying Lemma 1, we obtain the lemma. □