Prediction Under Imperfect Compression: A Theory of Approximate MDL

Qian Li[*]
SICIAM, Shenzhen Research Institute of Big Data
liqian.ict@gmail.com

Xinyu Mao[*]
University of Southern California
xinyumao.tcs@gmail.com

Shang-Hua Teng[*]
University of Southern California
shanghua@usc.edu

Guangxu Yang[*]
University of Southern California
guangxuy@usc.edu

Abstract

Minimum Description Length (MDL) is a framework for inductive inference and learning that formalizes Occam's razor as a compression principle based on minimizing the total description length: $L(\text{model}) + L(\text{data} \mid \text{model})$. In sequential prediction, MDL selects, at each step, a model that minimizes this objective on the observed prefix and uses it to predict the next outcome. Classical MDL theory shows that exact optimization yields a strong prediction guarantee: *the cumulative expected squared prediction error is finite*. In practice, however, machine-learning methods often rely on approximate optimization. To bridge this gap, we address the following fundamental question: *Under what conditions on approximation and regularization does MDL continue to guarantee reliable sequential prediction?* Our work characterizes when approximate MDL retains this finite-error guarantee, revealing a sharp distinction between additive and multiplicative approximations. For any additive approximation with slack $C \geq 0$, we prove that the cumulative expected squared prediction error is at most $16 \cdot 2^C$ times the classical bound for exact MDL. More generally, for the *balanced* MDL objective $\lambda \cdot L(\text{model}) + L(\text{data} \mid \text{model})$ with regularization parameter λ, we show that this factor generalizes to $16\lambda \cdot 2^{C/\lambda}$ when $\lambda \geq 1$. We complement these upper bounds with nearly matching lower bounds, showing that their dependence on C is essentially tight. Thus, the prediction guarantee is robust to any fixed additive optimization slack. We further show that the model-complexity regularization threshold $\lambda = 1$ is sharp: when $0 < \lambda < 1$, overfitting can produce infinite cumulative error over the universal class of estimable measures. Under multiplicative approximation, no regularization parameter $\lambda > 0$ provides a finite-error guarantee over the universal class. Thus, additive control of the optimization gap is sufficient—and, in this sense, essential—for retaining a finite-error guarantee.

1 Introduction

The Minimum Description Length (MDL) principle formalizes Occam's razor by selecting a model that minimizes the combined description length of the model itself and the observed data encoded under the model (Rissanen, 1978, 1998; Grünwald, 2007). MDL has become a central framework connecting model selection, universal coding, and prediction (Grünwald, 2007; Vitányi and Li, 2000). In sequential prediction over a family $\mathcal{M}$ of probabilistic generative models, MDL takes a particularly natural form: after observing a history $x_{<t}$, the learner selects a model $\nu \in \mathcal{M}$ that balances model

[*]The authors are listed in alphabetical order.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

complexity against how efficiently the model encodes the observed prefix $x_{<t}$, and then uses the conditional distribution $\nu(\cdot \mid x_{<t})$ to predict the next symbol.

This MDL objective takes the form $L(\text{model}) + L(\text{data} \mid \text{model})$, where $L()$ denotes the description length function of models. More generally, one may introduce a regularization parameter $\lambda > 0$ and consider objectives of the form $\lambda \cdot L(\text{model}) + L(\text{data} \mid \text{model})$. The theory of MDL under exact optimization is well developed. In the realizable setting with a countable model class, Poland and Hutter (2005) analyzed discrete MDL for online prediction and established finite bounds on the cumulative quadratic and Hellinger losses, together with almost-sure convergence. Hutter (2009) further showed that, for any countable model class containing the true measure, the distributions selected by discrete MDL eventually makes essentially the same predictions as the true environment in total variation. More recently, Milovanov (2024) proved that a universal predictor based on exact MDL achieves finite cumulative squared prediction error along every infinite Martin-Löf random sequence under sufficiently strong regularization. Taken together, these results formalize the classical exact-optimization guarantee underlying MDL prediction: exact minimization of a compression-based objective yields reliable sequential prediction.

Despite these theoretical guarantees, exact optimization is often unrealistic in modern machine learning. In practice, identifying the model that achieves the best compression requires solving a model-selection or training problem that may be computationally difficult or even intractable, often necessitating approximate solutions. This gap between ideal theory and computational reality presents a fundamental challenge. Minsky (2010), for example, highlighted both the appeal of algorithmic probability and the need for practical approximation:

> *It seems to me that the most important discovery since Gödel was the discovery by Chaitin, Solomonoff, and Kolmogorov of the concept called Algorithmic Probability... it is a beautiful theory... but it's got one problem, that is, that you cannot actually calculate what this theory predicts because it is too hard, it requires an infinite amount of work. However, it should be possible to make practical approximations... that would make better predictions than anything we have today.*

1.1 Our contributions

Minsky's call for practical approximations, together with the computational barriers to exact MDL optimization, motivates the central question of this paper:

Under what conditions on approximation and regularization does MDL continue to guarantee reliable sequential prediction?

Our work provides a principled characterization of when approximate MDL still guarantees finite cumulative expected squared prediction error, revealing a sharp distinction between additive and multiplicative approximations.

Formally, we consider realizable sequential prediction over a finite alphabet $\mathcal{X}$ of symbols (or tokens) and a countable class $\mathcal{M}$ of probability measures on $\mathcal{X}^\infty$, with the data generated by a true environment in $\mu \in \mathcal{M}$. We refer to each measure $\nu \in \mathcal{M}$ as an *environment*. For every $t \geq 1$, ν induces a distribution on $\mathcal{X}^{t-1}$; for $x_{<t} \in \mathcal{X}^{t-1}$, we write $\nu(x_{<t})$ for the probability that $x_{<t}$ forms the first $t-1$ tokens of the generated sequence. Whenever $\nu(x_{<t}) > 0$, the environment also defines the next-token distribution $\nu(\cdot \mid x_{<t})$. Following the standard exact-MDL prediction framework, assign positive weights $(w_\nu)_{\nu \in \mathcal{M}}$ satisfying

$$\sum_{\nu \in \mathcal{M}} w_\nu \leq 1.$$

These weights define a possibly subnormalized *model prior*, encoding an *inductive bias* over $\mathcal{M}$, and induce the ideal model description lengths $L_w(\nu) := -\log_2 w_\nu$.

An important example of a countable model class is the *universal class of estimable measures*, as studied in algorithmic probability theory and Solomonoff induction. Conventionally defined over the binary alphabet $\mathcal{X} = \{0,1\}$, this universal class $\mathcal{M}_U$ consists of all estimable probability measures on $\{0,1\}^\infty$ and is equipped with the natural algorithmic weights $w_\nu^U := 2^{-K(\nu)}$, $\nu \in \mathcal{M}_U$ where $K(\nu)$ denotes the *prefix Kolmogorov complexity* of the measure ν (see Definition 2.3 for the formal definition).

Each model ν can be represented using a prefix-free binary codeword of length at most $\lceil -\log_2 w_\nu \rceil$ (by the Kraft inequality). Similarly, the likelihood $\nu(x_{<t})$ induces the ideal data description length $L(x_{<t} \mid \nu) := -\log_2 \nu(x_{<t})$ with the convention that $-\log_2 \nu(x) = \infty$ if $\nu(x) = 0$. For any fixed t and model ν, Shannon-Fano coding represents each string $x_{<t} \in \mathcal{X}^{t-1}$ by a prefix-free binary codeword of length at most $\lceil -\log_2 \nu(x_{<t}) \rceil$.

Combining these two description lengths, the regularized MDL score of a model ν on the observed prefix $x_{<t}$ is then defined as:

$$L_{\lambda,w}(\nu; x_{<t}) := \lambda L_w(\nu) + L(x_{<t} \mid \nu) = \lambda \log_2 \frac{1}{w_\nu} + \log_2 \frac{1}{\nu(x_{<t})} \tag{1}$$

where $\lambda > 0$ controls the strength of model-complexity regularization. The standard MDL score corresponds to $\lambda = 1$, whereas $\lambda > 1$ imposes stronger model-complexity regularization.

Under *additive MDL approximation* with slack $C \geq 0$, at each time step $t = 1, 2, \ldots$, the predictor selects a model $\tilde{\rho}_t^+ \in \mathcal{M}$ whose regularized MDL score is within C of the minimum:

$$L_{\lambda,w}(\tilde{\rho}_t^+; x_{<t}) \leq \inf_{\nu \in \mathcal{M}} L_{\lambda,w}(\nu; x_{<t}) + C. \tag{2}$$

It then predicts x_t according to $\tilde{\rho}_t^+(\cdot \mid x_{<t})$. Relative to the true environment $\mu \in \mathcal{M}$ and its next-token distribution $\mu(\cdot \mid x_{<t})$, the squared prediction error at time t is

$$\sum_{a \in \mathcal{X}} \left(\tilde{\rho}_t^+(a \mid x_{<t}) - \mu(a \mid x_{<t})\right)^2 .$$

Similarly, under *multiplicative MDL approximation* with parameter $\varepsilon \geq 0$, the predictor selects a sequential predication model for $\tilde{\rho}_t^\times \in \mathcal{M}$ whose regularized MDL score satisfies:

$$L_{\lambda,w}(\tilde{\rho}_t^\times; x_{<t}) \leq (1 + \varepsilon) \inf_{\nu \in \mathcal{M}} L_{\lambda,w}(\nu; x_{<t}). \tag{3}$$

Exact MDL is recovered by setting $C = 0$ in the additive approximation or $\varepsilon = 0$ in the multiplicative approximation. In what follows, we denote the model selected by exact MDL at time t by $\hat{\rho}_t$. Classical exact-MDL theory (Poland and Hutter, 2005) shows that, for any true environment $\mu \in \mathcal{M}$, the predictor induced by the model sequence $\hat{\rho} = (\, \hat{\rho}_t \,)_{t \geq 1}$ has cumulative expected squared prediction error $O(w_\mu^{-1})$, and hence finite.

Summary of main results. Understanding whether the classical prediction guarantee is robust to approximation is therefore fundamental. In particular, we ask to what extent, and for which values of λ, approximate MDL retains the classical compression-to-prediction guarantee.

In this paper, we develop a theory of approximate MDL that addresses this question by characterizing prediction robustness across all regularization parameters $\lambda > 0$ under additive approximation and establishing a sharp separation between additive and multiplicative approximations.

- *Robustness under Additive Approximation*: We prove that, for any countable model class $\mathcal{M}$, true environment $\mu \in \mathcal{M}$, additive slack $C \geq 0$, and regularization parameter $\lambda \geq 1$, the predictor induced by the model sequence $\tilde{\rho}^+ = (\, \tilde{\rho}_t^+ \,)_{t \geq 1}$ has cumulative expected squared prediction error $O\left(\lambda 2^{C/\lambda} w_\mu^{-1}\right)$, and hence finite. Thus, MDL is robust to additive approximation. Specializing to the universal class $\mathcal{M}_U$ of estimable measures, this result establishes that additively approximate MDL remains a universal predictor. Furthermore, we complement these upper bounds with nearly matching lower bounds, showing that their dependence on the approximation slack is nearly tight. As a byproduct, our result also closes a gap in the exact-MDL setting: setting $C = 0$ yields finite cumulative expected squared prediction error for all $\lambda \geq 1$, including the previously unresolved range $\lambda \in (1, 2]$ left open by Milovanov (2024).
- *Model-Complexity Regularization Threshold*: We prove that, for the universal class $\mathcal{M}_U$ with its natural Kolmogorov-complexity-based weights, overfitting can lead to non-vanishing expected one-step squared prediction error, and hence infinite cumulative expected squared prediction error, when $0 < \lambda < 1$. Thus, the model-complexity regularization threshold $\lambda = 1$ is sharp, demonstrating that sufficiently strong model-complexity regularization is necessary for robust and reliable sequential prediction.

- *Sharp Distinction between Additive and Multiplicative Approximations*: We prove that, for the universal class $\mathcal{M}_U$ with natural weights $w_\nu = 2^{-K(\nu)}$, for every $\lambda > 0$ and $\epsilon > 0$, there exists a true environment $\mu \in \mathcal{M}_U$ and a valid $(1+\epsilon)$-multiplicative MDL approximation $\tilde{\rho}^\times = (\tilde{\rho}_t^\times)_{t\geq 1}$ whose induced predictor has non-vanishing per-step expected squared prediction error and hence infinite cumulative expected squared prediction error. Thus, multiplicative approximation fails to guarantee even vanishing expected one-step prediction error, regardless of the regularization strength $\lambda > 0$.

