An Information-Theoretic Analysis of In-Context Learning

Hong Jun Jeon [1] Jason D. Lee [2] Qi Lei [3] Benjamin Van Roy [4]

Abstract

Previous theoretical results pertaining to meta-learning on sequences build on contrived and convoluted mixing time assumptions. We introduce new information-theoretic tools that lead to a concise yet general decomposition of error for a Bayes optimal predictor into two components: meta-learning error and intra-task error. These tools unify analyses across many meta-learning challenges. To illustrate, we apply them to establish new results about in-context learning with transformers and corroborate existing results a simple linear setting. Our theoretical results characterize how error decays in both the number of training sequences and sequence lengths. Our results are very general; for example, they avoid contrived mixing time assumptions made by all prior results that establish decay of error with sequence length.

1. Introduction

In recent years, we have observed the capability of large language models (LLMs) to learn from data within just its context window. This puzzling phenomenon referred to as in-context learning (ICL) (Brown et al., 2020), has captured the attention of the theoretical machine learning community. As the data available in-context is dwarfed by the extensive pretraining set, meta-learning stands as a prevailing explanation for ICL (Xie et al., 2022).

As aforementioned, Xie et al. (2022) introduced the idea that ICL could be interpreted as implicit Bayesian inference within a mixture of HMMs. While their theoretical results rely on contrived assumptions and fail to explain how ICL

[1]Department of Computer Science, Stanford University, Stanford, CA, USA [2]Princeton University, Princeton, NJ, USA [3]New York University, New York City, NY, USA [4]Stanford University, Stanford, CA, USA. Correspondence to: Hong Jun Jeon <hjjeon@stanford.edu>.

is possible with such short sequences, their work *initiated* the study of modeling ICL as Bayesian inference or other thoroughly studied learning processes such as empirical risk minimization. As much of the theoretical community is most familiar with error analyses of empirical risk minimization, much of the existing results (Li et al., 2023a; Bai et al., 2023; Edelman et al., 2021) study the error of an ICL under the assumption that ICL is competitive in out-of-sample performance with empirical risk minimization. However, each of these error bounds is limited in some way such as exponential depth dependence (Edelman et al., 2021; Li et al., 2023a) or error which decays only with the number of sequences and not the length of the sequences (Edelman et al., 2021; Bai et al., 2023). The results which do demonstrate that error decays in both the number of training sequences and sequence length often rely on contrived mixing time assumptions (Zhang et al., 2023b) or stability conditions which are equivalent to fast mixing (Li et al., 2023a).

Our work revisits the idea of modeling ICL as Bayesian inference. In this work, we introduce new information-theoretic tools based on work by Jeon et al. (2023) which lead to a concise yet general decomposition of error in meta-learning from sequences. This decomposition *exactly* characterizes error of a Bayes optimal predictor as the sum of two components: meta-learning error, and intra-task error. Notably, our results (e.g. Theorem 4.2) show the meta-learning error is equal to

$$\frac{\mathbb{I}(\text{all data}; \text{meta parameters})}{\text{number of sequences} \times \text{sequence length}}$$

and intra-task error is equal to

$$\frac{\mathbb{I}(\text{task data}; \text{task parameters}|\text{meta parameters})}{\text{sequence length}}.$$

Our results also do not make contrived assumptions on the data generating process such as stability or fast mixing within the sequence. To demonstrate the use of our theory, we specialize it to produce novel results for meta-learning a sparse mixture of transformer models. The result provides a compelling narrative as to how ICL is possible with such few examples.

As some of our tools are non-standard to much of the community, we begin by introducing our framework in the sim-

1

pler setting of learning from a single sequence of data. In the following section, we naturally extend the analysis to meta-learning from many sequences and present our main result (Theorem 4.2). Since our results are general and abstract, we demonstrate the application of these results to several concrete problem instances. In the main text, we provide concrete examples that resemble learning from data generated by a deep transformer model, and in the appendix, we provide simpler problem instances for reference (logistic regression, linear representation learning). While our results in the main text are limited to the analysis of the Bayes optimal predictor, we provide results in Appendix C which provide insights about suboptimal predictors.

2. Related Works

In-context Learning and Transformer. LLMs based on the transformer architecture (Vaswani et al., 2023) have exhibited the ability to learn from data within the context of a prompt (Brown et al., 2020). This phenomenon, referred to as in-context learning (ICL), has received significant empirical investigation (Liu et al., 2021; Min et al., 2021; Lu et al., 2021; Zhao et al., 2021; Rubin et al., 2021; Elhage et al., 2021; Kirsch et al., 2022; Wei et al., 2023; Brown et al., 2020; Dong et al., 2022).

However, theoretical understanding of ICL is still relatively nascent (Xie et al., 2022; Garg et al., 2022; Von Oswald et al., 2023; Dai et al., 2022; Giannou et al., 2023; Li et al., 2023a; Raventos et al., 2023). Among the existing theoretical work, most focuses on the optimization dynamics (Tian et al., 2023a;b; Jelassi et al., 2022; Li et al., 2023b; Tarzanagh et al., 2023; Zhang et al., 2023a; Huang et al., 2023; Ahn et al., 2023; Mahankali et al., 2023) or the representation power (Sanford et al., 2023; Song & Zhong, 2023; Von Oswald et al., 2023; Giannou et al., 2023; Liu et al., 2022) regarding the transformer architecture. In the realm of statistical results, much of the existing work is confined to how transformers can perform ICL by simulating gradient descent (Von Oswald et al., 2023; Akyürek et al., 2022; Dai et al., 2022; Giannou et al., 2023). However, as they provide no concrete sample complexity results, they are not directly comparable to our work. The work that is perhaps most relevant to ours include those which analyze the sample complexity of ICL under the assumption that its performance is comparable to empirical risk minimization or Bayesian inference (Xie et al., 2022; Li et al., 2023a; Bai et al., 2023; Edelman et al., 2021; Zhang et al., 2023b). Despite their quantitative sample complexity results, as mentioned in the introduction, these results are ultimately limited by either their restrictive assumptions on mixing times of the data sequence or their inability to capture how sequence length contributes to reduction in error.

Meta-learning. As our work analyzes ICL under the lens of meta-learning, we provide a brief exposition of its existing work. Recent empirical advancements have sparked interest in the theoretical foundations of meta-learning (Baxter, 2000; Denevi et al., 2018; Finn et al., 2019). In settings such as tasks drawn from a shared meta-distribution, several works (Maurer, 2009; Pontil & Maurer, 2013; Maurer et al., 2016) have derived generalization bounds albeit for simplistic settings such as linear representation or linear classifiers. Under strong assumptions such as large margin or large number of tasks Srebro & Ben-David (2006); Aliakbarpour et al. (2023) were also able to establish such bounds. These results all rely on the assumption that the data *within* each meta-task is independently and identically distributed (iid) under an (unknown) probability distribution. However, in the context of LLMs, for which the meta-tasks are separate documents, the sequence of tokens within each document is certainly not iid.

Learning as Compression. Lastly, we believe that this work provides a mathematically rigorous connection between effective learning and *compression*, a viewpoint espoused by (Delétang et al., 2024). They colloquially and empirically draw the connection between a large language model's ability to compress data and its ability to make predictions. We believe that Theorem 3.4 establishes this connection by upper and lower bounding the predictive performance of the Bayes optimal predictor by the *rate-distortion function* which characterizes optimal lossy compression. We hope that this mathematical derivation lends further credence and elucidation to the "prediction is compression" viewpoint.

3. Learning from Sequential Data

We begin by introducing our general information-theoretic tools for the analysis of standard *supervised learning* on sequential data. Examples of such learning problems include but are not limited to natural language modeling and learning from video/audio data. Results from this section draw inspiration from (Jeon et al., 2023) which focused on the analysis of supervised learning from *iid* data.

We model all uncertain quantities as random variables. Each random variable is defined with respect to a common probability space $(\Omega, \mathbb{F}, \mathbb{P})$. Of particular interest to our analysis is a sequence $X_1, X_2, \ldots, X_T$ of discrete random variables which represent observations. This sequence is generated by an *autoregressive* model parameterized by a random variable θ such that for all $t \in \mathbb{Z}_+$, X_{t+1} may depend on θ and the entire history $X_1, \ldots, X_t$, which we abbreviate as H_t.

3.1. Bayesian Error

Our framework is *Bayesian* in the sense that it treats learning as the process of reducing uncertainty about the random variable θ. A learning algorithm produces, for each t, a *predictive distribution* P_t of X_{t+1} after observing the history H_t. We express such an algorithm in terms of a function π for which $P_t = \pi(H_t)$. For a horizon $T \in \mathbb{Z}_{++}$, it is common to quantify the error realized by predictions P_t for $t < T$ in terms of the average cumulative expected log-loss:

$$\mathbb{L}_{T,\pi} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_\pi \left[- \ln P_t(X_{t+1}) \right].$$

However, the minimum achievable log-loss is not 0. For instance, consider $P_t = \mathbb{P}(X_{t+1} \in \cdot | \theta, H_t)$. Even this *omniscient* predictor incurs a Bayesian error of $\mathbb{H}(H_T|\theta)/T$, where $\mathbb{H}$ denote the (conditional) entropy and $\mathbb{I}$ the (conditional) mutual information:

$$\mathbb{H}(X|Y) = \mathbb{E}\left[- \ln \mathbb{P}(X|Y) \right]$$
$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y).$$

As a result, we define *estimation error* as:

$$\mathcal{L}_{T,\pi} = \mathbb{L}_{T,\pi} - \underbrace{\frac{\mathbb{H}(H_T|\theta)}{T}}_{\substack{\text{irreducible} \\ \text{error}}}.$$

Note that $\mathbb{H}(H_T|\theta)$ is irreducible since it is independent of π. Estimation error represents error which is reducible via learning and should vanish as T goes to ∞ for a competent learning algorithm. Our analysis will focus on characterizing $\mathcal{L}_{T,\pi}$ for an optimal π.

3.2. Achievable Bayesian Error

The following result establishes that across all problem instances, the optimal algorithm π sets $P_t = \mathbb{P}(X_{t+1} \in \cdot | H_t)$ for all t. We denote this *Bayesian posterior* by $\hat{P}_t$.

Lemma 3.1. (Bayesian posterior is optimal) *For all $t \in \mathbb{Z}_+$,*

$$\mathbb{E}\left[- \ln \hat{P}_t(X_{t+1}) \right] = \min_\pi \mathbb{E}_\pi \left[- \ln P_t(X_{t+1}) \right].$$

Proof. We assume equality to hold *almost surely*.

$$\mathbb{E}\left[- \ln P_t(X_{t+1}) | H_t \right]$$
$$= \mathbb{E}\left[- \ln \hat{P}_t(X_{t+1}) + \ln \frac{\hat{P}_t(X_{t+1})}{P_t(X_{t+1})} \Big| H_t \right]$$
$$= \mathbb{E}\left[- \ln \hat{P}_t(X_{t+1}) \Big| H_t \right] + \mathbf{d}_{\mathrm{KL}}(\hat{P}_t \| P_t).$$

The result follows since KL divergence is non-negative. □

We use $\mathcal{L}_T$ to denote the *optimal* achievable Bayesian estimation error:

$$\mathcal{L}_T = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[- \ln \hat{P}_t(X_{t+1}) \right] - \frac{\mathbb{H}(H_T|\theta)}{T}.$$

In the main text we restrict our attention to the study of *optimal* achievable Bayesian error but we provide an extension to arbitrary predictors which depend on the history H_t in Appendix C. The following result provides an exact characterization of $\mathcal{L}_T$.

Theorem 3.2. (Bayesian estimation error) *For all $T \in \mathbb{Z}_+$,*

$$\mathcal{L}_T = \frac{\mathbb{I}(H_T;\theta)}{T}.$$

Proof.

$$\mathcal{L}_T = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[- \ln \hat{P}_t(X_{t+1}) \right] - \frac{\mathbb{H}(H_T|\theta)}{T}$$
$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\ln \frac{\mathbb{P}(X_{t+1}|H_t, \theta)}{\hat{P}_t(X_{t+1})} \right]$$
$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(X_{t+1} \in \cdot | H_t, \theta) \| \hat{P}_t(X_{t+1} \in \cdot)) \right]$$
$$\overset{(a)}{=} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta | H_t)$$
$$\overset{(b)}{=} \frac{\mathbb{I}(H_T;\theta)}{T},$$

where (a) follows from the KL-divergence formulation of conditional mutual information and (b) follows from the chain rule of conditional mutual information. □

Jeon et al. (2023) establish Theorem 3.2 in the setting in which the sequence is iid when conditioned on θ. Our result extends the result to general sequences. $\mathcal{L}_T$ will often vanish as $T \to \infty$. For instance, if $\mathbb{H}(\theta) < \infty$, then this will trivially be the case as $\mathbb{I}(H_t; \theta) \leq \mathbb{H}(\theta)$ for all t. However, even in problems for which $\mathbb{H}(\theta) = \infty$, for example, if θ is a continuous random variable, the estimation error will still often vanish as $T \to \infty$. Note that $\mathbb{H}(\theta)$ should not be confused with $\mathbf{h}(\theta)$, the *differential entropy* of θ. The differential entropy does not capture the same qualitative properties as discrete entropy, namely 1) invariance under change of variables, 2) non-negativity. While *differences* in differential entropy still provide meaningful insight via mutual information ($\mathbb{I}(X;Y) = \mathbf{h}(X) - \mathbf{h}(X|Y)$), the quantity itself is largely vacuous for the purposes of measuring information content and therefore deriving error bounds. The appropriate extension of discrete entropy to continuous random variables can be made via *rate-distortion theory*.