The key mathematical distinction is that additive and multiplicative approximations induce fundamentally different effective slack. An additive slack C remains fixed as the observed prefix grows and can therefore be controlled when $\lambda \geq 1$, with stronger regularization further mitigating its effect. In contrast, a multiplicative slack permits an additive error of order $\epsilon \cdot \mathrm{OPT}_\lambda(x_{<t})$, which can grow linearly with t in stochastic environments. This growing tolerance can allow systematically incorrect models to remain admissible indefinitely. The following table summarizes our prediction guarantees for universal induction alongside the relevant classical MDL results.

Setting / Approximation	**Regularization strength (λ)**	**Cumulative Error**
Exact (Poland and Hutter, 2005)	$\lambda = 1$ (Standard)	Finite
Exact (Milovanov, 2024)	$\lambda > 2$	Finite
Additive Approximation (Theorem 3.1)	$\lambda \geq 1$	Finite
Additive Approximation (Theorem 3.2)	$0 < \lambda < 1$	$+\infty$
Multiplicative Approximation (Theorem 4.1)	$\lambda > 0$	$+\infty$

1.2 Related work

The study of MDL and its connections to the broader problem of universal induction has a rich history (Solomonoff, 1964a,b; Rissanen, 1978; Vitányi and Li, 2000; Grünwald, 2007; Wallace and Dowe, 1999). Our work is most directly related to the literature on exact MDL prediction. Poland and Hutter (2005) analyze discrete MDL predictors for online prediction, while Hutter (2009) proves that, for any countable model class containing the true measure, the models selected by MDL asymptotically predict the true measure in total variation. Our positive result, Theorem 3.1, extends this line of exact-selection theory to additive approximate minimization. The closest technical precursor is the best-explanation predictor of Milovanov (2024), for which exact minimization of a complexity-regularized log-loss objective with sufficiently strong regularization yields finite cumulative squared prediction error. Our theorem shows that this phenomenon persists under additive approximation. In contrast, Theorem 4.1 shows that it does not extend to a uniform guarantee over all valid multiplicative approximate selectors.

A second relevant line is algorithmic statistics and best-explanation modeling. Two-part descriptions have long been used to formalize individual explanations through sufficient statistics and structure functions (Vitányi and Li, 2000; Gács et al., 2001; Vereshchagin and Vitányi, 2004). This viewpoint is especially relevant for our negative results: when the regularization parameter is small, an approximate MDL selector can explain the observed prefix by encoding the sample itself into the model description, which corresponds to a form of memorization or pointwise overfitting. Closely related in spirit is Hutter's analysis of monotone-complexity prediction (Hutter, 2003), which showed that universal deterministic or one-part MDL can succeed in deterministic estimable environments yet fail in probabilistic environments. Our additive counterexample for $0 < \lambda < 1$ (Theorem 3.2) exhibits a similar overfitting mechanism inside an approximate two-part MDL.

A third line concerns approximation and surrogate MDL objectives. Adriaans and Vitányi (2009) showed that along a sequence of shorter approximate two-part MDL codes, the corresponding models need not improve monotonically in goodness of fit. Conceptually, our results may be viewed as a sequential-prediction analogue of this warning: approximate improvement in description length is not, by itself, enough to guarantee reliable prediction. Grünwald and de Rooij (2005) study prequential maximum-likelihood codes and related universal coding criteria under misspecification. Their setting differs from ours: we keep the MDL objective itself fixed and vary only the optimization accuracy, measured either additively or multiplicatively at each prediction step.

Finally, MDL-style reasoning also appears in modern machine learning through description-length regularization, neural-network compression, and prequential coding (Hinton and van Camp, 1993; Blier and Ollivier, 2018; Bornschein et al., 2023; Delétang et al., 2024). Although these works do not address the selector-level robustness question studied here, their broader compression perspective motivates the study of approximate compression-based prediction and model selection, as pursued both in this paper and in recent work on the robustness of approximate Solomonoff universal induction (Li and Teng, 2026). We emphasize that Kolmogorov complexity is uncomputable, making universal MDL an idealized prediction principle. Allowing optimization error at each prediction step does not by itself remove this computational obstacle. A complementary approach is to use *time-bounded Kolmogorov complexity*: Given a prescribed computable time bound τ, define $K^\tau(x)$ as the length of a shortest program that outputs x within $\tau(|x|)$ computation steps. Unlike unrestricted Kolmogorov complexity, K^τ is computable by finite exhaustive search, although computing it need not be efficient. Pursuing this direction, Hirahara and Nanashima (2026) developed an "efficiently computable version" of Solomonoff induction.

2 Preliminaries: sequential prediction and universal induction

Let $\mathcal{X}$ be a finite alphabet. For $n \geq 0$, let $\mathcal{X}^n$ denote the set of strings of length n over $\mathcal{X}$, with $\mathcal{X}^0$ consisting of the empty string. Let

$$\mathcal{X}^* := \bigcup_{n \geq 0} \mathcal{X}^n$$

be the set of all finite strings over $\mathcal{X}$. For $x \in \mathcal{X}^*$, let $\ell(x)$ denote its length, namely, the unique integer $n \geq 0$ such that $x \in \mathcal{X}^n$. If $\ell(x) = n \geq 1$, we write

$$x = x_1 \cdots x_n, \qquad x_{1:n} = x, \qquad x_{<t} := x_1 \cdots x_{t-1} \quad (1 \leq t \leq n+1).$$

For $x \in \mathcal{X}^*$, define the cylinder set

$$[x] := \{\omega \in \mathcal{X}^\infty : \omega_{1:\ell(x)} = x\}.$$

By convention, the cylinder corresponding to the empty string is $\mathcal{X}^\infty$.

Let ν be a probability measure on $\mathcal{X}^\infty$, where $\mathcal{X}^\infty$ is equipped with the σ-algebra generated by the cylinder sets. We identify ν with its induced prefix-probability function on $\mathcal{X}^*$, writing $\nu(x) := \nu([x])$. These prefix probabilities satisfy the consistency condition

$$\nu(x) = \sum_{a \in \mathcal{X}} \nu(xa) \qquad \text{for all } x \in \mathcal{X}^*.$$

The prefix probability of the empty string is 1. Whenever $\nu(x) > 0$, the measure ν induces the *next-token probabilities*

$$\nu(a \mid x) := \frac{\nu(xa)}{\nu(x)}, \qquad a \in \mathcal{X}.$$

Sequential prediction. Let μ be a probability measure on $\mathcal{X}^\infty$ that generates a random infinite sequence

$$X_{1:\infty} = X_1 X_2 \cdots \sim \mu.$$

The measure μ is unknown to the predictor. At each time $t \geq 1$, after observing a realization

$$X_{<t} = x_{<t} := x_1 x_2 \cdots x_{t-1},$$

the predictor seeks to predict the next token X_t.

More precisely, whenever $\mu(x_{<t}) > 0$, the prediction target is the conditional distribution $\mu(\cdot \mid x_{<t})$, or, equivalently, the next-token probabilities

$$\mu(a \mid x_{<t}) := \frac{\mu(x_{<t}a)}{\mu(x_{<t})}, \qquad a \in \mathcal{X}.$$

A sequential predictor $\rho = (\rho_t)_{t \geq 1}$ is a sequence of next-token predictors, where, at each time t, ρ_t assigns to every history $x_{<t} \in \mathcal{X}^{t-1}$ a probability distribution $\rho_t(\cdot \mid x_{<t})$ over $\mathcal{X}$.

Cumulative expected squared prediction error. Following classical MDL theory, we measure the predictive performance of a sequential predictor $\rho = (\rho_t)_{t\geq 1}$ relative to the true measure μ based on the complete sequence of expected one-step squared prediction errors. At time t, define

$$e_t(\mu, \rho) := \mathbb{E}_{X_{<t}\sim\mu}\left[\sum_{a\in\mathcal{X}}\Big(\mu(a \mid X_{<t}) - \rho_t(a \mid X_{<t})\Big)^2\right], \tag{4}$$

where the expectation is taken over the random history $X_{<t}$ generated according to μ. The *cumulative expected squared prediction error* is

$$S_\infty(\mu, \rho) := \sum_{t=1}^{\infty} e_t(\mu, \rho). \tag{5}$$

Since $e_t(\mu, \rho) \geq 0$, finiteness of $S_\infty(\mu, \rho)$ implies $e_t(\mu, \rho) \to 0$. We say that ρ *learns* μ if $S_\infty(\mu, \rho)$ is finite.

Realizable sequential prediction and universal class of estimable measures. As in classical MDL theory, we consider the realizable setting, in which the true measure μ is assumed to belong to a countable class $\mathcal{M}$ of probability measures on $\mathcal{X}^\infty$ known to the predictor. We say that a sequential prediction algorithm *learns* $\mathcal{M}$ if, for every $\mu \in \mathcal{M}$, the predictor ρ induced by the algorithm learns μ from its observed prefixes.

An important example of a countable model class is the *universal class of estimable measures*, as studied in algorithmic probability theory and Solomonoff induction.

A probability measure is called *estimable* if there exists a Turing machine that can approximate its cylinder probabilities to any prescribed precision. The precise computational convention and the associated description length are given in Definition 2.3. We denote by $\mathcal{M}_U$ the class of all estimable probability measures on $\{0, 1\}^\infty$ and refer to it as the the *universal model class*.[2]

Description length and Kolmogorov complexity. The connection between an estimable measure and a Turing machine that computes its cylinder probabilities allows us to assign a description length to the measure. This is the principle underlying Kolmogorov complexity. We first recall the necessary concepts.

Definition 2.1 (Prefix Turing Machine and Self-Delimiting Program)**.** A *prefix Turing machine* T has a one-way read-only program tape, a separate read-only tape containing a finite input z, finitely many work tapes, and an output tape. We write $T(p; z)$ for the initialization where the program tape and the input tape are initiated with p and z respectively. For input z, we say a program $p \in \{0, 1\}^*$ is (T, z)-*self-delimiting* if $T(p; z)$ halts with the program head pointing exactly to the last symbol of p. We write

$$T(p; z) = a$$

if p is (T, z)-self-delimiting and $T(p, z)$ outputs a. Note that the set of all (T, z)-self-delimiting programs is prefix-free.

Definition 2.2 (Universal Prefix Turing Machine)**.** A prefix Turing machine U is universal if, for every prefix Turing machine T, there exists a finite binary string e_T such that for all input z and (T, z)-self-delimiting program q,

$$U(e_T q; z) = T(q; z). \tag{6}$$

Equation (6) implicitly says $e_T q$ is (U, z)-self-delimiting.

Such machines exist by the universal machine construction in the proof of Theorem 3.1.1 in (Li and Vitányi, 2008). We fix one such reference machine U throughout the paper. A program for any fixed machine T can therefore be simulated by U with an additive overhead of $|e_T|$ bits, independent of both the program and the auxiliary input.

Fix computable binary encodings of pairs (x, k) and of rational numbers. Rational outputs below are understood through these fixed encodings.

[2]The use of a binary alphabet is conventional. Any fixed finite alphabet can be encoded using binary strings, changing the relevant Kolmogorov complexities by at most an additive constant.

Definition 2.3 (Prefix Kolmogorov Complexity of an Estimable Measure)**.** Let ν be an estimable probability measure on $\{0,1\}^\infty$. A program p *computes* ν if, for every $x \in \{0,1\}^*$ and every precision parameter $r \in \mathbb{N}$, p is $(U, \langle x, r\rangle)$-self-delimiting and

$$|U(p; \langle x, r\rangle) - \nu(x)| \leq 2^{-r},$$

where we interpret the output $U(p; \langle x, r\rangle)$ as a rational number. The *prefix Kolmogorov complexity* of ν is

$$K_U(\nu) := \min\{|p| : p \text{ computes } \nu\}.$$

Since U is fixed, we also write $K(\nu) := K_U(\nu)$.

The universal model prior and MDL objective. We now connect the universal class $\mathcal{M}_U$ to the general MDL framework introduced in Section 1.1. Because $K(\nu)$ is defined using prefix-free descriptions, the algorithmic weights $w_\nu^U := 2^{-K(\nu)}$, $\nu \in \mathcal{M}_U$, satisfy Kraft's inequality:

$$\sum_{\nu \in \mathcal{M}_U} w_\nu^U \leq 1. \tag{7}$$

Thus, $(w_\nu^U)_{\nu \in \mathcal{M}_U}$ defines a subnormalized model prior over $\mathcal{M}_U$, encoding an inductive bias toward measures with shorter descriptions or equivalently, lower Kolmogorov complexity. Substituting these weights into the regularized MDL objective defined in Equation (1) of Section 1.1 gives

$$L_{\lambda, w^U}(\nu; x_{<t}) = \lambda K(\nu) - \log_2 \nu(x_{<t}).$$

Hence, universal MDL is a special case of the general weighted MDL framework introduced in Section 1.1. Although motivated by universal MDL, our results, like those of classical MDL theory, apply to any countable class $\mathcal{M}$ of probability measures on $\mathcal{X}^\infty$, where $\mathcal{X}$ is a finite alphabet, equipped with positive weights $(w_\nu)_{\nu \in \mathcal{M}}$ satisfying $\sum_{\nu \in \mathcal{M}} w_\nu \leq 1$. These weights encode an arbitrary inductive bias over $\mathcal{M}$. For example, if an enumeration $(\nu_i)_{i \geq 1}$ orders the models according to the desired bias, standard choices include the geometric weights $w_{\nu_i} = 2^{-i}$ and polynomial weights such as $w_{\nu_i} = 6/(\pi^2 i^2)$, see, for instance, Grünwald (2007).

Throughout this paper, we use the following terminology.

Definition 2.4 (Weighted Model Class)**.** A *weighted model class* is a triple $(\mathcal{X}, \mathcal{M}, w)$, where $\mathcal{X}$ is a finite alphabet, $\mathcal{M}$ is a countable class of probability measures on $\mathcal{X}^\infty$, and $w = (w_\nu)_{\nu \in \mathcal{M}}$ are positive weights satisfying $\sum_{\nu \in \mathcal{M}} w_\nu \leq 1$, We refer to w as the *model prior*, which may be subnormalized. We refer to $\left(\{0,1\}, \mathcal{M}_U, w^U\right)$ as the *universal weighted model class*.

3 Prediction guarantees and the regularization threshold under additive MDL approximation

In this section, we study additive approximations to the MDL objective $L_{\lambda,w}$ defined in Equation (1) and present the high-level proof ideas underlying our characterization results. Full proofs are provided in the appendix.

Approximation robustness: an upper bound. Extending classical exact-MDL prediction guarantees, our first result shows that additive MDL approximation is robust whenever $\lambda \geq 1$.

Theorem 3.1 (Robustness under Sufficient Regularization)**.** *For every $C \geq 0$, $\lambda \geq 1$, weighted model class $(\mathcal{X}, \mathcal{M}, w)$, and true environment $\mu \in \mathcal{M}$, every approximate MDL predictor $\tilde{\rho}^+ = (\tilde{\rho}_t^+)_{t \geq 1}$ with additive slack C, as defined in Equation (2), has finite cumulative expected squared prediction error. More precisely,*

$$\sum_{t \geq 1} \mathbb{E}_\mu \left[\sum_{a \in \mathcal{X}} \left(\mu(a \mid X_{<t}) - \tilde{\rho}_t^+(a \mid X_{<t}) \right)^2 \right] \leq 16 K_{\mathrm{PH}} \lambda \, 2^{C/\lambda} w_\mu^{-1}.$$

Here, K_{PH} is the constant appearing in the exact static MDL quadratic-loss bound of Poland and Hutter (2005); under the normalization used in that bound, one may take $K_{\mathrm{PH}} = 21$.

Proof idea of Theorem 3.1. For each $\nu \in \mathcal{M}$, define

$$c_\nu := 2^{-C} \left(\frac{w_\mu}{w_\nu} \right)^\lambda .$$

Whenever $\tilde{\rho}_t^+ = \nu$, the additive approximation condition, compared against the score of the true environment μ, implies

$$\frac{\nu(X_{<t})}{\mu(X_{<t})} \geq c_\nu .$$

We decompose the prediction error according to the selected model and then replace the event $\{\tilde{\rho}_t^+ = \nu\}$ by this necessary likelihood-ratio condition. This gives

$$S_\infty(\mu, \tilde{\rho}^+) \leq \sum_{\nu \in \mathcal{M}} \mathbb{E}_\mu \left[\sum_{t \geq 1} \mathbf{1} \left\{ \frac{\nu(X_{<t})}{\mu(X_{<t})} \geq c_\nu \right\} \sum_{a \in \mathcal{X}} \big(\mu(a \mid X_{<t}) - \nu(a \mid X_{<t}) \big)^2 \right] .$$

For each ν, an affinity-telescoping argument bounds the corresponding summand by

$$\frac{2}{\alpha(1-\alpha)} c_\nu^{-\alpha}, \qquad \alpha \in (0,1),$$

whereas a stopping argument based on the exact static MDL bound gives $2K_{\mathrm{PH}} c_\nu^{-1}$.

Taking $\alpha = 1/(2\lambda)$ and combining these two bounds according to the weight ratio $r_\nu := w_\nu / w_\mu$ bounds the contribution of each ν by $16 K_{\mathrm{PH}} \lambda \, 2^{C/\lambda} r_\nu$. Finally, $\sum_{\nu \in \mathcal{M}} r_\nu = (w_\mu)^{-1} \sum_{\nu \in \mathcal{M}} w_\nu \leq w_\mu^{-1}$, and hence

$$S_\infty(\mu, \tilde{\rho}^+) \leq 16 K_{\mathrm{PH}} \lambda \, 2^{C/\lambda} w_\mu^{-1}.$$

Full details are given in Appendix A.

Corollary 3.1 (Universal Induction under Additive MDL Approximation)**.** *For every $C \geq 0$, $\lambda \geq 1$, and true environment $\mu \in \mathcal{M}_U$, every approximate MDL predictor $\tilde{\rho}^+ = (\tilde{\rho}_t^+)_{t \geq 1}$ with additive slack C for the universal weighted model class $(\{0,1\}, \mathcal{M}_U, w^U)$ has vanishing expected one-step squared prediction error, that is,*

$$e_t(\mu, \tilde{\rho}^+) \longrightarrow 0 \qquad \text{as } t \to \infty.$$

Consequently, additive approximate MDL achieves universal induction over $\mathcal{M}_U$.

This follows directly from Theorem 3.1, since every nonnegative summable sequence converges to zero.

Near-sharpness of the upper bound. The next proposition shows that the dependence of the upper bound in Theorem 3.1 on C, λ, and w_μ is essentially sharp in the worst case. Write $z_+ := \max\{z, 0\}$.

Proposition 3.1 (Nearly Matching Lower Bound)**.** *There exists an absolute constant $c > 0$ such that the following holds. For every $\lambda \geq 1$, $C \geq 0$, and $\alpha \in (0,1)$, there exists a finite weighted model class $(\{0,1\}, \mathcal{M}, w)$, a true environment $\mu \in \mathcal{M}$ with $w_\mu = \alpha$, and a valid additive approximate MDL predictor $\tilde{\rho}^+ = (\tilde{\rho}_t^+)_{t \geq 1}$ with slack C such that*

$$\sum_{t \geq 1} \mathbb{E}_\mu \left[\sum_{a \in \{0,1\}} \Big(\mu(a \mid X_{<t}) - \tilde{\rho}_t^+(a \mid X_{<t}) \Big)^2 \right] \geq c \lambda \Big(2^{C/\lambda} (w_\mu^{-1} - 1) - 1 \Big)_+ .$$

The proof of Proposition 3.1 is given in Appendix A.

Model-complexity regularization threshold. Theorem 3.1 establishes that additive MDL approximation is robust throughout the regime $\lambda \geq 1$. This regime includes the classical MDL objective

$$L(\text{model}) + L(\text{data} \mid \text{model}),$$

which assigns equal weight to the model and data description lengths in accordance with their standard coding interpretation. Complementing this result, the next theorem shows that the classical choice $\lambda = 1$ is the sharp threshold for a universal prediction guarantee. Specifically, when $0 < \lambda < 1$, satisfying the additive approximate MDL criterion in Equation (2) is insufficient to guarantee universal induction because the criterion permits pathological pointwise overfitting to the observed prefix.

Theorem 3.2 (Universal Induction Guarantees Require Strong Model-Complexity Regularization)**.** *There exist universal constants $c_1, c_2 > 0$ such that the following holds. Write $C_\lambda := c_1 \cdot \log \frac{1}{1-\lambda} + c_2$. For every $\lambda \in (0,1)$, there exist a true environment $\mu \in \mathcal{M}_U$ and a valid additive approximate MDL predictor $\tilde{\rho}^+ = (\tilde{\rho}_t^+)_{t\geq 1}$ with slack C_λ satisfying Equation (2), whose expected squared prediction error equals $1/2$ at every time t. Consequently,*

$$\sum_{t\geq 1} \mathbb{E}_\mu \left[\sum_{a\in\{0,1\}} \Big(\mu(a \mid X_{<t}) - \tilde{\rho}_t^+(a \mid X_{<t}) \Big)^2 \right] = +\infty.$$

Proof idea of Theorem 3.2. For every $x \in \{0,1\}^*$, the universal MDL objective takes the form

$$L_{\lambda,w^U}(\nu; x) = \lambda K(\nu) - \log_2 \nu(x).$$

Lemma 3.1 below shows that when $0 < \lambda < 1$, a deterministic measure supported on an extension of the observed prefix can remain within a constant additive slack of the optimal MDL score. Taking the true environment μ to be the fair-coin measure, an additive approximate MDL predictor can select such a deterministic, overfitting measure after every history. Its next-step prediction is deterministic, whereas the true conditional distribution is always $(1/2, 1/2)$. Its squared prediction error is therefore $1/2$ at every step. Full details are provided in Appendix A.

Lemma 3.1 (Existence of a Feasible Deterministic Approximate Predictor)**.** *There exist universal constants $c_1, c_2 > 0$ such that the following holds. Write $C_\lambda := c_1 \cdot \log \frac{1}{1-\lambda} + c_2$. For every $\lambda \in (0,1)$ and $x \in \{0,1\}^*$, there exists a deterministic estimable measure $\nu_x \in \mathcal{M}_U$ satisfying $\nu_x(x) = 1$ and*

$$L_{\lambda,w^U}(\nu_x; x) \leq \inf_{\nu\in\mathcal{M}_U} L_{\lambda,w^U}(\nu; x) + C_\lambda.$$

Remark 3.1. In Theorem A.1, we provide a comprehensive characterization, showing that *for every $\lambda \in (0,1)$ and slack $C \geq 0$, every* valid approximate MDL predictor can lead to infinite accumulated error. The case $C = 0$ corresponds to exact MDL. The proof is more involved; it does not prove a lower bound on the per-step expected error, but shows that infinitely many steps incur constant error.

4 Multiplicative MDL approximation does not guarantee universal prediction

In this section, we show that, throughout the entire regularization regime $\lambda > 0$, multiplicative MDL approximation does not uniformly guarantee finite cumulative expected squared prediction error. This establishes a sharp contrast between additive and multiplicative MDL approximation. We focus on the universal weighted model class.

Theorem 4.1 (Failure of Universal-Induction Guarantee under Multiplicative MDL Approximation)**.** *For every $\lambda > 0$ and $\varepsilon > 0$, there exist a true environment $\mu \in \mathcal{M}_U$ and a valid multiplicative approximate MDL predictor $\tilde{\rho}^\times = (\tilde{\rho}_t^\times)_{t\geq 1}$ with approximation factor $1+\varepsilon$, satisfying Equation (3), such that*

$$\sum_{t\geq 1} \mathbb{E}_\mu \left[\sum_{a\in\{0,1\}} \Big(\mu(a \mid X_{<t}) - \tilde{\rho}_t^\times(a \mid X_{<t}) \Big)^2 \right] = +\infty.$$

Proof idea of Theorem 4.1. The proof separates two regimes.

When $0 < \lambda < 1$, the additive overfitting construction becomes multiplicatively admissible because the optimal MDL score diverges in probability along fair-coin sequences. Consequently, any constant additive slack is eventually absorbed by the multiplicative tolerance.

When $\lambda \geq 1$, a slightly biased rational Bernoulli measure remains within the multiplicative tolerance with probability bounded away from zero, because the effective additive slack permitted by the multiplicative approximation grows linearly with the optimal MDL score.

In both regimes, we construct a valid multiplicative approximate MDL predictor whose expected one-step squared prediction error is bounded away from zero for infinitely many time steps.