Definition 3.3. (rate-distortion function) Let $\epsilon \geq 0$, $\theta : \Omega \mapsto \Theta$ be a random variable, and ρ a distortion function which maps θ and a random variable $\tilde{\theta}$ to $\Re$. The rate-distortion function evaluated for random variable θ at tolerance ϵ takes the value:

$$\inf_{\tilde{\theta} \in \tilde{\Theta}_\epsilon} \mathbb{I}(\theta; \tilde{\theta}),$$

where

$$\tilde{\Theta}_\epsilon = \left\{ \tilde{\theta} : \rho(\theta, \tilde{\theta}) \leq \epsilon \right\}.$$

One can think of $\tilde{\theta}$ as a lossy *compression* of the random variable θ. The objective $\mathbb{I}(\theta; \tilde{\theta})$, referred to as the *rate*, characterizes the number of nets that $\tilde{\theta}$ retains about θ. Meanwhile, the distortion function ρ characterizes how lossy the compression is. When we apply rate-distortion theory to the analysis of machine learning, we restrict our attention to the case in which

$$\rho(\theta, \tilde{\theta})$$
$$= \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(X_{t+1} \in \cdot | \theta, H_t) \| \mathbb{P}(X_{t+1} \in \cdot | \tilde{\theta}, H_t))\right]$$
$$= \mathbb{I}(X_{t+1}; \theta | \tilde{\theta}, H_t).$$

We assume that $\tilde{\theta} \perp X_{t+1} | (\theta, H_t)$ (the compression $\tilde{\theta}$ does not contain exogenous information about X_{t+1}, such as aleatoric noise, which cannot be determined from (θ, H_t)). We use the notation $\mathbb{H}_{\epsilon, T}(\theta)$ to denote the rate-distortion function w.r.t. this KL-divergence distortion function averaged across horizon T:

$$\mathbb{H}_{\epsilon, T}(\theta) = \inf_{\tilde{\theta} \in \tilde{\Theta}_{\epsilon.T}} \mathbb{I}(\theta; \tilde{\theta}),$$

where

$$\tilde{\Theta}_{\epsilon, T} = \left\{ \tilde{\theta} : \tilde{\theta} \perp H_T | \theta; \quad \frac{\mathbb{I}(H_T; \theta | \tilde{\theta})}{T} \leq \epsilon \right\}.$$

With this notation established, we present the following result for sequential learning. The proof can be found in Appendix A.

Theorem 3.4. (rate-distortion estimation error bound) *For all $T \in \mathbb{Z}_+$,*

$$\sup_{\epsilon \geq 0} \min \left\{ \frac{\mathbb{H}_{\epsilon, T}(\theta)}{T}, \epsilon \right\} \leq \mathcal{L}_T \leq \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{\epsilon, T}(\theta)}{T} + \epsilon.$$

An interpretation of the above result is that the Bayesian posterior implicitly finds the compression $\tilde{\theta}$ that optimally trades off learning complexity $\mathbb{I}(\theta; \tilde{\theta})$ and distortion $\mathbb{I}(H_T; \theta | \tilde{\theta})$. While these results are very general, they remain abstract. In Appendix A.1 we provide a simple logistic regression example. In the main text, we provide an analysis for learning from a sequence generated by a deep transformer model.

3.3. Deep Transformer

In the transformer environment, we let $(X_1, X_2, \ldots)$ be a sequence in $\{1, \ldots, d\}$, where d denotes the size of the vocabulary. Each of the d outcomes is associated with a *known* embedding vector which we denote as Φ_j for $j \in \{1, \ldots, d\}$. We assume that for all j, $\|\Phi_j\|_2 = 1$. For brevity of notation, we let $\phi_t = \Phi_{X_t}$ i.e. the embedding associated with token X_t.

Let K denote the context length of the transformer, L denote it's depth, and r denote the attention dimension. We assume that the first token X_1 is sampled from an arbitrary pmf on $\{1, \ldots, d\}$ but subsequent tokens are sampled based on the previous K tokens within the context window and the weights of a depth L transformer model.

We use $U_{t,\ell}$ to denote the output of layer ℓ at time t ($U_{t,0} = \phi_{t-K+1:t}$) (the embeddings associated with the past K tokens). For all $t \leq T, \ell < L$, let

$$\mathrm{Attn}^{(\ell)}(U_{t,\ell-1}) = \sigma\left(\frac{U_{t,\ell-1}^\top A^{(\ell)} U_{t,\ell-1}}{\sqrt{r}}\right)$$

denote the attention matrix of layer ℓ where σ denotes the softmax function applied elementwise along the columns. The matrix $A^{(\ell)} \in \Re^{r \times r}$ can be interpreted as the product of the key and query matrices and without loss of generality, we assume that the elements of the matrices A_ℓ are distributed iid $\mathcal{N}(0, 1)$.

Subsequently, we let

$$U_{t,\ell} = \mathrm{Clip}\left(V^{(\ell)} U_{t,\ell-1} \mathrm{Attn}^{(\ell)}(U_{t,\ell-1})\right),$$

where Clip is applied element-wise across the columns of the matrix input and bounds each column's L_2 norm by 1. This plays a similar role to layer normalization (Ba et al., 2016) as it applies normalization for a layer across the *feature* dimension as opposed to the batch dimension. The matrix $V^{(\ell)}$ resembles the value matrix and without loss of generality, we assume that the elements of $V^{(\ell)}$ are distributed iid $\mathcal{N}(0, 1/r)$. For $\ell < L$, $V^{(\ell)} \in \Re^{r \times r}$ and $V^{(L)} \in \Re^{d \times r}$.

Finally, the next token is generated via sampling from the softmax of the final layer:

$$X_{t+1} \sim \sigma\left(U_{t,L}[-1]\right),$$

where $U_{t,L}[-1]$ denotes the right-most column of $U_{t,L}$. At each layer ℓ, the parameters $\theta^{(\ell)}$ consist of the matrices $A^{(\ell)}, V^{(\ell)}$. We will use the notation $\theta^{(i:j)}$ for $i \leq j$ to denote the collection $(\theta^{(i)}, \theta^{(i+1)}, \ldots, \theta^{(j)})$. While we assume Gaussianity of the weights for concreteness, the proof technique follows for any weight distribution with bounded expected squared operator norm. We now present an upper

bound on the estimation error of the Bayes optimal predictor.

Theorem 3.5. (estimation error bound) *For all* d, r, L, K, T, *if for all* t, X_t *is generated by the transformer environment, then*

$$\mathcal{L}_T \leq \frac{pL \ln\left(136eK^2\right) + p\ln\left(\frac{2KT^2}{L}\right)}{T},$$

where $p = 2r^2(L-1) + (dr + r^2)$ *denotes the parameter count of the transformer.*

We note that even if the sequence generated by the transformer is not iid, $\mathcal{L}_T$ decays linearly in T, the length of the sequence. Furthermore, we observe that $\mathcal{L}_T$ is upper bounded by a multiple of the product of parameter count and depth of the transformer model as in Bai et al. (2023). In the following section, we will draw the connection to ICL by studying meta-learning in a data generating process which resembles a sparse *mixture* of deep transformers.

4. Meta-Learning from Sequential Data

In this section, we analyze the achievable performance of *meta-learning* from sequences. The tools of the Bayesian framework apply exactly as they do in standard supervised learning from sequences. An example of meta-learning from sequences includes language model pretraining in which each "meta-task" can be interpreted as a separate document and the "sequence" as the tokens which comprise the document. We will use the terminology *document* going forward to refer to a "meta-task" in meta-learning.

4.1. Data Generating Process

We now consider sequential data which resembles a *corpus* of text documents. We assume that all documents in the corpus have an identical length denoted by T. For each document m, we let $D_m = (X_1^{(m)}, \ldots, X_T^{(m)})$ be the sequence of discrete random variables which resembles its constituent tokens. In meta-learning, we assume that the sequence $(D_1, D_2, \ldots)$ is *exchangable*. By De Finetti's theorem there exists a latent random variable ψ such that conditioned on ψ, $(D_1, D_2, \ldots)$ are iid. Note that ψ encodes information which is learnable *across* documents in a corpus. As such, ψ represents the *meta* parameters.

However, we further assume that each document additionally consists of *intra*-task parameters θ_m which encode information that is specific to document m. As in the previous section, we assume that the sequence D_m is produced by an autoregressive process. As such, for all t, the value of $X_{t+1}^{(m)}$ depends on θ_m and the prior tokens $(X_1^{(m)}, \ldots, X_t^{(m)})$ in D_m. We assume that $(\theta_1, \theta_2, \ldots)$ is also iid conditioned on ψ and that $D_m \perp \psi | \theta_m$. As a consequence, for all $m \neq n$,

$X_t^{(m)} \perp X_t^{(n)} | \psi$. Tokens *across* documents do not convey information beyond what is contained in ψ.

4.2. Bayesian Error

Our framework is *Bayesian* in the sense that it treats learning as the process of reducing uncertainty about $\theta_1, \ldots, \theta_m, \psi$, which are taken to be random variables. For a meta-learning problem with M documents each of length T, a learning algorithm produces, for each $(m,t) \in [M] \times [T]$, a *predictive distribution* $P_{m,t}$ of $X_{t+1}^{(m)}$ after observing the concatenated history which we denote by

$$H_{m,t} = (D_1, D_2, \ldots, D_{m-1}, X_1^{(m)}, \ldots, X_t^{(m)}).$$

$H_{m,t}$ consists of *all* tokens from documents $1, \ldots, m-1$ and up to the tth token of document m. We express our meta-learning algorithm in terms of a function π for which $P_{m,t} = \pi(H_{m,t})$. For all $M, T \in \mathbb{Z}_{++}$, it is common to measure the error realized by predictions $P_{m,t}$ for $(m,t) \in [M] \times [T]$ in terms of the average cumulative expected log-loss:

$$\mathbb{L}_{M,T,\pi} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}_\pi\left[-\ln P_{m,t}\left(X_{t+1}^{(m)}\right)\right].$$

We note that this objective largely resembles the objective LLMs minimize in the process of pre-training. However, again since this error is not baselined at 0, we instead study the following estimation error:

$$\mathcal{L}_{M,T,\pi} = \mathbb{L}_{M,T,\pi} - \mathbb{L}_{M,T}^*,$$

where $\mathbb{L}_{M,T}^* = \mathbb{H}(H_{M,T}|\psi, \theta_{1:M})/(MT)$ is irreducible error.

4.3. Achievable Bayesian Error

We are in particular interested in the algorithm π which minimizes Bayesian error. Just as in supervised learning from sequences, across all problem instances, the optimal algorithm π sets $P_{m,t} = \mathbb{P}(X_{t+1}^{(m)} \in \cdot | H_{m,t})$ for all m, t. We denote this *Bayesian posterior* by $\hat{P}_{m,t}$.

Lemma 4.1. (Bayesian posterior is optimal) *For all* $m, t \in \mathbb{Z}_+$,

$$\mathbb{E}\left[-\ln \hat{P}_{m,t}\left(X_{t+1}^{(m)}\right)\right] = \min_\pi \mathbb{E}_\pi\left[-\ln P_{m,t}\left(X_{t+1}^{(m)}\right)\right].$$

We use $\mathcal{L}_{M,T}$ to denote the *optimal* achievable Bayesian error:

$$\mathcal{L}_{M,T} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[-\ln \hat{P}_{m,t}\left(X_{t+1}^{(m)}\right)\right] - \mathbb{L}_{M,T}^*.$$

We will restrict our attention to the performance of the optimal predictor $\hat{P}_t$. We now present the main result of this paper which decomposes optimal Bayesian error into 3 intuitive terms. The following result provides an exact characterization of $\mathcal{L}_{M,T}$.