A key ingredient is the following technical lemma, which relates the optimal universal MDL score of the observed data to its description length under the true environment. For $0 < \lambda < 1$, we apply the lemma at $\lambda = 1$ and use the relation

$$\mathrm{OPT}_{\lambda,w^U}(x) \geq \lambda\, \mathrm{OPT}_{1,w^U}(x).$$

Full proofs of the lemma and theorem are given in Appendix B.

Lemma 4.1 (Optimal Universal MDL Value versus True Description Length)**.** *For $\lambda \geq 1$ and $x \in \{0,1\}^*$, define*

$$\mathrm{OPT}_{\lambda,w^U}(x) := \inf_{\nu \in \mathcal{M}_U} L_{\lambda,w^U}(\nu; x).$$

Let $X_{1:\infty} \sim \mu$ for any estimable true environment $\mu \in \mathcal{M}_U$. Then, for every $n \in \mathbb{N}$ and $c > 0$, with probability at least $1 - 2^{-c}$,

$$\mathrm{OPT}_{\lambda,w^U}(X_{1:n}) \geq \log_2 \frac{1}{\mu(X_{1:n})} - c.$$

In other words, the optimal universal MDL score is unlikely to be substantially smaller than the description length of the observed data under the true environment.

5 Conclusion

This paper characterizes when approximate MDL preserves reliable sequential prediction. For countable realizable model classes, any fixed additive optimization error preserves finite cumulative expected squared prediction error whenever $\lambda \geq 1$; our nearly matching lower bounds show that the dependence on the approximation slack is essentially sharp. In the universal setting, $\lambda = 1$ is the sharp model-complexity regularization threshold: when $0 < \lambda < 1$, additive approximation can permit persistent overfitting and infinite cumulative prediction error. Under multiplicative approximation, the approximation criterion alone does not guarantee reliable prediction for any $\lambda > 0$.

Our analysis isolates how optimization error affects prediction, but it does not provide an algorithm for achieving the required approximation or overcome the uncomputability of Kolmogorov complexity in the universal setting. We view this work as a first step toward the broader goal of bringing universal induction closer to practical prediction. Two complementary directions toward this goal are to study computationally restricted or structurally constrained model classes and to relax exact model selection through approximate optimization. This paper pursues the second direction within the MDL framework. Combining the two is an important next step: even when restrictions make prediction computable, exact optimization may remain computationally infeasible. Our characterization of which forms of approximation preserve prediction guarantees provides a foundation for studying such settings.

Another important direction is to connect approximate MDL theory with the training and performance of large language models. For example, could this theory be extended to help explain empirical scaling laws for LLMs? Could differences in compression and optimization across architectures help explain the empirical success of Transformers? Addressing these questions will require formally connecting practical training objectives to the forms of approximation and robustness characterized here, as well as developing a better understanding of the model-complexity regularization induced by modern learning architectures. Finally, our results assume realizability and passive prediction. Extending the analysis to agnostic and interactive settings is therefore another natural direction.

Acknowledgments. We thank the anonymous reviewers for their constructive comments and suggestions. Qian Li's work was supported by Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone Project (No.HZQSWS-KCCYB-2024016), the National Natural Science Foundation of China Grant Number 62672335, Guangdong Basic and Applied Basic Research Foundation Grant Number 2026A1515011078, and the Internal Project of Shenzhen Research Institute of Big Data under grant number J00120260007. Shang-Hua Teng's work was supported in part by a USC Provost Annual 5K Research Fund for USC University and Distinguished Professors.

We used AI tools to help develop the proofs of Proposition 3.1, Lemma A.3, and Theorem A.1. We also used these tools to improve the writing. We take full responsibility for the entire content of this work.

References

Pieter Adriaans and Paul M. B. Vitányi. Approximation of the two-part MDL code. *IEEE Transactions on Information Theory*, 55(1):444–457, 2009. doi: 10.1109/TIT.2008.2008152.

Léonard Blier and Yann Ollivier. The description length of deep learning models. In *Advances in Neural Information Processing Systems 31*, pages 2220–2230, 2018.

Jörg Bornschein, Yazhe Li, and Marcus Hutter. Sequential learning of neural networks for prequential MDL. In *The Eleventh International Conference on Learning Representations*, 2023.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*, 2024.

Péter Gács, John T. Tromp, and Paul M. B. Vitányi. Algorithmic statistics. *IEEE Transactions on Information Theory*, 47(6):2443–2463, 2001. doi: 10.1109/18.945257.

Gustavo L Gilardoni. On pinsker's and vajda's type inequalities for csiszár's f-divergences. *IEEE Transactions on Information Theory*, 56(11):5377–5386, 2010.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007. ISBN 9780262529631.

Peter D. Grünwald and Steven de Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005*, volume 3559 of *Lecture Notes in Computer Science*, pages 652–667. Springer, 2005. doi: 10.1007/11503415_44.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pages 5–13, 1993. doi: 10.1145/168304.168306.

Shuichi Hirahara and Mikito Nanashima. Complexity-theoretic universal inductive inference. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*, STOC '26, pages 377–385. Association for Computing Machinery, 2026. doi: 10.1145/3798129.3800757.

Marcus Hutter. Sequence prediction based on monotone complexity. In *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 506–521. Springer, 2003. doi: 10.1007/978-3-540-45167-9_37.

Marcus Hutter. Discrete MDL predicts in total variation. In *Advances in Neural Information Processing Systems 22*, pages 817–825, 2009.

Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

Qian Li and Shang-Hua Teng. Approximate Solomonoff induction: On the robustness of learned universal predictors, 2026. URL https://arxiv.org/abs/2610.29289.

Alexey Milovanov. Prediction and MDL for infinite sequences. *Theory of Computing Systems*, 68(5): 1124–1137, 2024. doi: 10.1007/s00224-024-10180-0.

Marvin Minsky. Remarks in panel discussion: The Limits of Understanding. World Science Festival, New York City, June 2010. URL https://www.worldsciencefestival.com/videos/the-limits-of-understanding/. Video. Accessed: 2026-05-02.

Jan Poland and Marcus Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005. doi: 10.1109/TIT.2005.856956.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. doi: 10.1016/0005-1098(78)90005-5.

Jorma Rissanen. *Stochastic complexity in statistical inquiry*, volume 15. World scientific, 1998.

Ray J. Solomonoff. A formal theory of inductive inference. Part I. *Information and Control*, 7(1): 1–22, 1964a. doi: 10.1016/S0019-9958(64)90223-2.

Ray J. Solomonoff. A formal theory of inductive inference. Part II. *Information and Control*, 7(2): 224–254, 1964b. doi: 10.1016/S0019-9958(64)90131-7.

Nikolai K. Vereshchagin and Paul M. B. Vitányi. Kolmogorov's structure functions and model selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, 2004. doi: 10.1109/TIT.2004.838346.

Paul M. B. Vitányi and Ming Li. Minimum description length induction, bayesianism, and kolmogorov complexity. *IEEE Transactions on Information Theory*, 46(2):446–464, 2000. doi: 10.1109/18.825807.

Chris S. Wallace and David L. Dowe. Minimum message length and kolmogorov complexity. *The Computer Journal*, 42(4):270–283, 1999.

A Proofs for additive MDL approximation: prediction guarantees and the regularization threshold

Universal models and exact minimizers. For the universal-model arguments below, we work with $(\{0,1\}, \mathcal{M}_U, w^U)$, where $\mathcal{M}_U$ is the class of all estimable measures and $w^U_\nu := 2^{-K(\nu)}$. By Kraft's inequality, $\sum_{\nu\in\mathcal{M}_U} w^U_\nu \le 1$. For every $\lambda > 0$ and $x \in \{0,1\}^*$, write

$$L_{\lambda,w^U}(\nu; x) := \lambda K(\nu) - \log_2 \nu(x), \qquad \mathrm{OPT}_{\lambda,w^U}(x) := \inf_{\nu\in\mathcal{M}_U} L_{\lambda,w^U}(\nu; x).$$

We fix an exact minimizer, with ties broken by a fixed rule,

$$\hat{\nu}^\lambda_x \in \arg\min_{\nu\in\mathcal{M}_U} L_{\lambda,w^U}(\nu; x) = \arg\max_{\nu\in\mathcal{M}_U} (w^U_\nu)^\lambda \nu(x). \tag{8}$$

Such a minimizer exists and satisfies $\hat{\nu}^\lambda_x(x) > 0$. Indeed, the deterministic measure concentrated on $x0^\infty$ belongs to $\mathcal{M}_U$ and has positive weighted likelihood, say $b > 0$. Any model with weighted likelihood at least b must satisfy $w^U_\nu \ge b^{1/\lambda}$. Since $\sum_\nu w^U_\nu \le 1$, only finitely many models satisfy this condition, and the global maximum is attained among them.

A.1 Proof of Theorem 3.1 and Proposition 3.1

We prove Theorem 3.1 using three auxiliary lemmas. Lemma A.1 provides the one-step inequality underlying Lemma A.2. Lemmas A.2 and A.3 give complementary bounds on cumulative prediction loss. Combining these bounds and summing over the model class yields the theorem.

The following elementary lemma bounds the squared loss in terms of an affinity gap and plays a key role in the proofs of our positive results. It also follows from a Pinsker-type inequality for Csiszár f-divergences; see, for example, (Gilardoni, 2010, Corollary 9).

Lemma A.1. *For any $\alpha \in (0,1)$ and distributions p, q on $\mathcal{X}$,*

$$\sum_{a\in\mathcal{X}} (p_a - q_a)^2 \le \frac{2}{\alpha(1-\alpha)} \left(1 - \sum_{a\in\mathcal{X}} p_a^{1-\alpha} q_a^\alpha\right).$$

Proof of Lemma A.1. Fix $\alpha \in (0,1)$. For $0 < x < y \le 1$, define $f(t) = t^{1-\alpha} y^\alpha$. For every $t \in (0, y]$,

$$f''(t) = -\alpha(1-\alpha) y^\alpha t^{-1-\alpha} \le -\frac{\alpha(1-\alpha)}{y} \le -\alpha(1-\alpha).$$

By Taylor's theorem with Lagrange remainder, there exists $\xi \in (x, y)$ such that

$$f(x) = f(y) + f'(y)(x-y) + \frac{1}{2} f''(\xi)(x-y)^2.$$

Using $f(y) = y$, $f'(y) = 1-\alpha$, and the above bound on $f''(\xi)$, we obtain

$$\begin{aligned} x^{1-\alpha} y^\alpha &\le y + (1-\alpha)(x-y) - \frac{\alpha(1-\alpha)}{2}(x-y)^2 \\ &= (1-\alpha)x + \alpha y - \frac{\alpha(1-\alpha)}{2}(x-y)^2. \end{aligned}$$

Rearranging gives

$$(1-\alpha)x + \alpha y - x^{1-\alpha} y^\alpha \ge \frac{\alpha(1-\alpha)}{2}(x-y)^2. \tag{9}$$

This inequality is immediate when $x = y$, extends to zero coordinates by continuity, and holds for $x > y$ by exchanging x, y and $\alpha, 1-\alpha$.

Now apply (9) with $x = p_a$ and $y = q_a$ for each $a \in \mathcal{X}$. Summing over $a \in \mathcal{X}$ yields

$$\sum_{a\in\mathcal{X}} \left((1-\alpha)p_a + \alpha q_a - p_a^{1-\alpha} q_a^\alpha\right) \ge \frac{\alpha(1-\alpha)}{2} \sum_{a\in\mathcal{X}} (p_a - q_a)^2.$$

Since p and q are probability distributions,

$$\sum_{a\in\mathcal{X}}\big((1-\alpha)p_a+\alpha q_a\big)=(1-\alpha)\sum_a p_a+\alpha\sum_a q_a=1.$$

As a result, the left-hand side is equal to $1-\sum_{a\in\mathcal{X}}p_a^{1-\alpha}q_a^{\alpha}$. Thus,

$$1-\sum_{a\in\mathcal{X}}p_a^{1-\alpha}q_a^{\alpha}\geq\frac{\alpha(1-\alpha)}{2}\sum_{a\in\mathcal{X}}(p_a-q_a)^2,$$

as desired. □

We next use the inequality from Lemma A.1 to derive a cumulative prediction-error bound restricted to times at which the prefix likelihood ratio exceeds a prescribed threshold. This is useful for additive approximate MDL because, whenever a model ν is selected, the approximation condition imposes a lower bound on its likelihood ratio relative to the true environment.