Theorem 4.2. (Main Result) *For all* $M, T \in \mathbb{Z}_+$ *and* $m \in \{1, 2, \ldots, M\}$,

$$\mathcal{L}_{M,T} = \underbrace{\frac{\mathbb{I}(H_{M,T}; \psi)}{MT}}_{\substack{meta \\ estimation \\ error}} + \underbrace{\frac{\mathbb{I}(D_m; \theta_m | \psi)}{T}}_{\substack{intra\text{-}document \\ estimation \\ error}}.$$

Proof.

$$\mathcal{L}_{M,T}$$
$$= \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[-\ln \hat{P}_{m,t}(X_{t+1}^{(m)})\right] - \mathbb{L}_*$$
$$\stackrel{(a)}{=} \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}^{(m)}; \psi, \theta_m | H_{m,t})$$
$$\stackrel{(b)}{=} \frac{1}{MT} \sum_{m=1}^{M} \mathbb{I}(D_m; \psi, \theta_m | H_{m-1,T})$$
$$\stackrel{(c)}{=} \frac{1}{MT} \sum_{m=1}^{M} \mathbb{I}(D_m; \psi | H_{m-1,T})$$
$$\quad + \frac{1}{MT} \sum_{m=1}^{M} \mathbb{I}(D_m; \theta_m | \psi, H_{m-1,T})$$
$$\stackrel{(d)}{=} \frac{\mathbb{I}(H_{M,T}; \psi)}{MT} + \frac{\mathbb{I}(D_m; \theta_m | \psi)}{T},$$

where (a) follows from Theorem 3.2, and $(b), (c), (d)$ follow from the chain rule of mutual information. $\square$

The meta-estimation error represents the statistical error incurred in the process of estimating the meta parameters ψ. Since all tokens across all documents contain information about ψ, it is intuitive that meta-estimation error term decays linearly in MT. Since M could in practice be very large (for example in a pretraining dataset), $\mathbb{L}_{M,T}$ could be small even for small T if significant learning complexity is contained in ψ.

Meanwhile, the intra-document estimation error represents the statistical error incurred in the process of learning θ_m after already conditioning on ψ. As only the data from document m (D_m) pertains to θ_m, this error intuitively decays linearly in T, the length of the document. As mentioned before, if much of the learning complexity is contained in ψ, then $\mathbb{I}(D_m; \theta_m | \psi)$ will be small and therefore the intra-document estimation error may be small even for short document length T. We will revisit this idea in section 4.5 when we analyze ICL within this framework.

We note that Theorem 4.2 holds for all data generating processes which meet the natural assumptions made in subsection 4.1. It is surprising that we can arrive at such a result which decays linearly in both M, the number of documents, and T, the lengths of the documents without any explicit reliance on stability or mixing assumptions.

While the main result is useful for conceptual understanding, we need further tools to facilitate the theoretical analysis of concrete meta-learning problem instances. To extend this result, we again use rate-distortion theory under the following modified rate-distortion functions:

$$\mathbb{H}_{\epsilon,T}(\theta_m | \psi) = \inf_{\tilde{\theta}_m \in \tilde{\Theta}_{\epsilon,T}} \mathbb{I}(\theta_m; \tilde{\theta}_m | \psi),$$

where

$$\tilde{\Theta}_{\epsilon,T} = \left\{\tilde{\theta} : \tilde{\theta} \perp H_{M,T} | \theta_m; \quad \frac{\mathbb{I}(D_m; \theta_m | \tilde{\theta}, \psi)}{T} \leq \epsilon\right\},$$

and

$$\mathbb{H}_{\epsilon,M,T}(\psi) = \inf_{\tilde{\psi} \in \tilde{\Psi}_{\epsilon,M,T}} \mathbb{I}(\psi; \tilde{\psi}),$$

where

$$\tilde{\Psi}_{\epsilon,M,T} = \left\{\tilde{\psi} : \tilde{\psi} \perp H_{M,T} | \psi; \quad \frac{\mathbb{I}(H_{M,T}; \psi | \tilde{\psi})}{MT} \leq \epsilon\right\}.$$

With this notation in place, we establish the following upper and lower bounds on $\mathcal{L}_{M,T}$ in terms of the above rate distortion functions.

Theorem 4.3. (rate-distortion estimation error bound) *For all* $M, T \in \mathbb{Z}_+$, *and* $m \in \{1, \ldots, M\}$,

$$\mathcal{L}_{M,T} \leq \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{\epsilon,M,T}(\psi)}{MT} + \epsilon + \inf_{\epsilon' \geq 0} \frac{\mathbb{H}_{\epsilon',T}(\theta_m | \psi)}{T} + \epsilon',$$

and

$$\mathcal{L}_{M,T} \geq \sup_{\epsilon \geq 0} \min\left\{\frac{\mathbb{H}_{\epsilon,M,T}(\psi)}{MT}, \epsilon\right\}$$
$$+ \sup_{\epsilon' \geq 0} \min\left\{\frac{\mathbb{H}_{\epsilon',T}(\theta_m | \psi)}{T}, \epsilon'\right\}.$$

A direct consequence of Theorem 4.3 is an upper bound on Bayesan error with respect to *entropy* (by setting ϵ, ϵ' to 0). While the utility of such a bound is limited to settings in which $\psi, \theta_{1:M}$ are discrete random variables, it may be useful to the reader conceptually. The bound is captured in the following Corollary:

Corollary 4.4. (entropy estimation error bound) *For all* $M, T \in \mathbb{Z}_+$, *and* $m \in \{1, \ldots, M\}$

$$\mathcal{L}_{M,T} \leq \frac{\mathbb{H}(\psi)}{MT} + \frac{\mathbb{H}(\theta_m | \psi)}{T}.$$

In the following section, we will apply Theorem 4.3 to derive error bounds for a sparse mixture of (deep) transformers. For a simpler linear representation learning example, we refer the reader to Appendix B.1.

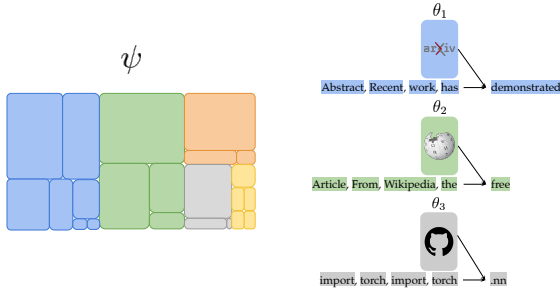### 4.4. Sparse Mixture of Transformers

Figure 1. The above diagram depicts the sparse mixture of transformers problem. ψ designates a mixture of many transformer models, each which is responsible for producing documents of a particular type. Each document is generated by sampling a transformer model θ_m from the mixture ψ, and generating a sequence $(X_1^{(m)}, X_2^{(m)}, X_3^{(m)}, \ldots)$ autoregressively.

In the sparse mixture of transformers environment, for all documents m, we let its tokens $(X_1^{(m)}, X_2^{(m)}, \ldots)$ be a sequence in $\{1, \ldots, d\}$, where d denotes the size of the vocabulary. Each of the d outcomes is associated with a *known* embedding vector which we denote as Φ_j for $j \in \{1, \ldots, d\}$. We assume that for all j, $\|\Phi_j\|_2 = 1$. For brevity of notation, we let $\phi_t^{(m)} = \Phi_{X_t^{(m)}}$ i.e. the embedding associated with token $X_t^{(m)}$.

Each document is generated by a transformer model which is sampled iid from a mixture. We assume that sampling is performed according to a categorical distribution parameterized by α with prior distribution $\mathbb{P}(\alpha \in \cdot) = \text{Dirichlet}(N, [R/N, \ldots, R/N])$ for a scale parameter $R \ll N$. Under this prior distribution, the expected number of unique outcomes grows linearly in R and only logarithmically in the number of draws (M in our case). As a result, we permit the size of the mixture N to potentially be exponentially large, but we assume that the mixture's complexity is controlled by the sparsity parameter R.

Each of the N elements of the mixture corresponds to a deep transformer network as outlined in Section 3.3. Let K denote the context lengths of the transformers, L denote their depths, and r their attention dimensions. Let $\psi_1, \ldots, \psi_N$ denote the weights of the transformer mixture elements. As a result, the meta-parameters ψ consist of $(\alpha, \psi_1, \ldots, \psi_N)$ i.e. the mixture sampling probabilities and the mixture model weights.

We assume that for all documents, the first token $X_1^{(m)}$ is sampled from an arbitrary pmf on $\{1, \ldots, d\}$ but subsequent tokens are sampled based on the previous K tokens within the context window and the weights of the sampled transformer model. The tokens of each document are generated according to the weights of the sampled transformer and the previous K tokens. The generation of token $X_{t+1}^{(m)}$ will depend on θ_m and $X_{t-K+1}^{(m)}, \ldots, X_t^{(m)}$. For all m, t, we let $(U_{t,0}^{(m)} = \phi_{t-K+1:t}^{(m)})$ refer to the embeddings associated with the past K tokens. For $\ell > 0$, we let $U_{t,\ell}^{(m)}$ denote the output of layer ℓ of the transformer with input $U_{t,0}^{(m)}$. For all $t \leq T, \ell < L, m \leq M$, let

$$\text{Attn}^{(\ell)}(U_{t,\ell-1}^{(m)}) = \sigma\left(\frac{U_{t,\ell-1}^{(m)\top} A_\ell^{(m)} U_{t,\ell-1}^{(m)}}{\sqrt{r}}\right)$$

denote the attention matrix of layer ℓ for document m where σ denotes the softmax function applied elementwise along the columns. The matrix $A_\ell^{(m)} \in \Re^{r \times r}$ can be interpreted as the product of the key and query matrices and without loss of generality, we assume that the elements of the matrices $A_\ell^{(m)}$ are distributed iid $\mathcal{N}(0, 1)$.

Subsequently, we let

$$U_{t,\ell}^{(m)} = \text{Clip}\left(V_\ell^{(m)} U_{t,\ell-1}^{(m)} \text{Attn}^{(\ell)}(U_{t,\ell-1}^{(m)})\right),$$

where Clip is applied elementwise on the columns of the matrix input and ensures that all columns have L_2 norm at most 1. The matrix $V_\ell^{(m)}$ resembles the value matrix and without loss of generality, we assume that the elements of $V_\ell^{(m)}$ are distributed iid $\mathcal{N}(0, 1/r)$.

Finally, the next token is generated via sampling from the softmax of the final layer:

$$X_{t+1}^{(m)} \sim \sigma\left(U_{t,L}^{(m)}[-1]\right),$$

where $U_{t,L}^{(m)}[-1]$ denotes the right-most column of $U_{t,L}^{(m)}$. At each layer ℓ, the parameters $\theta_{m,\ell}$ consist of the matrices $A_\ell^{(m)}, V_\ell^{(m)}$. While we assume Gaussianity of the weights for concreteness, the proof technique follows for any weight distribution with bounded expected squared operator norm.

We provide the following novel result which upper bounds the error of the optimal Bayesian learner when learning from data generated by the sparse mixture of transformers.

Theorem 4.5. (estimation error bound) *For all* $d, r, K, L, M, N, R, T \in \mathbb{Z}_{++}$, *if for all* $(m, t) \in [M] \times [T]$, $X_t^{(m)}$ *is generated according to the sparse mixture of*

transformers environment, then

$$\mathcal{L}_{M,T} \leq \frac{pRL \ln\left(1 + \frac{M}{R}\right) \ln(136eK^2)}{MT}$$
$$+ \frac{pR \ln\left(1 + \frac{M}{R}\right) \ln\left(\frac{2KMT^2}{L}\right)}{MT}$$
$$+ \frac{\ln(N)}{T},$$

where $p = 2r^2(L-1) + (dr + r^2)$ denotes the parameter count of each transformer in the mixture.

We now provide some qualitative comments about this result. The first and second terms denote the meta estimation error, and the third term denotes the intra-document estimation error.

The first two terms capture the error incurred in the process of learning ψ, the probabilities and model weights which comprise of the mixture. Note that even if there are N models in the mixture, due to the Dirichlet prior, the error depends linearly on the scale parameter R and only logarithmically on M. This is because even if a model is resampled for every document, several documents may still be generated by the *same* model from the mixture. As a result, the dependence is linear in R and only logarithmic in M. The remaining terms are linear in the product of parameter count and depth, which corroborates the results of Bai et al. (2023). However, our result decays linearly in MT as opposed to just M as in (Bai et al., 2023). Intuitively, the error from the meta-parameters ought to decrease in both the number of documents M and the *length* of the documents T. This is an advantage of the Bayesian framework as it does not rely on a uniform convergence argument which requires mixing time assumption on the tokens within the document to obtain decay in T.