Lemma A.2 (Power Threshold-Loss Bound)**.** *Let $\mathcal{X}$ be finite or countable. Let P and Q be probability measures on $\mathcal{X}^\infty$. Define the prefix likelihood ratio*

$$\Lambda(x):=\frac{Q(x)}{P(x)}$$

whenever $P(x)>0$, and set $\Lambda(x)=0$ otherwise. For $X_{1:\infty}\sim P$, write $\Lambda_t:=\Lambda(X_{<t})$. For each measure, conditional distributions on histories to which it assigns zero probability are defined arbitrarily. Then, for every $\alpha\in(0,1)$ and every $c>0$,

$$\mathbb{E}_{X\sim P}\sum_{t\geq1}\mathbf{1}\{\Lambda_t\geq c\}\sum_{a\in\mathcal{X}}\big(P(a\mid X_{<t})-Q(a\mid X_{<t})\big)^2\leq\frac{2}{\alpha(1-\alpha)}c^{-\alpha}.$$

Proof. Fix $\alpha\in(0,1)$ and $c>0$, and set $R_t:=\Lambda_t^\alpha$. Given $X_{<t}$, write

$$p_a:=P(a\mid X_{<t}),\qquad q_a:=Q(a\mid X_{<t}),\qquad d_t:=1-\sum_{a\in\mathcal{X}}p_a^{1-\alpha}q_a^{\alpha}.$$

All expectations below are under P, and histories of zero P-probability may be ignored.

First, we reduce to a bound on the weighted affinity gaps. By Lemma A.1,

$$\sum_{a\in\mathcal{X}}(p_a-q_a)^2\leq\frac{2}{\alpha(1-\alpha)}d_t.$$

Moreover, $R_t\geq c^\alpha$ whenever $\Lambda_t\geq c$, and hence $\mathbf{1}\{\Lambda_t\geq c\}\leq c^{-\alpha}R_t$. Combining these inequalities, for every finite $T\geq1$ we obtain

$$\mathbb{E}_P\sum_{t=1}^{T}\mathbf{1}\{\Lambda_t\geq c\}\sum_{a\in\mathcal{X}}(p_a-q_a)^2\leq\frac{2}{\alpha(1-\alpha)}\cdot c^{-\alpha}\cdot\mathbb{E}_P\sum_{t=1}^{T}R_td_t.$$

Thus, it remains to prove that $\mathbb{E}_P\sum_{t=1}^{T}R_td_t\leq1$.

Next, we prove $\mathbb{E}_P\sum_{t=1}^{T}R_td_t\leq1$ by telescoping. On the event $p_{X_t}>0$, which holds P-almost surely, the likelihood ratio satisfies

$$\Lambda_{t+1}=\Lambda_t\frac{q_{X_t}}{p_{X_t}},\qquad R_{t+1}=R_t\left(\frac{q_{X_t}}{p_{X_t}}\right)^{\alpha}.$$

Taking conditional expectation given $X_{<t}$ therefore yields

$$\mathbb{E}_P[R_{t+1}\mid X_{<t}]=R_t\sum_{a:p_a>0}p_a\left(\frac{q_a}{p_a}\right)^{\alpha}=R_t\sum_{a\in\mathcal{X}}p_a^{1-\alpha}q_a^{\alpha}=R_t(1-d_t).$$

If $Q(X_{<t})=0$, then $R_t=R_{t+1}=0$, so this identity remains valid regardless of the chosen conditional distribution q.

The empty history has probability 1 under both measures, so $R_1 = 1$. Since $d_t \geq 0$, the preceding identity implies inductively that $\mathbb{E}_P[R_t] \leq 1$ for every t. Thus all R_t are integrable, and rearranging gives

$$\mathbb{E}_P[R_t - R_{t+1} \mid X_{<t}] = R_t d_t.$$

Taking expectations in the identity and summing from $t = 1$ to T, we obtain the telescoping bound

$$\mathbb{E}_P \sum_{t=1}^{T} R_t d_t = \sum_{t=1}^{T} \big(\mathbb{E}_P[R_t] - \mathbb{E}_P[R_{t+1}]\big) = 1 - \mathbb{E}_P[R_{T+1}] \leq 1,$$

where the last inequality uses $R_{T+1} \geq 0$.

Finally, letting $T \to \infty$ and applying monotone convergence to the nonnegative loss terms proves the claim. □

Lemma A.2 bounds the thresholded cumulative squared loss by a multiple of $c^{-\alpha}$ for every $\alpha \in (0, 1)$. The following lemma complements this result with a bound proportional to c^{-1}, obtained through a likelihood-ratio stopping argument and the classical exact static-MDL quadratic-loss bound.

Lemma A.3 (Thresholded Likelihood-Ratio Loss)**.** *Let P and Q be probability measures on $\mathcal{X}^\infty$. Write*

$$\Lambda(x) := \frac{Q(x)}{P(x)}, \qquad \Lambda_t := \Lambda(X_{<t}),$$

where $X_{1:\infty} \sim P$, and set $\Lambda(x) = 0$ when $P(x) = 0$. As above, each measure's conditional distributions are defined arbitrarily on its zero-probability histories. There exists a universal constant $K_{\mathrm{PH}} < \infty$ such that, for every $c > 0$,

$$\mathbb{E}_{X\sim P} \sum_{t\geq 1} \mathbf{1}\{\Lambda_t \geq c\} \sum_{a\in\mathcal{X}} \big(P(a \mid X_{<t}) - Q(a \mid X_{<t})\big)^2 \leq \frac{2K_{\mathrm{PH}}}{c}.$$

Using the static-MDL quadratic-loss bound of (Poland and Hutter, 2005, Corollary 14), one may take $K_{\mathrm{PH}} = 21$.

Proof. Let $\tau := \inf\{t \geq 1 : \Lambda_t \geq c\}$ be the first threshold-crossing time. If $\tau = \infty$, the thresholded loss is zero. Condition on $\{\tau < \infty\}$ and on $X_{<\tau} = x$. Let P^x and Q^x be the conditional continuation measures after history x. For a future string y with $P(xy) > 0$,

$$\frac{Q^x(y)}{P^x(y)} = \frac{Q(xy)/Q(x)}{P(xy)/P(x)} = \frac{\Lambda(xy)}{\Lambda(x)}.$$

Therefore the rule that uses Q exactly when $\Lambda(xy) \geq c$ is equivalent to exact static MDL on the two-model class $\{P^x, Q^x\}$ with unnormalized weights

$$\bar{w}_P = 1, \qquad \bar{w}_Q = \frac{\Lambda(x)}{c}.$$

Indeed, this static MDL rule selects Q^x exactly when

$$\bar{w}_Q Q^x(y) \geq \bar{w}_P P^x(y),$$

which is equivalent to $\Lambda(xy) \geq c$.

By the exact static-MDL quadratic-loss bound of (Poland and Hutter, 2005, Corollary 14), the conditional expected cumulative squared loss of this two-model static MDL predictor is at most

$$K_{\mathrm{PH}} \left(1 + \frac{\Lambda_\tau}{c}\right),$$

because the normalized weight of the true model P^x is $\frac{1}{1+\Lambda_\tau/c}$.

It remains to average over the stopping event. With the convention $\Lambda_t = 0$ on P-null histories, $(\Lambda_t)_{t\geq 1}$ is a nonnegative P-supermartingale with $\Lambda_1 = 1$. Optional stopping applied to $\Lambda_{\tau\wedge n}$, followed by Fatou's lemma, gives

$$\mathbb{E}_P[\Lambda_\tau \cdot \mathbf{1}_{\{\tau<\infty\}}] \leq \mathbb{E}_P[\liminf_{n\to\infty} \Lambda_{\tau\wedge n} \cdot \mathbf{1}_{\{\tau<\infty\}}] \leq \liminf_{n\to\infty} \mathbb{E}_P[\Lambda_{\tau\wedge n} \cdot \mathbf{1}_{\{\tau<\infty\}}] \leq 1.$$

Moreover, since $\Lambda_\tau \geq c$ on $\{\tau < \infty\}$, $\mathbb{P}_P(\tau < \infty) \leq \frac{1}{c}$. Hence

$$\mathbb{E}_P \sum_{t\geq 1} \mathbf{1}\{\Lambda_t \geq c\} \sum_{a\in\mathcal{X}} \big(P(a \mid X_{<t}) - Q(a \mid X_{<t})\big)^2$$
$$\leq K_{\mathrm{PH}} \left(\mathbb{P}_P(\tau < \infty) + \frac{\mathbb{E}_P[\Lambda_\tau \cdot \mathbf{1}_{\{\tau<\infty\}}]}{c}\right) \leq \frac{2K_{\mathrm{PH}}}{c}. \qquad \square$$

We now combine Lemmas A.2 and A.3 to prove Theorem 3.1.

Proof of Theorem 3.1. Let $\tilde{\rho}^+$ satisfy Equation (2). For each $\nu \in \mathcal{M}$, define

$$\Lambda_t^\nu := \frac{\nu(X_{<t})}{\mu(X_{<t})}, \qquad c_\nu := 2^{-C} \left(\frac{w_\mu}{w_\nu}\right)^\lambda,$$

and

$$J_\nu = \mathbb{E}_\mu \sum_{t\geq 1} \mathbf{1}\{\Lambda_t^\nu \geq c_\nu\} \sum_{a\in\mathcal{X}} \big(\mu(a \mid X_{<t}) - \nu(a \mid X_{<t})\big)^2$$

Histories of zero μ-probability may be ignored.

Whenever $\tilde{\rho}_t^+ = \nu$, the additive approximation condition gives

$$w_\nu^\lambda \nu(X_{<t}) \geq 2^{-C} w_\mu^\lambda \mu(X_{<t}),$$

and hence $\Lambda_t^\nu \geq c_\nu$. Decomposing the loss according to the selected model, we have

$$S_\infty(\mu, \tilde{\rho}^+) \leq \sum_{\nu\in\mathcal{M}} \mathbb{E}_\mu \sum_{t\geq 1} \mathbf{1}\{\Lambda_t^\nu \geq c_\nu\} \sum_{a\in\mathcal{X}} \big(\mu(a \mid X_{<t}) - \nu(a \mid X_{<t})\big)^2 = \sum_{\nu\in\mathcal{M}} J_\nu. \tag{10}$$

This is the common threshold template for all $\lambda \geq 1$.

First, apply Lemma A.2 with $P = \mu$, $Q = \nu$, $c = c_\nu$, and $\alpha := 1/(2\lambda)$. For $0 \leq c_\nu \leq 1$, this gives

$$J_\nu \leq \frac{2}{\alpha(1-\alpha)} c_\nu^{-\alpha} \leq 8\lambda c_\nu^{-\alpha} \leq 8\lambda c_\nu^{-1/\lambda}$$

where the last inequality uses $\alpha = 1/(2\lambda)$ and $1 - \alpha \geq 1/2$. Next, apply Lemma A.3 with $P = \mu$, $Q = \nu$, and $c = c_\nu$. For $c_\nu > 1$, we have

$$J_\nu \leq 2K_{\mathrm{PH}} c_\nu^{-1} \leq 2K_{\mathrm{PH}} c_\nu^{-1/\lambda}$$

. Thus, since $\sum_{\nu\in\mathcal{M}} w_\nu \leq 1$,$c_\nu := 2^{-C} \left(\frac{w_\mu}{w_\nu}\right)^\lambda$ and (10), we obtain

$$S_\infty(\mu, \tilde{\rho}^+) \leq \sum_{\nu\in\mathcal{M}} \max\{8\lambda, 2K_{\mathrm{PH}}\}\, c_\nu^{-1/\lambda} \leq 16K_{\mathrm{PH}} 2^{C/\lambda} \frac{\sum_{\nu\in\mathcal{M}} w_\nu}{w_\mu} \leq 16K_{\mathrm{PH}} \lambda\, 2^{C/\lambda} w_\mu^{-1}.$$

The last inequality uses $\lambda \geq 1$ and $K_{\mathrm{PH}} \geq 1$, as satisfied by the stated choice $K_{\mathrm{PH}} = 21$. This proves the theorem. $\square$

We next specialize Theorem 3.1 to the universal weighted model class. The resulting summability of the expected squared prediction errors yields Corollary 3.1.