Finally, the third term is the intra-document estimation error which is the error incurred in the process of learning which model from the mixture generated each document. Since there are N different elements in the mixture, the $\log(N)/T$ is straightforward. The longer the document length T, the more certain we should be about which model generated the document, hence lower error. In the following section, we explicitly outline the connection between this example and ICL.

4.5. In-context Learning as Meta-Learning from Sequences

We now explicitly draw the connection between ICL and meta-learning from sequences. We assume that the pretraining dataset consists of M documents, each of length T. We assume that a new $(M+1)$th document type is drawn and an in-context learner is described by an algorithm which produces for each t a predictive distribution P_t^{in} of $X_{t+1}^{(M+1)}$

after observing the history $H_{M+1,t}$ which consists of the pretraining data an the t provided in the current context. We let $D_{M+1} = (X_1^{(M+1)}, \ldots, X_\tau^{(M+1)})$ denote the entire in-context sequence. Note that we have summarize the effect of pretraining by allowing $P_t^{(in)}$ to depend on the pretraining history $H_{M,T}$. We quantify error realized by predictions P_t^{in} in terms of the average cumulative expected log-loss:

$$\mathbb{L}_{M,T,\tau,\pi} = \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}_\pi \left[-\log P_t^{in}\left(X_{t+1}^{(M+1)}\right) \right],$$

where τ denotes the full length of the in-context sequence. We assume that $\tau \leq T$ as τ can be at most K, the context-length of the transformer and the document lengths T in pretraining are often much larger than K. We appropriately baseline the log-loss with the following:

$$\mathcal{L}_{M,T,\tau,\pi} = \mathbb{L}_{M,T,\tau,\pi} - \mathbb{L}_{M,T,\tau}^*,$$

where $\mathbb{L}_{M,T,\tau}^* = \mathbb{H}(D_{M+1}|\theta_{M+1})/\tau$ is the irreducible error.

As before, we establish that $\hat{P}_t(X_{t+1}^{(M+1)} \in \cdot) = \mathbb{P}(X_{t+1}^{(M+1)} \in \cdot | H_{M+1,t})$ minimizes the loss.

Lemma 4.6. *For all $M, T, t \in \mathbb{Z}_+$,*

$$\mathbb{E}\left[-\log \hat{P}_t(X_{t+1}^{(M+1)}) \right] = \min_\pi \mathbb{E}_\pi \left[\log P_t^{in}(X_{t+1}^{(M+1)}) \right].$$

Going forward, we will restrict our attention to the performance of $\hat{P}_t$ which we denote as:

$$\mathcal{L}_{M,T,\tau} = \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}\left[-\log \hat{P}_t(X_{t+1}^{(M+1)}) \right] - \mathbb{L}_{M,T,\tau}^*.$$

With this notation in place, we present an upper bound for the ICL error. A proof can be found in Appendix B.3.

Theorem 4.7. (in context learning error bound) *For all $M, T, \tau \in \mathbb{Z}_{++}$, if $\tau \leq T$, then*

$$\mathcal{L}_{M,T,\tau} \leq \underbrace{\frac{\mathbb{I}(H_{M,T}; \psi)}{M\tau}}_{\substack{meta \\ estimation \\ error}} + \underbrace{\frac{\mathbb{I}(D_{M+1}; \theta_{M+1}|\psi)}{\tau}}_{\substack{in\text{-}context \\ estimation \\ error}}.$$

Note that if M is large i.e. the number of pretraining documents is large, then almost all of the error will be attributed to the in-context estimation error:

Remark 4.8. For sufficiently large M (number of pretraining documents),

$$\mathcal{L}_{M,T,\tau} \lesssim \frac{\mathbb{I}(D_{M+1}; \theta_{M+1}|\psi)}{\tau}.$$

4.6. Discussion of Results

If each pretraining document is generated by a transformer model which is drawn from a mixture as in the previous section, the above remark suggests for a sufficiently large pretraining set, the in-context error can be small for even modest values of τ. The in-context error is upper bounded by $\log(N)/\tau$ where N is the size of the mixture. Effectively, the in-context data only needs to distinguish which model from the mixture generated the current sequence. As a result, the complexity is at most $\log(N)$ and the error decays linearly in the length of the in-context sequence τ. This corroborates work by Min et al. (2022) which established that an in-context sequence largely augments performance via providing information about the distributions of the inputs and labels as well as the format of the sequence. The LLMs is not literally learning from the examples, as even when the labels of examples were randomly scrambled, performance on downstream tasks was only marginally impacted. This lends credence to the hypothesis that ICL pinpoints which model from the mixture is most suitable for the given in-context sequence.

Much of the existing literature in this space is unable to obtain an error bound which decays in both M, the number of documents, and T, the lengths of the documents (Bai et al., 2023; Edelman et al., 2021). Li et al. (2023a) are able to establish such a result in the frequentist setting after assuming stability of the in-context sequence and stability of the ERM algorithm, albeit with exponential dependence on the depth of the transformer model. Zhang et al. (2023b) were also able to establish a result which decays in both M and T without such stability assumptions but instead with a dependence on the mixing time of the Markov chain induced by their data generating process. Our results are unique in that they are able to establish an error bound which diminishes in both M and T without reliance on contrived assumptions on the data generating process.

5. Conclusion

In this work, we introduced novel information-theoretic tools to analyze the error of meta-learning from sequences. Our tools produced very general and intuitive results which suggest that the error should decay in both the number of training sequences and the sequence lengths. Notably, these results hold without relying on contrived mixing time assumptions as common in existing work. By applying these tools, we developed novel results about ICL in transformers and a plausible mathematical hypothesis for how learning is possible even when only a small amount of data is provided in-context. While the results of the main text are limited to exact Bayesian inference, we provide results in the Appendix which extend to *suboptimal* algorithms as well. A further rigorous investigation into the mechanisms

by which transformers may be implementing a mixture of models would provide stronger credence to the hypothesis and results provided in this work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. While there are many potential societal consequences in working towards that goal, we don't strongly feel that a specific concern ought to be highlighted here.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Aliakbarpour, M., Bairaktari, K., Brown, G., Smith, A., and Ullman, J. Metalearning with very few samples per task. *arXiv preprint arXiv:2312.13978*, 2023.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. Transformers as statisticians: Provable in-context learning with in-context algorithm selection, 2023.

Baxter, J. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Dai, D., Sun, Y., Dong, L., Hao, Y., Sui, Z., and Wei, F. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, M., Orseau, L., Hutter, M., and Veness, J. Language modeling is compression, 2024.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. Incremental learning-to-learn with statistical guarantees. *arXiv preprint arXiv:1803.08089*, 2018.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Edelman, B. L., Goel, S., Kakade, S. M., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. *CoRR*, abs/2110.10090, 2021. URL https://arxiv.org/abs/2110.10090.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 2021.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. In *International Conference on Machine Learning*, pp. 1920–1930. PMLR, 2019.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Giannou, A., Rajput, S., Sohn, J.-y., Lee, K., Lee, J. D., and Papailiopoulos, D. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.

Huang, Y., Cheng, Y., and Liang, Y. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

Jelassi, S., Sander, M., and Li, Y. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Jeon, H. J., Zhu, Y., and Van Roy, B. An information-theoretic framework for supervised learning, 2023.

Kirsch, L., Harrison, J., Sohl-Dickstein, J., and Metz, L. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, 2022.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/li23l.html.

Li, Y., Li, Y., and Risteski, A. How do transformers learn topic structure: Towards a mechanistic understanding. *arXiv preprint arXiv:2303.04245*, 2023b.

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., and Zhang, C. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Mahankali, A., Hashimoto, T. B., and Ma, T. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

Maurer, A. Transfer bounds for linear feature learning. *Machine learning*, 75(3):327–350, 2009.

Maurer, A., Pontil, M., and Romera-Paredes, B. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work?, 2022.

Pontil, M. and Maurer, A. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pp. 55–76. PMLR, 2013.

Raventos, A., Paul, M., Chen, F., and Ganguli, S. The effects of pretraining task diversity on in-context learning of ridge regression. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Rubin, O., Herzig, J., and Berant, J. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.

Sanford, C., Hsu, D., and Telgarsky, M. Representational strengths and limitations of transformers. *arXiv preprint arXiv:2306.02896*, 2023.

Song, J. and Zhong, Y. Uncovering hidden geometry in transformers via disentangling position and context. *arXiv preprint arXiv:2310.04861*, 2023.

Srebro, N. and Ben-David, S. Learning bounds for support vector machines with learned kernels. In *International Conference on Computational Learning Theory*, pp. 169–183. Springer, 2006.

Tarzanagh, D. A., Li, Y., Thrampoulidis, C., and Oymak, S. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

Tian, Y., Wang, Y., Chen, B., and Du, S. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arXiv:2305.16380*, 2023a.

Tian, Y., Wang, Y., Zhang, Z., Chen, B., and Du, S. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*, 2023b.

Tripuraneni, N., Jin, C., and Jordan, M. Provable meta-learning of linear representations. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10434–10443. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tripuraneni21a.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference, 2022.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023a.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization, 2023b.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

A. Learning from Sequential Data

In the following, we will provide a result which requires a change of measure. For all random variables $X : \Omega \mapsto \mathcal{X}, Y : \Omega \mapsto \mathcal{Y}$ and realizations $y \in \mathcal{Y}$, one may consider the distribution $\mathbb{P}(X \in \cdot | Y = y)$. Let function $f(y) = \mathbb{P}(X \in \cdot | Y = y)$. Then, for any random variable $Z : \Omega \mapsto \mathcal{Z}$ for which $\mathcal{Z} \subseteq \mathcal{Y}$, we use $\mathbb{P}(X \in \cdot | Y \leftarrow Z)$ to denote $f(Z)$.

Lemma A.1. *If $\tilde{\theta} : \Omega \mapsto \tilde{\Theta}$ is a random variable such that for all $t \in \mathbb{Z}_+$, $Y_{t+1} \perp \tilde{\theta} | \theta, H_t$ and $\tilde{\Theta} \subseteq \Theta$, then*

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y_{t+1} \in \cdot | \theta, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | \tilde{\theta}, H_t)) \right] \leq \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y_{t+1} \in \cdot | \theta, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | \theta \leftarrow \tilde{\theta}, H_t)) \right].$$

Proof.

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y_{t+1} \in \cdot | \theta, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | \theta \leftarrow \tilde{\theta}, H_t)) \right]$$

$$= \mathbb{E}\left[\ln \frac{\mathbb{P}(Y_{t+1} | \theta, H_t)}{\mathbb{P}(Y_{t+1} | \theta \leftarrow \tilde{\theta}, H_t)} \right]$$

$$= \mathbb{E}\left[\ln \frac{\mathbb{P}(Y_{t+1} | \theta, H_t)}{\mathbb{P}(Y_{t+1} | \tilde{\theta}, H_t)} \right] + \mathbb{E}\left[\ln \frac{\mathbb{P}(Y_{t+1} | \tilde{\theta}, H_t)}{\mathbb{P}(Y_{t+1} | \theta \leftarrow \tilde{\theta}, H_t)} \right]$$

$$= \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y_{t+1} \in \cdot | \theta, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | \tilde{\theta}, H_t)) \right] + \mathbb{E}\left[\mathbb{E}\left[\ln \frac{\mathbb{P}(Y_{t+1} | \tilde{\theta}, H_t)}{\mathbb{P}(Y_{t+1} | \theta \leftarrow \tilde{\theta}, H_t)} \Big| \tilde{\theta}, H_t \right] \right]$$

$$= \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y_{t+1} \in \cdot | \theta, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | \tilde{\theta}, H_t)) \right] + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(Y_{t+1} \in \cdot | \tilde{\theta}, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | \theta \leftarrow \tilde{\theta}, H_t) \right) \right]$$

$$\overset{(a)}{\geq} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y_{t+1} \in \cdot | \theta, H_t) \| \mathbb{P}(Y_{t+1} \in \cdot | \tilde{\theta}, H_t)) \right],$$

where (a) follows from the fact that KL-divergence is non-negative. $\qquad\square$

Lemma 3.1. (Bayesian posterior is optimal) *For all $t \in \mathbb{Z}_+$,*

$$\mathbb{E}\left[-\ln \hat{P}_t(X_{t+1}) \right] = \min_{\pi} \mathbb{E}_{\pi}\left[-\ln P_t(X_{t+1}) \right].$$

Proof. In the below proof take all equality to hold *almost surely*.

$$\mathbb{E}\left[-\ln P_t(X_{t+1}) | H_t \right]$$

$$= \mathbb{E}\left[-\ln \hat{P}_t(X_{t+1}) + \ln \frac{\hat{P}_t(X_{t+1})}{P_t(X_{t+1})} \Big| H_t \right]$$

$$= \mathbb{E}\left[-\ln \hat{P}_t(X_{t+1}) \Big| H_t \right] + \mathbf{d}_{\mathrm{KL}}(\hat{P}_t \| P_t).$$

The result follows since KL divergence is non-negative. $\qquad\square$

Theorem 3.2. (Bayesian estimation error) *For all $T \in \mathbb{Z}_+$,*

$$\mathcal{L}_T = \frac{\mathbb{I}(H_T; \theta)}{T}.$$

Proof.

$$\mathbb{L}_T = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[-\ln \hat{P}_j(X_{t+1})\right]$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[-\ln \mathbb{P}(X_{t+1}|H_t, \theta) + \ln \frac{\mathbb{P}(X_{t+1}|H_t, \theta)}{\hat{P}_t(X_{t+1})}\right]$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{H}(X_{t+1}|\theta, H_t) + \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(X_{t+1} \in \cdot | H_t, \theta) \| \hat{P}_t(X_{t+1} \in \cdot))\right]$$

$$\stackrel{(a)}{=} \frac{\mathbb{H}(H_T|\theta)}{T} + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta|H_t)$$

$$\stackrel{(b)}{=} \frac{\mathbb{H}(H_T|\theta)}{T} + \frac{\mathbb{I}(H_T; \theta)}{T},$$

where (a) and (b) follow from the chain rule of conditional mutual information. $\square$

Theorem 3.4. **(rate-distortion estimation error bound)** *For all $T \in \mathbb{Z}_+$,*

$$\sup_{\epsilon \geq 0} \min\left\{\frac{\mathbb{H}_{\epsilon,T}(\theta)}{T}, \epsilon\right\} \leq \mathcal{L}_T \leq \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{\epsilon,T}(\theta)}{T} + \epsilon.$$

Proof.

$$\mathcal{L}_T = \frac{\mathbb{I}(H_T; \theta)}{T}$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta|H_t)$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta, \tilde{\theta}|H_t)$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \tilde{\theta}|H_t) + \mathbb{I}(X_{t+1}; \theta|\tilde{\theta}, H_t)$$

$$= \frac{\mathbb{I}(H_T; \tilde{\theta})}{T} + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta|\tilde{\theta}, H_t)$$

$$\leq \inf_{\tilde{\theta}} \frac{\mathbb{I}(\theta; \tilde{\theta})}{T} + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta|\tilde{\theta}, H_t)$$

$$\leq \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{\epsilon,T}(\theta)}{T} + \epsilon$$

Suppose that $\mathbb{I}(H_T; \theta) < \mathbb{H}_{\epsilon,T}$. Let $\tilde{\theta} = \tilde{H}_T \notin \tilde{\Theta}_{\epsilon,T}$ where $\tilde{H}_T$ is another history sampled in the same manner as H_T.

$$
\begin{aligned}
\mathbb{I}(H_T; \theta) &= \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta | H_t) \\
&\stackrel{(a)}{\geq} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta | \tilde{H}_t, H_t) \\
&= \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta | \tilde{\theta}, H_t) \\
&\stackrel{(b)}{\geq} \epsilon T,
\end{aligned}
$$

where (a) follows from the fact that conditioning reduces entropy and that $X_{t+1} \perp \tilde{H}_t | (\theta, H_t)$ and (b) follows from the fact that $\tilde{\theta} \notin \tilde{\Theta}_{\epsilon,T}$. Therefore, for all $\epsilon \geq 0$, $\mathbb{I}(H_T; \theta) \geq \min\{H_{\epsilon,T}, \epsilon T\}$. The result follows. $\qquad \square$

A.1. Logistic Regression

We introduce a simple logistic regression problem as a concrete instance to demonstrate an application of the general aforementioned results. We assume that $X_0 = \bar{X}_0$ and $X_t = (Y_t, \bar{X}_t)$ for all $t \geq 1$. The "inputs" $(\bar{X}_0, \ldots, \bar{X}_T)$ are generated according to an iid random process for which $X_j \sim \mathcal{N}(0, I_d)$. Meanwhile, we assume that Y_{t+1} is generated by the following process:

$$
Y_{t+1} = \begin{cases} 1 & \text{w.p. } \frac{1}{1+e^{-\theta^\top X_t}} \\ -1 & \text{otherwise} \end{cases},
$$

where θ denotes the parameters of the logistic model and we assume the prior distribution $\mathbb{P}(\theta \in \cdot) = \mathrm{Unif}(\{\nu \in \Re^d : \|\nu\|_2 \leq 1\})$.

In this environment, θ is the only unknown quantity and as such, the distributions of all random variables are *known* to the algorithm designer. In this example, the sequence is iid once conditioned on θ. We begin with this example for simplicity and to demonstrate that our analytical tools are general enough to subsume the analysis of supervised learning from iid data.

Theorem A.2. (logistic regression Bayesian error bounds) *For all $d, T \in \mathbb{Z}_{++}$, if θ, H_T follow the logistic regression environment, then*

$$
\mathcal{L}_T \leq \frac{d}{2T}\left(1 + \ln\left(1 + \frac{T}{4d}\right)\right).
$$

Proof. From Theorem 3.4, it suffices to upper bound the rate-distortion function. Let $\tilde{\theta} = \theta + Z$ where $Z \perp \theta$ and $Z \sim \mathcal{N}(0, 8\epsilon/d)$. Then,

$$
\begin{aligned}
&\mathbb{I}(Y; \theta | \tilde{\theta}, X) \\
&= \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot | \theta, X) \| \mathbb{P}(Y \in \cdot | \tilde{\theta}, X))\right] \\
&\stackrel{(a)}{\leq} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot | \theta, X) \| \mathbb{P}(Y \in \cdot | \theta \leftarrow \tilde{\theta}, X))\right] \\
&= \mathbb{E}\left[\frac{\ln\left(\frac{e^{-\tilde{\theta}^\top X}}{e^{-\theta^\top X}}\right)}{1 + e^{-\theta^\top X}} + \frac{\ln\left(\frac{e^{\tilde{\theta}^\top X}}{e^{\theta^\top X}}\right)}{1 + e^{\theta^\top X}}\right] \\
&\stackrel{(b)}{\leq} \frac{\mathbb{E}\left[\left(\theta^\top X - \tilde{\theta}^\top X\right)^2\right]}{8} \\
&= \frac{\mathbb{E}\left[\|\theta - \tilde{\theta}\|_2^2\right]}{8} \\
&= \epsilon,
\end{aligned}
$$

14

where (a) follows from Lemma 3.1 and (b) follows from the fact that for all $x, y \in \Re$,

$$\frac{\ln\left(\frac{1+e^{-y}}{1+e^{-x}}\right)}{1+e^{-x}} + \frac{\ln\left(\frac{1+e^{y}}{1+e^{x}}\right)}{1+e^{x}} \leq \frac{(x-y)^2}{8}.$$

Therefore, $\theta \in \Theta_\epsilon$ so it suffices to upper bound the rate $\mathbb{I}(\theta; \tilde{\theta})$.

$$\begin{aligned}
\mathbb{I}(\theta; \tilde{\theta}) &= \mathbf{h}(\tilde{\theta}) - \mathbf{h}(\tilde{\theta}|\theta) \\
&= \mathbf{h}(\tilde{\theta}) - \mathbf{h}(Z|\theta) \\
&= \mathbf{h}(\tilde{\theta}) - \mathbf{h}(Z) \\
&\leq \frac{d}{2}\ln\left(2\pi e\left(\frac{1+8\epsilon}{d}\right)\right) - \frac{d}{2}\ln\left(2\pi e\frac{8\epsilon}{d}\right) \\
&= \frac{d}{2}\ln\left(1 + \frac{1}{8\epsilon}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{L}_T &\overset{(a)}{\leq} \inf_{\epsilon \geq 0}\left(\frac{d}{2T}\ln\left(1 + \frac{1}{8\epsilon}\right) + \epsilon\right) \\
&\overset{(b)}{\leq} \frac{d}{2T}\ln\left(1 + \frac{T}{4d}\right) + \frac{d}{2T},
\end{aligned}$$

where (a) follows from Theorem 3.4 and (b) follows by setting $\epsilon = d/(2T)$. $\qquad\square$

As one would expect, the above result establishes that the Bayesian error of an optimal learning algorithm is $\mathcal{O}(\frac{d}{T}\log\frac{T}{d})$. The proof illustrates a common technique for bounding the rate-distortion function i.e. considering a compression $\tilde{\theta} = \theta + Z$ where Z is independent zero-mean Gaussian noise with tunable variance. In the following section, we use the same set of tools to analyze a much more complex supervised learning problem involving a sequence generated by a deep transformer model.

A.2. Transformers

Lemma A.3. *For all* $t, L \in \mathbb{Z}_{++}$ *and* $\ell \in \{1, \ldots, L\}$*, if* $\theta^{(i)} \perp \theta^{(j)}$, $\tilde{\theta}^{(i)} \perp \tilde{\theta}^{(j)}$, *and* $\theta^{(i)} \perp \tilde{\theta}^{(j)}$ *for* $i \neq j$*, then*

$$\mathbb{I}(X_{t+1}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \tilde{\theta}^{(1:\ell)}, H_t) \leq \mathbb{I}(H_{t+1}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \theta^{(1:\ell-1)}, \tilde{\theta}^{(\ell)}, X_0).$$

Proof.

$$\begin{aligned}
\mathbb{I}(X_{t+1}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \tilde{\theta}^{(1:\ell)}, H_t) &\overset{(a)}{=} \mathbb{I}(H_{t+1}, \tilde{\theta}^{(1:\ell-1)}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \tilde{\theta}^{(\ell)}) - \mathbb{I}(H_t, \tilde{\theta}^{(1:\ell-1)}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \tilde{\theta}^{(\ell)}) \\
&\overset{(b)}{\leq} \mathbb{I}(H_{t+1}, \tilde{\theta}^{(1:\ell-1)}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \tilde{\theta}^{(\ell)}, X_0) \\
&\overset{(c)}{\leq} \mathbb{I}(H_{t+1}, \theta^{(1:\ell-1)}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \tilde{\theta}^{(\ell)}, X_0) \\
&\overset{(d)}{=} \mathbb{I}(H_{t+1}, \theta^{(1:\ell-1)}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \tilde{\theta}^{(\ell)}, X_0) - \mathbb{I}(\theta^{(1:\ell-1)}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \tilde{\theta}^{(\ell)}, X_0) \\
&\overset{(e)}{=} \mathbb{I}(H_{t+1}; \theta^{(\ell)}|\theta^{(\ell+1:L)}, \theta^{(1:\ell-1)}, \tilde{\theta}^{(\ell)}, X_0)
\end{aligned}$$

where (a) follows from the chain rule of mutual information, (b) follows from the independence assumptions, (c) follows from the data processing inequality applied to the markov chain $\theta_i \perp \tilde{\theta}^{(1:\ell-1)}|(H_{t+1}, \theta^{(\ell+1:L)}, \theta^{(1:\ell-1)}, X_0^K)$, (d) follows from the fact that $\mathbb{I}(\theta^{(1:\ell-1)}; \theta_i|\theta^{(\ell+1:L)}, \tilde{\theta}_i, X_0) = 0$, and (e) follows from the chain rule of mutual information. $\qquad\square$

Lemma A.4. (transformer layer Lipschitz) *For all* $d, r, K, \ell \in \mathbb{Z}_{++}$, *if*

$$\tilde{U}^{(\ell)} = \text{Clip}\left(V^{(\ell)}\tilde{U}^{(\ell-1)}\text{Attn}^{(\ell)}\left(\tilde{U}^{(\ell-1)}\right)\right),$$

then

$$\left\|U^{(\ell)} - \tilde{U}^{(\ell)}\right\|_F^2 \leq 2K\|V^{(\ell)}\|_\sigma^2\left(1 + \frac{4K\|A^{(\ell)}\|_\sigma^2}{r}\right) \cdot \left\|U^{(\ell-1)} - \tilde{U}^{(\ell-1)}\right\|_F^2.$$

Proof.