Proof of Corollary 3.1. For the universal class with $w_\nu^U = 2^{-K(\nu)}$, Kraft's inequality for the underlying prefix machine gives $\sum_{\nu\in\mathcal{M}_U} w_\nu^U \leq 1$. Since the true estimable environment μ belongs to $\mathcal{M}_U$, Theorem 3.1 yields

$$S_\infty(\mu, \tilde{\rho}^+) = \sum_{t\geq 1} e_t(\mu, \tilde{\rho}^+) < \infty.$$

As each $e_t(\mu, \tilde{\rho}^+) \geq 0$, the terms of this convergent nonnegative series satisfy $e_t(\mu, \tilde{\rho}^+) \to 0$. $\square$

We next examine the sharpness of the upper bound in Theorem 3.1. Proposition 3.1 provides a nearly matching lower bound by constructing a finite weighted model class and a suitable additive approximate MDL predictor.

Proof of Proposition 3.1. Fix $\lambda \geq 1$, $C \geq 0$, and $w_0 \in (0,1)$. Let μ be the deterministic all-zero measure and set $w_\mu = w_0$.

The idea is to cycle through m false models for k rounds along the all-zero sequence. Each selected model predicts a fair coin, incurring squared loss $1/2$. We first derive conditions under which these selections satisfy the additive MDL approximation condition, and then choose the parameters to make the resulting loss large enough.

Step 1: Construct the model class. Let $m, k \geq 1$ be integers and let $r > 0$, all to be chosen in Step 4. Let $\mathcal{M} = \{\mu, \nu_1, \ldots, \nu_m\}$ and assign each false model the weight $w_{\nu_i} = r w_0$. The total weight is $w_0(1 + mr)$, so the weight constraint holds provided

$$mr \leq w_0^{-1} - 1. \tag{11}$$

For $1 \leq i \leq m$ and $1 \leq j \leq k$, define

$$t_{i,j} := (j-1)m + i.$$

Under ν_i, the symbols at times $t_{i,1}, \ldots, t_{i,k}$ are independent fair coin flips, and all other symbols are 0. Thus each false model has k designated times at which its prediction along the all-zero path differs from that of μ. Conditional distributions on histories of zero probability are defined arbitrarily.

Step 2: Derive the condition for admissible selections. At history $0^{t_{i,j}-1}$, select ν_i. At every other history x, select a model maximizing $w_\eta^\lambda \eta(x)$ over $\eta \in \mathcal{M}$. Such a maximizer exists because $\mathcal{M}$ is finite.

To verify the prescribed choices, fix $x = 0^{t_{i,j}-1}$. At this point, the models $\nu_1, \ldots, \nu_{i-1}$ have each had j fair coin times, whereas $\nu_i, \ldots, \nu_m$ have each had $j-1$. Since their weights are equal, ν_i has the largest weighted likelihood among the false models. Using $\mu(x) = 1$ and $\nu_i(x) = 2^{-(j-1)}$, we obtain

$$\frac{w_{\nu_i}^\lambda \nu_i(x)}{\max_{\eta \in \mathcal{M}} w_\eta^\lambda \eta(x)} = \frac{r^\lambda 2^{-(j-1)}}{\max\{1, r^\lambda 2^{-(j-1)}\}} = \min\{1, r^\lambda 2^{-(j-1)}\}.$$

Since $j \leq k$ and $2^{-C} \leq 1$, every prescribed choice satisfies Equation (2) provided

$$r^\lambda 2^{-(k-1)} \geq 2^{-C}. \tag{12}$$

All other choices are exact maximizers of the weighted likelihood.

Step 3: Count the prediction errors. At each designated time, the selected model predicts $(1/2, 1/2)$, whereas μ predicts $(1, 0)$, giving squared loss $1/2$. Since μ is deterministic and there are km such times,

$$S_\infty(\mu, \tilde{\rho}^+) \geq \frac{km}{2}.$$

Thus any $m, k \geq 1$ and $r > 0$ satisfying (11) and (12) yield a valid predictor with cumulative squared loss at least $km/2$. It remains to choose such parameters to obtain the claimed lower bound.

Step 4: Choose the parameters. Set $A := 2^{C/\lambda}(w_0^{-1} - 1)$. If $A \leq 1$, the claimed lower bound is zero, and the single-model class $\{\mu\}$ with the predictor that always selects μ suffices. Hence assume $A > 1$. For a given k, choose the smallest weight ratio allowed by (12), and then the largest number of false models allowed by (11):

$$r := 2^{(k-1-C)/\lambda}, \qquad m := \left\lfloor \frac{w_0^{-1} - 1}{r} \right\rfloor = \left\lfloor A 2^{(1-k)/\lambda} \right\rfloor.$$

These choices satisfy both constraints. We now choose k so that $m \geq 1$ and the bound from Step 3,

$$S_\infty(\mu, \tilde{\rho}^+) \geq \frac{k}{2} \left\lfloor A 2^{(1-k)/\lambda} \right\rfloor, \tag{13}$$

is sufficiently large. If $1 < A \leq 2$, choose $k := \lceil \lambda \log_2 A \rceil$. Then $k \geq 1$ and $k \leq 1 + \lambda \log_2 A$, so $A 2^{(1-k)/\lambda} \geq 1$ and hence $m \geq 1$. By (13),

$$S_\infty(\mu, \tilde{\rho}^+) \geq \frac{k}{2} \geq \frac{\lambda}{2} \log_2 A \geq \frac{\lambda}{2}(A-1),$$

where the last inequality follows from the concavity of $\log_2$ on $[1, 2]$.

If $A > 2$, choose $k := \lceil \lambda \rceil$. Since $\lambda \le k \le \lambda + 1$, $m = \lfloor A 2^{(1-k)/\lambda} \rfloor \ge \lfloor \frac{A}{2} \rfloor \ge \frac{A}{4}$, using $\lfloor x \rfloor \ge x/2$ for $x \ge 1$. In particular, $m \ge 1$, and

$$S_\infty(\mu, \tilde{\rho}^+) \ge \frac{km}{2} \ge \frac{\lambda A}{8} \ge \frac{\lambda(A-1)}{8}.$$

Combining these cases with $A \le 1$ proves

$$S_\infty(\mu, \tilde{\rho}^+) \ge \frac{\lambda}{8}\Big(2^{C/\lambda}(w_\mu^{-1} - 1) - 1\Big)_+.$$

Thus the proposition holds with $c = 1/8$. □

A.2 Proof of Lemma 3.1 and Theorem 3.2

We first establish Lemma A.4, which constructs a deterministic model consistent with the observed history while controlling its description length. For $0 < \lambda < 1$, this bound yields the deterministic approximate minimizers in Lemma 3.1. Selecting these models sequentially in a fair-coin environment then gives the counterexample in Theorem 3.2.

Lemma A.4 (Deterministic Refinement of an Estimable Measure)**.** *There exists a universal constant $c_{\mathrm{det}} < \infty$ such that for every estimable measure ν and every finite binary string x with $\nu(x) > 0$, there is a deterministic estimable measure δ_ω satisfying $\delta_\omega(x) = 1$ and*

$$K(\delta_\omega) \le K(\nu) + \log_2 \nu(x)^{-1} + c_{\mathrm{det}} \log_2\big(\log_2 \nu(x)^{-1} + 2\big) + c_{\mathrm{det}}.$$

Proof. Let p be a shortest program computing ν in the sense of Definition 2.3, so that $|p| = K_U(\nu)$. In particular, the same program p is $(U, \langle y, r\rangle)$-self-delimiting for every y and r. Write ϵ for the empty string and set $B := -\log_2 \nu(x) \ge 0$.

Step 1: Construct computable arithmetic intervals. Set $I_\epsilon = [0, 1)$. Given $I_y = [L_y, R_y)$, define

$$I_{y0} := [L_y, L_y + \nu(y0)), \qquad I_{y1} := [L_y + \nu(y0), R_y).$$

The consistency identity $\nu(y) = \nu(y0) + \nu(y1)$ implies inductively that $|I_y| = \nu(y)$ and that the two child intervals partition I_y. Their endpoints are computable uniformly from p and y, since each endpoint is obtained through finitely many additions of cylinder probabilities computed using p.

Step 2: Choose a short description of an interval inside I_x. Let $m := \lceil B \rceil + 2$. Since $|I_x| = \nu(x) \ge 4 \cdot 2^{-m}$, there is an integer $j \in \{0, \dots, 2^m - 1\}$ such that

$$J = [\alpha, \beta) := [j2^{-m}, (j+1)2^{-m}), \qquad \overline{J} \subset \mathrm{int}(I_x).$$

Let s be the m-bit binary representation of j, including leading zeros. Then s determines J and

$$|s| = m \le B + 3.$$

This is an existential choice of s. The decoder recovers $m = |s|$ from s; it does not need to compute B or $\lceil B \rceil$, or to know x or $\ell(x)$.

Step 3: Compute a deterministic extension of x. Given p, s, and a requested length $n \ge 0$, start with $y = \epsilon$. If $n = 0$, return ϵ. Otherwise, repeatedly perform the following step, maintaining $|J \cap I_y| > 0$. For each $a \in \{0, 1\}$, the real number

$$h_a(y) := \min\{\beta, R_{ya}\} - \max\{\alpha, L_{ya}\}$$

is computable uniformly from p, s, y, a, and is positive exactly when $J \cap I_{ya}$ has positive length. For $k = 1, 2, \dots$, compute rational approximations for both bits satisfying

$$|\widehat{h}_{a,k}(y) - h_a(y)| \le 2^{-k}.$$

At the first k for which $\widehat{h}_{a,k}(y) > 2^{-k}$ for some a, choose the smallest such bit and replace y by ya. Repeat until $\ell(y) = n$.

The invariant holds initially. Since the child intervals partition I_y, at least one child has positive-length intersection with J. Its positivity is eventually certified by the above test, so each iteration terminates

and preserves the invariant. The selection rule is independent of n; hence the outputs are consistent prefixes of a single computable sequence ω. Because $J \subset \mathrm{int}(I_x)$, before x has been completed only the child consistent with x can intersect J in positive length. Thus ω extends x. Let δ_ω be the probability measure concentrated on ω; then $\delta_\omega(x) = 1$.

Step 4: Encode a program computing the cylinder probabilities. Fix a computable prefix-free encoding $s \mapsto \overline{s}$ with

$$|\overline{s}| \le |s| + 2\log_2(|s|+1) + c_{\mathrm{enc}},$$

where c_{enc} is an absolute constant. Such an encoding is obtained by encoding $|s|+1$ self-delimitingly and then appending s.

Let $z_0 := \langle \epsilon, 1 \rangle$. Define a fixed machine W as follows. On every auxiliary input, it first simulates U on its external program tape with auxiliary input z_0, recording the program prefix up to the simulated head position when the simulation halts. For our chosen p, this identifies and stores exactly p, because p is (U, z_0)-self-delimiting. The machine then reads and decodes $\overline{s}$, leaving its external program head at the last symbol of $\overline{s}$.

On auxiliary input $\langle y, r \rangle$, the machine uses Step 3 to compute the first $\ell(y)$ bits of ω and outputs the rational number 1 if these bits equal y, and 0 otherwise. All probability computations simulate U using the stored copy of p on work tapes; the external program head does not move after the initial parsing. This parsing is performed on every query, including $y = \epsilon$.

For each auxiliary input z, the set of (W, z)-self-delimiting programs is a subset of

$$D_{U,z_0}\{\overline{s} : s \in \{0,1\}^*\},$$

where D_{U,z_0} is the set of (U, z_0)-self-delimiting programs. Both factors are prefix-free, so their concatenation is prefix-free. The machine need not determine whether a parsed program computes a measure; on other programs, its computation may fail to halt. For the chosen p, s, however, Step 3 guarantees that it halts on every query with its program head at the last symbol of $p\overline{s}$, and

$$W(p\overline{s}; \langle y, r \rangle) = \delta_\omega(y) \qquad \text{for every } y \text{ and } r.$$

Step 5: Apply universality and bound the program length. By (6), there is a fixed compiler prefix e_W such that $p_\omega := e_W p\overline{s}$ is $(U, \langle y, r \rangle)$-self-delimiting and

$$\left|U(p_\omega; \langle y, r \rangle) - \delta_\omega(y)\right| = 0 \le 2^{-r} \qquad \text{for every } y \text{ and } r.$$

Thus δ_ω is estimable and p_ω computes it in the sense of Definition 2.3. Consequently,

$$\begin{aligned} K_U(\delta_\omega) &\le |e_W| + |p| + |\overline{s}| \\ &\le K_U(\nu) + m + 2\log_2(m+1) + O(1) \\ &\le K_U(\nu) + B + 2\log_2(B+2) + O(1), \end{aligned}$$

where the last inequality uses $m \le B + 3$ and $m + 1 \le 2(B+2)$. All constants depend only on the fixed reference machine and coding conventions. Substituting $B = \log_2 \nu(x)^{-1}$ and choosing c_{det} sufficiently large proves the claim. □

We next apply Lemma A.4 to an exact MDL minimizer. When $0 < \lambda < 1$, the reduction in log-loss offsets the additional model complexity up to a constant depending on λ but not on the observed prefix.

Proof of Lemma 3.1. Set $\nu^\star := \hat{\nu}_x^\lambda$, as fixed in (8), and write

$$B := \log_2 \nu^\star(x)^{-1} < \infty.$$

By Lemma A.4, there exists a deterministic estimable measure ν_x with $\nu_x(x) = 1$ and

$$K(\nu_x) \le K(\nu^\star) + B + c_{\mathrm{det}} \log_2(B+2) + c_{\mathrm{det}}.$$

Therefore,

$$L_{\lambda, w^U}(\nu_x; x) = \lambda K(\nu_x) \le \lambda K(\nu^\star) + \lambda B + \lambda c_{\mathrm{det}} \log_2(B+2) + \lambda c_{\mathrm{det}}.$$

Since

$$\mathrm{OPT}_{\lambda, w^U}(x) = L_{\lambda, w^U}(\nu^\star; x) = \lambda K(\nu^\star) + B,$$

we obtain

$$L_{\lambda,w^U}(\nu_x;x) - \mathrm{OPT}_{\lambda,w^U}(x) \le \lambda c_{\mathrm{det}} \log_2(B+2) + \lambda c_{\mathrm{det}} - (1-\lambda)B.$$

Using $\lambda < 1$ and the elementary inequality $\ln z \le z - 1$ for $z > 0$, we obtain

$$\begin{aligned}
\lambda c_{\mathrm{det}} \log_2(B+2) + \lambda c_{\mathrm{det}} - (1-\lambda)B &\le c_{\mathrm{det}} \log_2(B+2) + c_{\mathrm{det}} - (1-\lambda)B \\
&\le c_{\mathrm{det}} \log_2 \frac{1}{1-\lambda} + c_{\mathrm{det}} \log_2 \frac{c_{\mathrm{det}}}{\ln 2} + c_{\mathrm{det}} - \frac{c_{\mathrm{det}}}{\ln 2} + 2 \\
&= c_1 \log_2 \frac{1}{1-\lambda} + c_2,
\end{aligned}$$

where $c_1 := c_{\mathrm{det}}$ and $c_2 := c_{\mathrm{det}} \log_2 \frac{c_{\mathrm{det}}}{\ln 2} + c_{\mathrm{det}} - \frac{c_{\mathrm{det}}}{\ln 2} + 2$. Both constants depend only on the fixed reference machine and coding conventions. Therefore, setting

$$C_\lambda := c_1 \log_2 \frac{1}{1-\lambda} + c_2,$$

we conclude that

$$L_{\lambda,w^U}(\nu_x;x) \le \mathrm{OPT}_{\lambda,w^U}(x) + C_\lambda.$$

□

Lemma 3.1 provides an admissible deterministic model after every history. We obtain the sequential counterexample by choosing the fair-coin measure as the true environment.

Proof of Theorem 3.2. Fix $\lambda \in (0,1)$. Let the true environment be the fair-coin measure

$$\mu := \mathrm{Bern}(1/2)^\infty.$$

This measure is estimable, so $\mu \in \mathcal{M}_U$. For each history $x \in \{0,1\}^*$, let $S(x) := \nu_x$ be the deterministic measure constructed in Lemma 3.1. The lemma implies that for every history x,

$$L_{\lambda,w^U}(S(x);x) \le \mathrm{OPT}_{\lambda,w^U}(x) + C_\lambda.$$

Hence S is a valid additive-C_λ Approx-MDL selector. At history $x_{<t}$, select the model $\tilde{\rho}_t^+ := S(x_{<t})$, whose next-token probabilities are

$$\tilde{\rho}_t^+(\cdot \mid x_{<t}) := S(x_{<t})(\cdot \mid x_{<t}).$$

Now fix any history x. Since $S(x)$ is deterministic, there exists a bit $b(x) \in \{0,1\}$ such that

$$S(x)(\cdot \mid x) = \delta_{b(x)}.$$

On the other hand, the true environment μ always predicts

$$\mu(\cdot \mid x) = (1/2, 1/2).$$

Therefore, the one-step squared loss is exactly

$$\sum_{a\in\{0,1\}} \Big(\mu(a \mid x) - S(x)(a \mid x)\Big)^2 = \left(\frac{1}{2} - 1\right)^2 + \left(\frac{1}{2} - 0\right)^2 = \frac{1}{2}.$$

This holds for every history x, so for every $t \ge 1$, the expected error remains constant:

$$\mathbb{E}_\mu\left[\sum_{a\in\{0,1\}} \Big(\mu(a \mid X_{<t}) - \tilde{\rho}_t^+(a \mid X_{<t})\Big)^2\right] = \frac{1}{2}.$$

Hence, the cumulative squared error diverges:

$$S_\infty(\mu, \tilde{\rho}^+) = \sum_{t\ge 1} \frac{1}{2} = +\infty.$$

Because the expected single-step prediction error fails to converge to zero, the predictor $\tilde{\rho}^+$ does not universally learn the estimable environment μ. □

Theorem A.1. *Consider the universal model class $\mathcal{M}_U$ with weights $w_\nu^U = 2^{-K(\nu)}$, and let $\mu = \mathrm{Bern}(1/2)^\infty$ be the fair-coin measure. For every $\lambda \in (0,1)$ and every finite $C \geq 0$, there exist an integer $N \geq 1$ and a constant $c > 0$, depending only on λ, C, the fixed reference machine, and the fixed coding conventions, such that every valid additive approximate MDL predictor $\tilde{\rho}^+ = (\tilde{\rho}_t^+)_{t\geq 1}$ satisfying Equation (2) with slack C obeys*

$$\sum_{t=1}^{T} e_t(\mu, \tilde{\rho}^+) \geq c \left\lfloor \frac{T}{N} \right\rfloor \qquad \text{for every integer } T \geq 1.$$

Consequently,

$$S_\infty(\mu, \tilde{\rho}^+) = +\infty, \qquad \limsup_{t\to\infty} e_t(\mu, \tilde{\rho}^+) \geq \frac{c}{N} > 0.$$

In particular, these conclusions hold for exact MDL, corresponding to $C = 0$, and the same true environment μ works for all $\lambda \in (0,1)$ and all finite $C \geq 0$.

Proof. Fix $\lambda \in (0,1)$ and a finite $C \geq 0$. Throughout the proof, write

$$L_\lambda(\nu; x) := L_{\lambda, w^U}(\nu; x) = \lambda K(\nu) - \log_2 \nu(x), \qquad \mathrm{OPT}_\lambda(x) := \inf_{\nu\in\mathcal{M}_U} L_\lambda(\nu; x).$$

Call a model ν *admissible at* x if $L_\lambda(\nu; x) \leq \mathrm{OPT}_\lambda(x) + C$. The infimum is finite and attained at every history: the deterministic measure supported on $x0^\infty$ has a finite score R_x, whereas any model with score at most R_x has $K(\nu) \leq R_x/\lambda$, and there are only finitely many measures of bounded complexity. Thus admissible models exist even when $C = 0$, and all admissible models have $\nu(x) > 0$.

Let U denote the fixed universal prefix machine. The conditional prefix complexity of a string x given y is defined as $K_{\mathrm{f}}(x \mid y) = \min\{|p| : U(p; y) = x\}$. In particular, $K_{\mathrm{f}}(x) := K(x \mid \epsilon)$ where ϵ is the empty string. This notation is distinct from the measure complexity $K(\nu)$. The prefix complexity addition theorem gives

$$K_{\mathrm{f}}(x, y) = K_{\mathrm{f}}(x) + K_{\mathrm{f}}\big(y \mid x, K_{\mathrm{f}}(x)\big) + O(1), \tag{14}$$

where the constant is independent of y, k; see, e.g., Li and Vitányi (2008).

Step 1: Every admissible model assigns a uniformly positive probability to the observed history. Let ν be admissible at x and put $B := -\log_2 \nu(x) \geq 0$. By Lemma A.4, there is a deterministic estimable measure δ_ω such that $\delta_\omega(x) = 1$ and

$$K(\delta_\omega) \leq K(\nu) + B + c_{\mathrm{det}} \log_2(B + 2) + c_{\mathrm{det}}.$$

Comparing the admissible score with this candidate gives

$$\lambda K(\nu) + B \leq \mathrm{OPT}_\lambda(x) + C \leq \lambda K(\delta_\omega) + C \leq \lambda K(\nu) + \lambda B + \lambda c_{\mathrm{det}} \log_2(B + 2) + \lambda c_{\mathrm{det}} + C.$$

Hence

$$(1 - \lambda)B \leq \lambda c_{\mathrm{det}} \log_2(B + 2) + \lambda c_{\mathrm{det}} + C.$$

Since $1 - \lambda > 0$, the left-hand side grows linearly in B, whereas the right-hand side grows logarithmically. There is therefore an integer $b_0 \geq 0$, depending only on the fixed parameters, such that every admissible pair (x, ν) has $B \leq b_0$. Set $\beta := 2^{-b_0} > 0$. Then

$$\nu(x) \geq \beta \qquad \text{whenever } \nu \text{ is admissible at } x. \tag{15}$$

Step 2: A history at which an admissible model predicts both bits with substantial probability has a short description. Fix the rational number $\tau := \beta/8$, and for $\nu \in \mathcal{M}_U$ define

$$E_\nu := \{z \in \{0,1\}^* : \nu(z0) > \tau \text{ and } \nu(z1) > \tau\}.$$

We first show that $|E_\nu| < 1/\tau$. Consider the finite prefix tree consisting of all paths from the root to the children $z0, z1$, $z \in E_\nu$. Every member of E_ν is a internal node, so the tree has at least $|E_\nu| + 1$ leaves. Each leaf is a child of some member of E_ν, and therefore has ν-probability greater than τ. Since the leaf cylinders are disjoint, $(|E_\nu| + 1)\tau < 1$.

Given a program computing ν, we can effectively enumerate E_ν without repetition: dovetail over strings and precision levels, and enumerate a string once both strict inequalities have been certified.

Describe $z \in E_\nu$ by a shortest program for ν, followed by its rank in this enumeration. The rank uses a fixed number of bits depending only on τ. To parse this description under Definition 2.3, the decoder first simulates the measure program on the fixed query $\langle \epsilon, 1 \rangle$, where ϵ is the empty string, and records the program bits read. Self-delimitation identifies the end of the program, after which the decoder reads the fixed-length rank and uses the recorded program for all subsequent simulations. It halts when the specified rank is enumerated; it need not determine when the enumeration is complete. Thus there is a constant d_τ, independent of ν, z, such that

$$z \in E_\nu \quad \Longrightarrow \quad K_{\mathrm{f}}(z) \le K(\nu) + d_\tau. \tag{16}$$

Call ν *balanced at* x if $\nu(1 \mid x) \in [1/4, 3/4]$. If ν is admissible and balanced at x, then (15) implies $\nu(xa) \ge \beta/4 > \tau$ for both $a \in \{0, 1\}$. Consequently,

$$K_{\mathrm{f}}(x) \le K(\nu) + d_\tau \qquad \text{if } \nu \text{ is admissible and balanced at } x. \tag{17}$$

Step 3: Only boundedly many positions along a zero continuation admit a balanced model. Fix a finite string y ending in 1, and let $\eta_y := \delta_{y0^\infty}$. A fixed compiler transforms a prefix program for y into a program computing all cylinder probabilities of η_y. Thus there is a constant d_0, independent of y, such that

$$K(\eta_y) \le K_{\mathrm{f}}(y) + d_0.$$

For every integer $k \ge 1$, we have $\eta_y(y0^k) = 1$. If ν is admissible at $y0^k$, then

$$\lambda K(\nu) \le L_\lambda(\nu; y0^k) \le \mathrm{OPT}_\lambda(y0^k) + C \le \lambda K(\eta_y) + C \le \lambda K_{\mathrm{f}}(y) + \lambda d_0 + C.$$

If ν is also balanced there, then (17) yields

$$K_{\mathrm{f}}(y0^k) \le K_{\mathrm{f}}(y) + d_0 + C/\lambda + d_\tau.$$

Because y ends in 1, the pair (y, k) can be computably recovered from $y0^k$ by stripping its trailing zeros. Combining this observation with (14), we obtain an integer $D \ge 0$, independent of y, k, and the selected model, such that

$$K_{\mathrm{f}}\big(k \mid y, K_{\mathrm{f}}(y)\big) \le D$$

whenever some admissible model is balanced at $y0^k$.