$$\left\|U^{(\ell)} - \tilde{U}^{(\ell)}\right\|_F^2$$

$$= \left\|\text{Clip}\left(V^{(\ell)}U^{(\ell-1)}\sigma\left(\frac{U^{(\ell-1)\top}A^{(\ell)}U^{(\ell-1)}}{\sqrt{r}}\right)\right) - \text{Clip}\left(V^{(\ell)}\tilde{U}^{(\ell-1)}\sigma\left(\frac{\tilde{U}^{(\ell-1)\top}A^{(\ell)}\tilde{U}^{(\ell-1)}}{\sqrt{r}}\right)\right)\right\|_F^2$$

$$\overset{(a)}{\leq} \left\|V^{(\ell)}U^{(\ell-1)}\sigma\left(\frac{U^{(\ell-1)\top}A^{(\ell)}U^{(\ell-1)}}{\sqrt{r}}\right) - V^{(\ell)}\tilde{U}^{(\ell-1)}\sigma\left(\frac{\tilde{U}^{(\ell-1)\top}A^{(\ell)}\tilde{U}^{(\ell-1)}}{\sqrt{r}}\right)\right\|_F^2$$

$$\overset{(b)}{=} \sum_{k=1}^{K}\|V^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)}\sigma\left(\frac{U^{(\ell-1)\top}A^{(\ell)}U_k^{(\ell-1)}}{\sqrt{r}}\right) - \tilde{U}^{(\ell-1)}\sigma\left(\frac{\tilde{U}^{(\ell-1)\top}A^{(\ell)}\tilde{U}_k^{(\ell-1)}}{\sqrt{r}}\right)\right\|_2^2$$

$$\leq \sum_{k=1}^{K}2\|V^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)}\sigma\left(\frac{U^{(\ell-1)\top}A^{(\ell)}U_k^{(\ell-1)}}{\sqrt{r}}\right) - \tilde{U}^{(\ell-1)}\sigma\left(\frac{U^{(\ell-1)\top}A^{(\ell)}U_k^{(\ell-1)}}{\sqrt{r}}\right)\right\|_2^2$$

$$+ \sum_{k=1}^{K}2\|V^{(\ell)}\|_\sigma^2\left\|\tilde{U}^{(\ell-1)}\sigma\left(\frac{U^{(\ell-1)\top}A^{(\ell)}U_k^{(\ell-1)}}{\sqrt{r}}\right) - \tilde{U}^{(\ell-1)}\sigma\left(\frac{\tilde{U}^{(\ell-1)\top}A^{(\ell)}\tilde{U}_k^{(\ell-1)}}{\sqrt{r}}\right)\right\|_2^2$$

$$\overset{(c)}{\leq} \sum_{k=1}^{K}2\|V^{(\ell)}\|_\sigma^2\|U^{(\ell-1)} - \tilde{U}^{(\ell-1)}\|_F^2$$

$$+ \sum_{k=1}^{K}\frac{2K}{r}\|V^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)\top}A^{(\ell)}U_k^{(\ell-1)} - \tilde{U}^{(\ell-1)\top}A^{(\ell)}\tilde{U}_k^{(\ell-1)}\right\|_2^2$$

$$\overset{(d)}{\leq} 2K\|V^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)} - \tilde{U}^{(\ell-1)}\right\|_F^2 + \frac{4K}{r}\sum_{k=1}^{K}\|V^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)\top}A^{(\ell)}U_k^{(\ell-1)} - U^{(\ell-1)\top}A^{(\ell)}\tilde{U}_k^{(\ell-1)}\right\|_2^2$$

$$+ \frac{4K}{r}\sum_{k=1}^{K}\|V^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)\top}A^{(\ell)}\tilde{U}_k^{(\ell-1)} - \tilde{U}^{(\ell-1)\top}A^{(\ell)}\tilde{U}_k^{(\ell-1)}\right\|_2^2$$

$$\overset{(e)}{\leq} 2K\|V^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)} - \tilde{U}^{(\ell-1)}\right\|_F^2 + \frac{4K^2}{r}\sum_{k=1}^{K}\|V^{(\ell)}\|_\sigma^2\|A^{(\ell)}\|_\sigma^2\left\|U_k^{(\ell-1)} - \tilde{U}_k^{(\ell-1)}\right\|_2^2$$

$$+ \frac{4K}{r}\sum_{k=1}^{K}\|V^{(\ell)}\|_\sigma^2\|A^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)} - \tilde{U}^{(\ell-1)}\right\|_F^2$$

$$= 2K\|V^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)} - \tilde{U}^{(\ell-1)}\right\|_F^2 + \frac{8K^2}{r}\|V^{(\ell)}\|_\sigma^2\|A^{(\ell)}\|_\sigma^2\left\|U^{(\ell-1)} - \tilde{U}^{(\ell-1)}\right\|_F^2$$

$$= 2K\|V^{(\ell)}\|_\sigma^2\left(1 + \frac{4K\|A^{(\ell)}\|_\sigma^2}{r}\right) \cdot \left\|U^{(\ell-1)} - \tilde{U}^{(\ell-1)}\right\|_F^2,$$

where (a) follows from the fact that Clip is a contraction mapping, where in (b), $U_k^{(\ell-1)}$ denotes the kth column of $U^{(\ell-1)} \in \Re^{d \times K}$, (c) follows from the fact $\|\tilde{U}^{(\ell-1)}\|_\sigma^2 \leq K$ and the fact that softmax is 1-Lipschitz, (d) follows from the fact that $(x+z)^2 \leq 2(x-y)^2 + 2(y+z)^2$, and (e) follows from the fact that $\|U^{(\ell-1)}U^{(\ell-1)\top}\|_\sigma \leq K$. $\qquad\square$

Lemma A.5. *For all $d, r, K \in \mathbb{Z}_{++}$ and $\epsilon \geq 0$, if $V \in \Re^{r \times r}$ consists of elements distributed iid $\mathcal{N}(0, 1/r)$, $A \in \Re^{r \times r}$ consists of elements distributed $\mathcal{N}(0, 1)$, for all $\ell \in [L]$,*

$$\tilde{U}^{(\ell)} = \text{Clip}\left(\tilde{V}^{(\ell)} U^{(\ell-1)} \tilde{\text{Attn}}^{(\ell)}\left(U^{(\ell-1)}\right)\right),$$

where for all $\ell < L$, $\mathbb{E}[\|V^{(\ell)} - \tilde{V}^{(\ell)}\|_F^2] \leq \epsilon$, $\mathbb{E}[\|A^{(\ell)} - \tilde{A}^{(\ell)}\|_F^2] \leq \epsilon$, and $\mathbb{E}[\|V^{(L)} - \tilde{V}^{(L)}\|_F^2] \leq \epsilon$, $\mathbb{E}[\|A^{(L)} - \tilde{A}^{(L)}\|_F^2] \leq r\epsilon/d$, then

$$\mathbb{E}\left[\|U^{(\ell)} - \tilde{U}^{(\ell)}\|_F^2\right] \leq \begin{cases} 2K^2 (1 + K)\epsilon & \text{if } \ell < L \\ 2K (1 + K)\epsilon & \text{if } \ell = L \end{cases}.$$

Proof.

$$\mathbb{E}\left[\left\|U^{(\ell)} - \tilde{U}^{(\ell)}\right\|_F^2\right]$$

$$\leq \mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V^{(\ell)} u \text{Attn}^{(\ell)}(u) - \tilde{V}^{(\ell)} u \tilde{\text{Attn}}^{(\ell)}(u)\right\|_F^2\right]$$

$$= \mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V u \sigma\left(\frac{u^\top A u}{\sqrt{r}}\right) - \tilde{V} u \sigma\left(\frac{u^\top \tilde{A} u}{\sqrt{r}}\right)\right\|_F^2\right]$$

$$\stackrel{(a)}{\leq} 2\mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|\left(V - \tilde{V}\right) u \sigma\left(\frac{u^\top \tilde{A} u}{\sqrt{r}}\right)\right\|_F^2\right] + 2\mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V u \left(\sigma\left(\frac{u^\top A u}{\sqrt{r}}\right) - \sigma\left(\frac{u^\top \tilde{A} u}{\sqrt{r}}\right)\right)\right\|_F^2\right]$$

$$\stackrel{(b)}{\leq} 2\mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V - \tilde{V}\right\|_F^2 \left\|u \sigma\left(\frac{u^\top \tilde{A} u}{\sqrt{r}}\right)\right\|_F^2\right] + 2\mathbb{E}\left[\sup_{u \in \mathcal{U}} \|V\|_F^2 \left\|u \sigma\left(\frac{u^\top A u}{\sqrt{r}}\right) - u \sigma\left(\frac{u^\top \tilde{A} u}{\sqrt{r}}\right)\right\|_F^2\right]$$

$$\stackrel{(c)}{\leq} 2\epsilon \cdot \sup_{u \in \mathcal{U}} \|u\|_F^2 \cdot \left\|\sigma\left(\frac{u^\top \tilde{A} u}{\sqrt{r}}\right)\right\|_F^2 + 2\mathbb{E}\left[\|V\|_F^2 \cdot \sup_{u \in \mathcal{U}} \|u\|_F^2 \left\|\sigma\left(\frac{u^\top A u}{\sqrt{r}}\right) - \sigma\left(\frac{u^\top \tilde{A} u}{\sqrt{r}}\right)\right\|_F^2\right]$$

$$\stackrel{(d)}{\leq} 2\epsilon K^2 + 2\mathbb{E}\left[K r \cdot \sup_{u \in \mathcal{U}} \left\|\frac{u^\top A u}{\sqrt{r}} - \frac{u^\top \tilde{A} u}{\sqrt{r}}\right\|_F^2\right]$$

$$\stackrel{(e)}{=} 2\epsilon K^2 + 2K \cdot \mathbb{E}\left[\sup_{u \in \mathcal{U}} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(u_i^\top (A - \tilde{A}) u_j\right)^2\right]$$

$$\leq 2\epsilon K^2 + 2K \cdot \mathbb{E}\left[\sum_{i=1}^{K} \sum_{j=1}^{K} \left\|A - \tilde{A}\right\|_F^2\right]$$

$$= 2K^2 \epsilon + 2K^3 \epsilon,$$

where (a) follows from the fact that $\|a + b\|_F^2 \leq 2\|a\|_F^2 + 2\|b\|_F^2$ for all matrices a, b, (b) follows from the fact that $\|ab\|_F^2 \leq \|a\|_\sigma^2 \|b\|_F^2$ and $\|a\|_\sigma^2 \leq \|a\|_F^2$ for all matrices a, b, (c) follows from the fact that $\mathbb{E}\left[\|V - \tilde{V}\|_F^2\right] = \epsilon$, (d) follows from the fact that $\mathbb{E}[\|V\|_F^2] = r$, and the fact that softmax is 1-Lipschitz, and where in (e), x_i denotes the ith column of matrix x.

For $\ell = L$,

$$\mathbb{E}\left[\left\|U^{(L)} - \tilde{U}^{(L)}\right\|_F^2\right]$$

$$\leq \mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V^{(L)}u\mathrm{Attn}^{(L)}(u)[-1] - \tilde{V}^{(L)}u\tilde{\mathrm{Attn}}^{(L)}(u)[-1]\right\|_F^2\right]$$

$$= \mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V^{(L)}u\sigma\left(\frac{u^\top A^{(L)}u[-1]}{\sqrt{r}}\right) - \tilde{V}^{(L)}u\sigma\left(\frac{u^\top \tilde{A}^{(L)}u[-1]}{\sqrt{r}}\right)\right\|_F^2\right]$$

$$\stackrel{(a)}{\leq} 2\mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|\left(V^{(L)} - \tilde{V}^{(L)}\right)u\sigma\left(\frac{u^\top \tilde{A}^{(L)}u[-1]}{\sqrt{r}}\right)\right\|_F^2\right]$$

$$+ 2\mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V^{(L)}u\left(\sigma\left(\frac{u^\top A^{(L)}u[-1]}{\sqrt{r}}\right) - \sigma\left(\frac{u^\top \tilde{A}^{(L)}u[-1]}{\sqrt{r}}\right)\right)\right\|_F^2\right]$$

$$\stackrel{(b)}{\leq} 2\mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V^{(L)} - \tilde{V}^{(L)}\right\|_F^2 \left\|u\sigma\left(\frac{u^\top \tilde{A}^{(L)}u[-1]}{\sqrt{r}}\right)\right\|_F^2\right]$$

$$+ 2\mathbb{E}\left[\sup_{u \in \mathcal{U}} \left\|V^{(L)}\right\|_F^2 \left\|u\sigma\left(\frac{u^\top A^{(L)}u[-1]}{\sqrt{r}}\right) - u\sigma\left(\frac{u^\top \tilde{A}^{(L)}u[-1]}{\sqrt{r}}\right)\right\|_F^2\right]$$

$$\stackrel{(c)}{\leq} 2\epsilon \cdot \sup_{u \in \mathcal{U}} \|u\|_F^2 \cdot \left\|\sigma\left(\frac{u^\top \tilde{A}^{(L)}u[-1]}{\sqrt{r}}\right)\right\|^2$$

$$+ 2\mathbb{E}\left[\|V^{(L)}\|_F^2 \cdot \sup_{u \in \mathcal{U}} \|u\|_F^2 \left\|\sigma\left(\frac{u^\top A^{(L)}u[-1]}{\sqrt{r}}\right) - \sigma\left(\frac{u^\top \tilde{A}^{(L)}u[-1]}{\sqrt{r}}\right)\right\|_F^2\right]$$

$$\stackrel{(d)}{\leq} 2\epsilon K + 2\mathbb{E}\left[Kd \cdot \sup_{u \in \mathcal{U}} \left\|\frac{u^\top A^{(L)}u[-1]}{\sqrt{r}} - \frac{u^\top \tilde{A}^{(L)}u[-1]}{\sqrt{r}}\right\|_F^2\right]$$

$$\stackrel{(e)}{=} 2\epsilon K + \frac{2Kd}{r} \cdot \mathbb{E}\left[\sup_{u \in \mathcal{U}} \sum_{i=1}^K \left(u_i^\top (A^{(L)} - \tilde{A}^{(L)})u[-1]\right)^2\right]$$

$$\leq 2\epsilon K + \frac{2Kd}{r} \cdot \mathbb{E}\left[\sum_{i=1}^K \left\|A^{(L)} - \tilde{A}^{(L)}\right\|_F^2\right]$$

$$= 2K\epsilon + 2K^2\epsilon,$$

where (a) follows from the fact that $\|a + b\|_F^2 \leq 2\|a\|_F^2 + 2\|b\|_F^2$ for all matrices a, b, (b) follows from the fact that $\|ab\|_F^2 \leq \|a\|_\sigma^2 \|b\|_F^2$ and $\|a\|_\sigma^2 \leq \|a\|_F^2$ for all matrices a, b, (c) follows from the fact that $\mathbb{E}\left[\|V - \tilde{V}\|_F^2\right] = \epsilon$, (d) follows from the fact that $\mathbb{E}[\|V\|_F^2] = d$, and the fact that softmax is 1-Lipschitz, and where in (e), u_i denotes the ith column of matrix u. $\qquad\square$

Lemma A.6. (transformer layer distortion upper bound) *For all $d, r, t, K, L \in \mathbb{Z}_{++}$, $0 \leq \epsilon$, and $\ell \in [L]$, if*

$$\tilde{\theta}^{(\ell)} = (\tilde{V}^{(\ell)}, \tilde{A}^{(\ell)}) = (V^{(\ell)} + Z^{(\ell)}, A^{(\ell)} + B^{(\ell)}),$$

where

$$(Z^{(\ell)}, B^{(\ell)}) \perp (V^{(\ell)}, A^{(\ell)}), \quad Z^{(\ell)} \stackrel{iid}{\sim} \begin{cases} \mathcal{N}(0, \epsilon/r^2) & \text{if } \ell < L \\ \mathcal{N}(0, \epsilon/(rd)) & \text{if } \ell = L \end{cases}; \quad B^{(\ell)} \stackrel{iid}{\sim} \mathcal{N}(0, \epsilon/r^2),$$

then

$$\mathbb{I}(H_{t+1}; \theta^{(\ell)} | \theta^{(\ell+1:L)}, \theta^{(1:\ell-1)}, \tilde{\theta}^{(\ell)}, X_0) \leq \epsilon(t+1)K\left(8K(1+16K)\right)^{L-\ell+1}.$$

Proof. We begin with the base cases $\ell = L$

$$
\begin{aligned}
\mathbb{I}(X_{t+1}; \theta^{(L)} | \theta^{(1:L-1)}, \tilde{\theta}^{(L)}, H_t) &= \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(X_{t+1} \in \cdot | \theta^{(1:L)}, H_t) \,\|\, \mathbb{P}(X_{t+1} \in \cdot | \theta^{(1:L-1)}, \tilde{\theta}^{(L)}, H_t)\right)\right] \\
&\leq \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(X_{t+1} \in \cdot | \theta^{(1:L)}, H_t) \,\|\, \mathbb{P}(X_{t+1} \in \cdot | \theta^{(1:L-1)}, \theta^{(L)} \leftarrow \tilde{\theta}^{(L)}, H_t)\right)\right] \\
&\leq \mathbb{E}\left[\left\| f_{\theta^{(1:L)}}(X_{t:t-K+1}) - f_{\tilde{\theta}^{(L)}}\left(f_{\theta^{(1:L-1)}}(X_{t:t-K+1})\right) \right\|_2^2\right] \\
&\leq 2K(1+K)\epsilon,
\end{aligned}
$$

For $\ell < L$, we have:

$$
\begin{aligned}
&\mathbb{I}(X_{t+1}; \theta^{(\ell)} | \theta^{(1:\ell-1)}, \theta^{(\ell+1:L)}, \tilde{\theta}^{(\ell)}, H_t) \\
&= \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(X_{t+1} \in \cdot | \theta^{(1:L)}, H_t\right) \,\|\, \mathbb{P}\left(X_{t+1} \in \cdot | \theta^{(1:\ell-1)}, \theta^{(\ell+1:L)}, \tilde{\theta}^{(\ell)}, H_t\right)\right)\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(X_{t+1} \in \cdot | \theta^{(1:L)}, H_t\right) \,\|\, \mathbb{P}\left(X_{t+1} \in \cdot | \theta^{(1:\ell-1)}, \theta^{(\ell+1:L)}, \theta^{(\ell)} \leftarrow \tilde{\theta}^{(\ell)}, H_t\right)\right)\right] \\
&\overset{(b)}{\leq} \mathbb{E}\left[\left\| f_{\theta^{(1:L)}}(X_{t:t-K+1}) - \left(f_{\theta^{(\ell+1:L)}} \circ f_{\tilde{\theta}^{(\ell)}} \circ f_{\theta^{(1:\ell-1)}}\right)(X_{t:t-K+1}) \right\|_2^2\right] \\
&\overset{(c)}{\leq} \mathbb{E}\left[2K\|V^{(L)}\|_\sigma^2 \left(1 + \frac{4K\|A^{(L)}\|_\sigma^2}{r}\right) \cdot \left\| f_{\theta^{(1:L-1)}}(X_{t:t-K+1}) - \left(f_{\theta^{(\ell+1:L-1)}} \circ f_{\tilde{\theta}^{(\ell)}} \circ f_{\theta^{(1:\ell-1)}}\right)(X_{t:t-K+1}) \right\|_2^2\right] \\
&\overset{(d)}{\leq} \mathbb{E}\left[\left(\prod_{l=\ell+1}^{L} 2K\|V^{(l)}\|_\sigma^2 \left(1 + \frac{4K\|A^{(l)}\|_\sigma^2}{r}\right)\right) \cdot \left\| f_{\theta^{(1:\ell)}}(X_{t:t-K+1}) - \left(f_{\tilde{\theta}^{(\ell)}} \circ f_{\theta^{(1:\ell-1)}}\right)(X_{t:t-K+1}) \right\|_2^2\right] \\
&\overset{(e)}{\leq} \mathbb{E}\left[\left(\prod_{l=\ell+1}^{L} 2K\|V^{(l)}\|_\sigma^2 \left(1 + \frac{4K\|A^{(l)}\|_\sigma^2}{r}\right)\right) \cdot 2K^2(1+K)\epsilon\right] \\
&\overset{(f)}{\leq} (8K(1+16K))^{L-\ell} \cdot 2K^2(1+K)\epsilon \\
&\leq \epsilon K \left(8K(1+16K)\right)^{L-\ell+1}
\end{aligned}
$$

where (a) follows from Lemma A.1, (b) follows from Lemma B.2, (c) and (d) follow from Lemma A.4, (e) follows from Lemma A.5 and the fact that $(V^{(i)}, A^{(i)} \perp (V^{(j)}, A^{(j)}))$ and $(V^{(i)}, A^{(i)} \perp (\tilde{V}^{(j)}, \tilde{A}^{(j)}))$ for $i \neq j$, and (f) follows from the fact that $A^{(L)} \perp V^{(L)}$ and $\mathbb{E}\left[\|A^{(L)}\|_\sigma^2\right] \leq 4r$ and $\mathbb{E}\left[\|V^{(L)}\|_\sigma^2\right] \leq 4$.

With this inequality in place, we have the following.

$$
\begin{aligned}
\mathbb{I}(H_{t+1}; \theta^{(\ell)} | \theta^{(\ell+1:L)}, \theta^{(1:\ell-1)}, \tilde{\theta}^{(\ell)}, X_0) &= \sum_{k=0}^{t} \mathbb{I}(X_{k+1}; \theta^{(\ell)} | \theta^{(\ell+1:L)}, \theta^{(1:\ell-1)}, \tilde{\theta}^{(\ell)}, H_k) \\
&\overset{(a)}{\leq} \sum_{k=0}^{t} \epsilon K \left(8K(1+16K)\right)^{L-\ell+1} \\
&= \epsilon(t+1)K \left(8K(1+16K)\right)^{L-\ell+1},
\end{aligned}
$$

where (a) follows from the above result. $\qquad \square$

Lemma A.7. (transformer distortion upper bound) *For all $d, r, t, K, L \in \mathbb{Z}_{++}$, and $\ell \in [L]$, if*

$$
\tilde{\theta}^{(\ell)} = (\tilde{V}^{(\ell)}, \tilde{A}^{(\ell)}) = (V^{(\ell)} + Z^{(\ell)}, A^{(\ell)} + B^{(\ell)}),
$$

where

$$
(Z^{(\ell)}, B^{(\ell)}) \perp (V^{(\ell)}, A^{(\ell)}), \quad Z^{(\ell)} \overset{iid}{\sim} \begin{cases} \mathcal{N}(0, \epsilon/r^2) & \text{if } \ell < L \\ \mathcal{N}(0, \epsilon/(rd)) & \text{if } \ell = L \end{cases}; \quad B^{(\ell)} \overset{iid}{\sim} \mathcal{N}(0, \epsilon/r^2),
$$

then

$$
\mathbb{I}(X_{t+1}; \theta^{(1:L)} | \tilde{\theta}^{(1:L)}, H_t) \leq \epsilon K L(t+1) \left(8K(1+16K)\right)^{L}.
$$

Proof.

$$\mathbb{I}(X_{t+1}; \theta^{(1:L)} | \tilde{\theta}^{(1:L)}, H_t) = \sum_{\ell=1}^{L} \mathbb{I}(X_{t+1}; \theta^{(\ell)} | \tilde{\theta}^{(1:L)}, \theta^{(\ell+1:L)} H_t)$$

$$\overset{(a)}{\leq} \sum_{\ell=1}^{L} \mathbb{I}(H_{t+1}; \theta^{(\ell)} | \theta^{(\ell+1:L)}, \theta^{(1:\ell-1)}, \tilde{\theta}^{(\ell)}, X_0)$$

$$\overset{(b)}{\leq} \epsilon K L (t+1) \left(8K(1+16K)\right)^L,$$

where (a) follows from Lemma A.3, and (b) follows from Lemma A.6. $\qquad\square$

Lemma A.8. (transformer rate-distortion upper bound) *For all $d, r, t, K, L \in \mathbb{Z}_{++}$, if for all t, X_t is generated by the transformer process, then*

$$\mathbb{H}_{\epsilon,T}(\theta^{(1:L)}) \leq r \cdot \max\{r, d\} L \ln\left(1 + \frac{r \cdot \max\{r, d\} KLT(8K(1+16K))^L}{\epsilon}\right).$$

Proof. For all $\ell \in [L]$, let

$$\tilde{\theta}^{(\ell)} = (\tilde{V}^{(\ell)}, \tilde{A}^{(\ell)}) = (V^{(\ell)} + Z^{(\ell)}, A^{(\ell)} + B^{(\ell)}),$$

where

$$(Z^{(\ell)}, B^{(\ell)}) \perp (V^{(\ell)}, A^{(\ell)}), \quad Z^{(\ell)} \overset{iid}{\sim} \begin{cases} \mathcal{N}(0, \epsilon'/r^2) & \text{if } \ell < L \\ \mathcal{N}(0, \epsilon'/(rd)) & \text{if } \ell = L \end{cases}; \quad B^{(\ell)} \overset{iid}{\sim} \mathcal{N}(0, \epsilon'/r^2).$$