For each fixed auxiliary input $(y, K_{\mathrm{f}}(y))$, fewer than 2^{D+1} binary programs have length at most D. Set $M := 2^{D+1}$ and $L := M + 1$. At most M positive integers k can admit an admissible balanced model at $y0^k$. Therefore, for every y ending in 1, there is a $k_y \in \{1, \dots, L\}$ such that *every* admissible model at $y0^{k_y}$ is unbalanced. The constants are uniform in y; the argument does not require computing $K_{\mathrm{f}}(y)$ or finding k_y effectively.

Step 4: Every time block incurs a fixed positive expected prediction error. Fix any valid additive approximate MDL predictor $\tilde{\rho}^+$ with slack C. For each history x, define

$$\Delta(x) := \sum_{a \in \{0,1\}} \left(\frac{1}{2} - \tilde{\rho}^+_{\ell(x)+1}(a \mid x) \right)^2 = 2 \left(\tilde{\rho}^+_{\ell(x)+1}(1 \mid x) - \frac{1}{2} \right)^2.$$

At $y0^{k_y}$, every admissible model is unbalanced, so

$$\Delta(y0^{k_y}) \ge \frac{1}{8} \qquad \text{for every } y \text{ ending in } 1.$$

Set $N := L + 2$, which will be the fixed length of each time block. This constant depends only on λ, C, the reference machine, and the coding conventions; it is independent of the starting time, the observed history, and the predictor. The two additional steps account for the initial 1 and the prediction following the last zero in the continuation 10^L. Fix an integer $n \ge 0$ and condition on an arbitrary history $X_{1:n} = h$. Under the fair-coin measure, the event

$$A_h := \{X_{n+1:n+L+1} = 10^L\}$$

has conditional probability 2^{-L-1}. On this event, with $y := h1$, the predictor visits every history $y0^k$, $1 \le k \le L$. In particular, the prediction following $y0^{k_y}$ occurs at time $n + k_y + 2 \le n + N$. Nonnegativity of the other losses therefore gives

$$\mathbb{E}_\mu \left[\sum_{t=n+1}^{n+N} \Delta(X_{<t}) \,\middle|\, X_{1:n} = h \right] \ge \frac{1}{8} \, \mathbb{P}_\mu(A_h \mid X_{1:n} = h) = 2^{-L-4}.$$

This bound is uniform in n, h, and the predictor. Let $c := 2^{-L-4} > 0$. Applying the bound to disjoint blocks starting at $n = jN$ and taking expectations yields

$$\sum_{t=1}^{T} e_t(\mu, \tilde{\rho}^+) \geq \sum_{j=0}^{\lfloor T/N \rfloor - 1} \sum_{t=jN+1}^{(j+1)N} e_t(\mu, \tilde{\rho}^+) \geq c \left\lfloor \frac{T}{N} \right\rfloor.$$

Thus $S_\infty(\mu, \tilde{\rho}^+) = +\infty$ and

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} e_t(\mu, \tilde{\rho}^+) \geq \frac{c}{N} > 0.$$

Since $0 \leq e_t(\mu, \tilde{\rho}^+) \leq 1/2$, its limit superior is at least the limit inferior of its Cesàro means. This proves the remaining assertion. □

B Proofs for multiplicative MDL approximation: failure of universal prediction guarantees

We first establish Lemma 4.1, which gives a high-probability lower bound on the optimal universal MDL score in terms of the description length under the true environment. This bound will be used to show that models with persistent prediction error can satisfy the multiplicative approximation condition with probability bounded away from zero.

Proof of Lemma 4.1. Define the mixture over all estimable measures weighted by their prior:

$$\xi_\lambda(x) := \sum_{\nu \in \mathcal{M}_U} (w_\nu^U)^\lambda \nu(x) = \sum_{\nu \in \mathcal{M}_U} 2^{-\lambda K(\nu)} \nu(x).$$

For every n,

$$\sum_{x \in \{0,1\}^n} \xi_\lambda(x) = \sum_{\nu \in \mathcal{M}_U} 2^{-\lambda K(\nu)} \sum_{x \in \{0,1\}^n} \nu(x) \leq \sum_{\nu \in \mathcal{M}_U} 2^{-\lambda K(\nu)} \leq \sum_{\nu \in \mathcal{M}_U} 2^{-K(\nu)} \leq 1,$$

where we used $\lambda \geq 1$ and Kraft's inequality for the prefix-free codes underlying the Kolmogorov complexity. Thus ξ_λ is a semimeasure.

For any $\nu \in \mathcal{M}_U$ and any string x, $\xi_\lambda(x) \geq 2^{-\lambda K(\nu)} \nu(x)$, so

$$-\log_2 \xi_\lambda(x) \leq \lambda K(\nu) + \log_2 \nu(x)^{-1} = L_{\lambda, w^U}(\nu; x).$$

Taking the minimum over ν gives

$$-\log_2 \xi_\lambda(x) \leq \mathrm{OPT}_{\lambda, w^U}(x). \tag{18}$$

Now define the nonnegative random variable $Z_n := \frac{\xi_\lambda(X_{1:n})}{\mu(X_{1:n})}$ with the convention $Z_n := 0$ whenever $\mu(X_{1:n}) = 0$. Then

$$\mathbb{E}_\mu[Z_n] = \sum_{\substack{x \in \{0,1\}^n \\ \mu(x) > 0}} \xi_\lambda(x) \leq \sum_{x \in \{0,1\}^n} \xi_\lambda(x) \leq 1.$$

By Markov's inequality, $\Pr_\mu[Z_n \geq 2^c] \leq 2^{-c}$. Equivalently, with probability at least $1 - 2^{-c}$,

$$\xi_\lambda(X_{1:n}) \leq 2^c \mu(X_{1:n}),$$

that is,

$$-\log_2 \mu(X_{1:n}) \leq -\log_2 \xi_\lambda(X_{1:n}) + c.$$

Combining this with (18) yields

$$\mathrm{OPT}_{\lambda, w^U}(X_{1:n}) \geq \log_2 \mu(X_{1:n})^{-1} - c$$

with probability at least $1 - 2^{-c}$. □

We now prove Theorem 4.1. The lower bound from Lemma 4.1 will allow us to show that models with persistent prediction error satisfy the multiplicative approximation condition with probability tending to one. We construct such models separately for $0 < \lambda < 1$ and $\lambda \geq 1$, and then use a common selection rule to obtain a valid predictor with divergent cumulative expected squared error.

Proof of Theorem 4.1. Fix $\lambda > 0$ and $\varepsilon > 0$, and let $\mu = \mathrm{Bern}(1/2)^\infty$. For brevity, write $L_\beta = L_{\beta,w^U}$ and $\mathrm{OPT}_\beta = \mathrm{OPT}_{\beta,w^U}$.

For every fixed $\beta \geq 1$ and $\delta \in (0,1)$, Lemma 4.1, applied with $c = \delta n$, gives

$$\mathbb{P}_\mu[\mathrm{OPT}_\beta(X_{1:n}) < (1-\delta)n] \leq 2^{-\delta n},$$

since $-\log_2 \mu(X_{1:n}) = n$. Also, using μ itself as a candidate yields $\mathrm{OPT}_\beta(X_{1:n}) \leq n + \beta K(\mu)$. Consequently,

$$\frac{\mathrm{OPT}_\beta(X_{1:n})}{n} \xrightarrow{\mathbb{P}_\mu} 1.$$

We next choose candidate measures ν_x whose one-step squared prediction error under μ is a fixed positive constant, while

$$A_n := \{L_\lambda(\nu_{X_{1:n}}; X_{1:n}) \leq (1+\varepsilon)\,\mathrm{OPT}_\lambda(X_{1:n})\}$$

satisfies $\mathbb{P}_\mu(A_n) \to 1$.

We split the proof into two cases.

Case 1: $0 < \lambda < 1$**.** The construction in the proof of Lemma 3.1 gives a constant $C_\lambda < \infty$ and, for every history x, a deterministic estimable measure ν_x such that

$$\nu_x(x) = 1, \qquad L_\lambda(\nu_x; x) \leq \mathrm{OPT}_\lambda(x) + C_\lambda.$$

Since $-\log_2 \nu(x) \geq 0$, we have $L_\lambda(\nu; x) \geq \lambda L_1(\nu; x)$ for every ν and x. Taking infima and using Step 1 therefore gives

$$\mathrm{OPT}_\lambda(X_{1:n}) \geq \lambda\,\mathrm{OPT}_1(X_{1:n}) \xrightarrow{\mathbb{P}_\mu} +\infty.$$

It follows that

$$\mathbb{P}_\mu(A_n) \geq \mathbb{P}_\mu[\mathrm{OPT}_\lambda(X_{1:n}) \geq C_\lambda/\varepsilon] \longrightarrow 1.$$

Each ν_x predicts the next bit deterministically, so its one-step squared error against the fair-coin law is $1/2$.

Case 2: $\lambda \geq 1$**.** Choose a rational $\theta \in (0,1) \setminus \{1/2\}$ sufficiently close to $1/2$ that

$$h_\theta := -\frac{1}{2}\log_2 \theta - \frac{1}{2}\log_2(1-\theta) < 1 + \varepsilon.$$

Set $\nu_x = \nu_\theta := \mathrm{Bern}(\theta)^\infty$ for every x. This measure is estimable, and the law of large numbers gives

$$\frac{L_\lambda(\nu_\theta; X_{1:n})}{n} = \frac{\lambda K(\nu_\theta)}{n} - \frac{1}{n}\log_2 \nu_\theta(X_{1:n}) \xrightarrow{\mathbb{P}_\mu} h_\theta.$$

Combining this with Step 1 yields

$$\frac{L_\lambda(\nu_\theta; X_{1:n}) - (1+\varepsilon)\,\mathrm{OPT}_\lambda(X_{1:n})}{n} \xrightarrow{\mathbb{P}_\mu} h_\theta - (1+\varepsilon) < 0.$$

Thus $\mathbb{P}_\mu(A_n) \to 1$. The candidate's one-step squared error is the positive constant $2(\theta - 1/2)^2$ at every history.

Using the candidates from the applicable case, define

$$S(x) := \begin{cases} \nu_x, & L_\lambda(\nu_x; x) \leq (1+\varepsilon)\,\mathrm{OPT}_\lambda(x), \\ \hat{\nu}^\lambda_x, & \text{otherwise}, \end{cases}$$

where $\hat{\nu}^\lambda_x$ is the exact minimizer from (8). At history $x_{1:n}$, let $\tilde{\rho}^\times_{n+1} := S(x_{1:n})$. By construction, this predictor satisfies the multiplicative approximation condition at every history. All selected measures assign positive probability to the observed history, so their conditional predictions are well defined.

On A_n, the selected model is $\nu_{X_{1:n}}$, whose squared prediction error equals a constant $d > 0$: namely, $d = 1/2$ in Case 1 and $d = 2(\theta - 1/2)^2$ in Case 2. Hence

$$\mathbb{E}_\mu\left[\sum_{a\in\{0,1\}}\left(\mu(a \mid X_{1:n}) - \tilde{\rho}^{\times}_{n+1}(a \mid X_{1:n})\right)^2\right] \geq d\,\mathbb{P}_\mu(A_n).$$

Since $\mathbb{P}_\mu(A_n) \to 1$, the expected one-step squared error is at least $d/2$ for all sufficiently large n. Summing over n proves $S_\infty(\mu, \tilde{\rho}^{\times}) = +\infty$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main finite-error guarantee and failure results, and the formal theorem statements and proofs appear in Sections 3 and 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses the theoretical scope, realizability/countability/estimability assumptions, and the absence of empirical algorithms or experiments.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in the definitions and theorem statements; complete proofs appear in Sections 3, 4, and Appendices A and B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: The paper does not include experiments; the main contributions are theoretical results and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not introduce experiments requiring code or datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experimental training or test settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments, so statistical significance reporting is not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments that require compute-resource disclosure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is theoretical and does not involve human subjects, private data, released models, or other activities that would require special ethical safeguards.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is a foundational research and not tied to particular applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with high misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use datasets, codebases, or models as research assets; it cites prior scholarly work in the bibliography.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, codebases, or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subjects research or crowdsourcing, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The usage of LLMs is described in the acknowledgement.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.