Let $\epsilon' = \frac{\epsilon}{KLT(8K(1+16K))^L}$. Then,

$$\mathbb{I}(\theta^{(1:L)}; \tilde{\theta}^{(1:L)}) = \mathbf{h}(\tilde{\theta}^{(1:L)}) - \mathbf{h}(\tilde{\theta}^{(1:L)} | \theta^{(1:L)})$$

$$= \sum_{\ell=1}^{L} \mathbf{h}(\tilde{\theta}^{(\ell)}) - \mathbf{h}(\tilde{\theta}^{(\ell)} | \theta^{(\ell)})$$

$$= \sum_{\ell=1}^{L} \left(\mathbf{h}(\tilde{V}^{(\ell)}) - \mathbf{h}(\tilde{V}^{(\ell)} | V^{(\ell)}) + \mathbf{h}(\tilde{A}^{(\ell)}) - \mathbf{h}(\tilde{A}^{(\ell)} | A^{(\ell)}) \right)$$

$$\leq \sum_{\ell=1}^{L-1} \left(\frac{r^2}{2} \ln\left(2\pi e \left(\frac{1}{r} + \frac{\epsilon'}{r^2}\right)\right) - \frac{r^2}{2} \ln\left(2\pi e \left(\frac{\epsilon'}{r^2}\right)\right) \right)$$

$$+ \left(\frac{rd}{2} \ln\left(2\pi e \left(\frac{1}{r} + \frac{\epsilon'}{rd}\right)\right) - \frac{rd}{2} \ln\left(2\pi e \left(\frac{\epsilon'}{rd}\right)\right) \right)$$

$$+ \sum_{\ell=1}^{L} \left(\frac{r^2}{2} \ln\left(2\pi e \left(1 + \frac{\epsilon'}{r^2}\right)\right) - \frac{r^2}{2} \ln\left(2\pi e \left(\frac{\epsilon'}{r^2}\right)\right) \right)$$

$$\leq \frac{Lr \cdot \max\{r, d\}}{2} \ln\left(1 + \frac{\max\{r, d\} KLT(8K(1+16K))^L}{\epsilon}\right)$$

$$+ \frac{Lr^2}{2} \ln\left(1 + \frac{r^2 KLT(8K(1+16K)))^L}{\epsilon}\right)$$

$$\leq Lr \cdot \max\{r, d\} \ln\left(1 + \frac{r \cdot \max\{r, d\} KLT(8K(1+16K))^L}{\epsilon}\right),$$

We now verify that the distortion is $\leq \epsilon$.

$$
\begin{aligned}
\frac{\mathbb{I}(H_T; \theta^{(1:L)} | \tilde{\theta}^{(1:L)})}{T} &= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}; \theta^{(1:L)} | \tilde{\theta}^{(1:L)}, H_t) \\
&\leq \frac{1}{T} \sum_{t=0}^{T-1} \epsilon' KL(t+1) \left(8K(1+16K)\right)^L \\
&\leq \epsilon' KLT \left(8K(1+16K)\right)^L \\
&= \epsilon.
\end{aligned}
$$

$\square$

Theorem 3.5. (estimation error bound) *For all d, r, L, K, T, if for all t, X_t is generated by the transformer environment, then*

$$
\mathcal{L}_T \leq \frac{pL \ln\left(136eK^2\right) + p\ln\left(\frac{2KT^2}{L}\right)}{T},
$$

where $p = 2r^2(L-1) + (dr + r^2)$ denotes the parameter count of the transformer.

Proof.

$$
\begin{aligned}
\mathcal{L}_T &\overset{(a)}{\leq} \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{\epsilon, T}(\theta^{(1:L)})}{T} + \epsilon \\
&\overset{(b)}{\leq} \inf_{\epsilon \geq 0} \frac{r \cdot \max\{r, d\} L \ln\left(1 + \frac{r \cdot \max\{r, d\} KLT (8K(1+16K))^L}{\epsilon}\right)}{T} + \epsilon \\
&\overset{(c)}{\leq} \frac{r \cdot \max\{r, d\} L \ln\left(1 + \frac{r \cdot \max\{r, d\} KLT (8K(1+16K))^L}{\frac{r \cdot \max\{r, d\} L^2}{T}}\right)}{T} + \frac{r \cdot \max\{r, d\} L^2}{T} \\
&\leq \frac{r \cdot \max\{r, d\} L \ln\left(\frac{2KT^2 (8K(1+16K))^L}{L}\right)}{T} + \frac{r \cdot \max\{r, d\} L^2}{T} \\
&= \frac{r \cdot \max\{r, d\} L^2 \ln\left(8eK(1+16K)\right)}{T} + \frac{r \cdot \max\{r, d\} L \ln\left(\frac{2KT^2}{L}\right)}{T},
\end{aligned}
$$

where (a) follows from Theorem 3.4, (b) follows from Lemma A.8, (c) follows by setting $\epsilon = r \cdot \max\{r, d\} L^2 / T$ $\quad\square$

B. Meta-Learning from Sequential Data

Lemma 4.1. (Bayesian posterior is optimal) *For all $m, t \in \mathbb{Z}_+$,*

$$
\mathbb{E}\left[-\ln \hat{P}_{m,t}\left(X_{t+1}^{(m)}\right)\right] = \min_{\pi} \mathbb{E}_{\pi}\left[-\ln P_{m,t}\left(X_{t+1}^{(m)}\right)\right].
$$

Proof. In the below proof take all equality to hold *almost surely*.

$$
\begin{aligned}
&\mathbb{E}\left[-\ln P_{m,t}\left(X_{t+1}^{(m)}\right) | H_{m,t}\right] \\
&= \mathbb{E}\left[-\ln \hat{P}_{m,t}\left(X_{t+1}^{(m)}\right) + \ln \frac{\hat{P}_{m,t}(X_{t+1}^{(m)})}{P_{m,t}(X_{t+1}^{(m)})} \Big| H_{m,t}\right] \\
&= \mathbb{E}\left[-\ln \hat{P}_{m,t}\left(X_{t+1}^{(m)}\right) \Big| H_{m,t}\right] + \mathbf{d}_{\mathrm{KL}}(\hat{P}_{m,t} \| P_{m,t}).
\end{aligned}
$$

The result follows from the fact that $\mathbf{d}_{\mathrm{KL}}(\hat{P}_{m,t} \| P_{m,t}) > 0$ for all $P_{m,t} \neq \hat{P}_{m,t}$. $\quad\square$

Theorem 4.2. (Main Result) *For all $M, T \in \mathbb{Z}_+$ and $m \in \{1, 2, \ldots, M\}$,*

$$\mathcal{L}_{M,T} = \underbrace{\frac{\mathbb{I}(H_{M,T}; \psi)}{MT}}_{\substack{meta \\ estimation \\ error}} + \underbrace{\frac{\mathbb{I}(D_m; \theta_m|\psi)}{T}}_{\substack{intra\text{-}document \\ estimation \\ error}}.$$

Proof.

$$\mathbb{L}_{M,T} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[-\ln \hat{P}_{m,t}(X_{t+1}^{(m)})\right]$$

$$= \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{H}(X_{t+1}^{(m)}|\theta_m, X_1^{(m)}, \ldots, X_t^{(m)})$$

$$+ \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(X_{t+1}^{(m)} \in \cdot|\psi, \theta_m, X_1^{(m)}, \ldots, X_t^{(m)})\|\mathbb{P}(X_{t+1}^{(m)} \in \cdot|H_{m,t})\right)\right]$$

$$= \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}^{(m)}; \psi, \theta_m|H_{m,t}) + \mathbb{H}(X_{t+1}^{(m)}|\theta_m, X_1^{(m)}, \ldots, X_t^{(m)})$$

$$= \frac{1}{MT} \sum_{m=1}^{M} \mathbb{I}(D_m; \psi, \theta_m|H_{m-1,T}) + \mathbb{H}(D_m|\theta_m)$$

$$= \frac{1}{MT} \sum_{m=1}^{M} \mathbb{I}(D_m; \psi|H_{m-1,T}) + \mathbb{I}(D_m; \theta_m|\psi, H_{m-1,T}) + \mathbb{H}(D_m|\theta_m)$$

$$= \frac{\mathbb{I}(H_{M,T}; \psi)}{MT} + \frac{\mathbb{I}(D_m; \theta_m|\psi)}{T} + \frac{1}{MT} \sum_{m=1}^{M} \mathbb{H}(D_m|\theta_m).$$

$\square$

Theorem 4.3. (rate-distortion estimation error bound) *For all $M, T \in \mathbb{Z}_+$, and $m \in \{1, \ldots, M\}$,*

$$\mathcal{L}_{M,T} \leq \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{\epsilon,M,T}(\psi)}{MT} + \epsilon + \inf_{\epsilon' \geq 0} \frac{\mathbb{H}_{\epsilon',T}(\theta_m|\psi)}{T} + \epsilon',$$

and

$$\mathcal{L}_{M,T} \geq \sup_{\epsilon \geq 0} \min\left\{\frac{\mathbb{H}_{\epsilon,M,T}(\psi)}{MT}, \epsilon\right\}$$
$$+ \sup_{\epsilon' \geq 0} \min\left\{\frac{\mathbb{H}_{\epsilon',T}(\theta_m|\psi)}{T}, \epsilon'\right\}.$$

22

Proof. We begin by showing the upper bound:

$$
\begin{aligned}
\mathcal{L}_{M,T} &= \frac{\mathbb{I}(H_{M,T};\psi)}{MT} + \frac{\mathbb{I}(D_m;\theta_m|\psi)}{T} \\
&= \frac{\mathbb{I}(H_{M,T};\psi,\tilde{\psi})}{MT} + \frac{\mathbb{I}(D_m;\theta_m,\tilde{\theta}_m|\psi)}{T} \\
&= \frac{\mathbb{I}(H_{M,T};\tilde{\psi})}{MT} + \frac{\mathbb{I}(H_{M,T};\psi|\tilde{\psi})}{MT} + \frac{\mathbb{I}(D_m;\theta_m,\tilde{\theta}_m|\psi)}{T} \\
&\overset{(a)}{\leq} \frac{\mathbb{I}(\psi;\tilde{\psi})}{MT} + \frac{\mathbb{I}(H_{M,T};\psi|\tilde{\psi})}{MT} + \frac{\mathbb{I}(D_m;\theta_m,\tilde{\theta}_m|\psi)}{T} \\
&= \frac{\mathbb{I}(\psi;\tilde{\psi})}{MT} + \frac{\mathbb{I}(H_{M,T};\psi|\tilde{\psi})}{MT} + \frac{\mathbb{I}(D_m;\tilde{\theta}_m|\psi)}{T} + \frac{\mathbb{I}(D_m;\theta_m|\tilde{\theta}_m,\psi)}{T} \\
&\overset{(b)}{\leq} \frac{\mathbb{I}(\psi;\tilde{\psi})}{MT} + \frac{\mathbb{I}(H_{M,T};\psi|\tilde{\psi})}{MT} + \frac{\mathbb{I}(\theta_m;\tilde{\theta}_m|\psi)}{T} + \frac{\mathbb{I}(D_m;\theta_m|\tilde{\theta}_m,\psi)}{T} \\
&\overset{(c)}{\leq} \frac{\mathbb{H}_{\epsilon,M,T}(\psi)}{MT} + \epsilon + \frac{\mathbb{H}_{\epsilon',M,T}(\tilde{\theta}_m|\psi)}{T} + \epsilon',
\end{aligned}
$$

where (a) and (b) follow from the data processing inequality and (c) follows from the definition of the rate-distortion functions. The upper bound follows from the fact that inequality (c) holds for all $\epsilon \geq 0$.

We now prove the lower bound. Suppose that $\mathbb{I}(H_{M,T};\psi) < \mathbb{H}_{\epsilon,M,T}(\psi)$ Let $\tilde{\psi} = \tilde{H}_{M,T} \notin \tilde{\Psi}_{\epsilon,M,T}$ where $\tilde{H}_{M,T}$ is another history sampled in the same manner as $H_{M,T}$.

$$
\begin{aligned}
\mathbb{I}(H_{M,T};\psi) &= \sum_{m=1}^{M}\sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}^{(m)};\psi|H_{m,t}) \\
&\overset{(a)}{\geq} \sum_{m=1}^{M}\sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}^{(m)};\psi|\tilde{H}_{M,T},H_{m,t}) \\
&= \sum_{m=1}^{M}\sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}^{(m)};\psi|\tilde{\psi},X_1^{(m)},\ldots,X_t^{(m)}) \\
&\overset{(b)}{\geq} \epsilon MT,
\end{aligned}
$$

where (a) follows from the fact that conditioning reduces entropy and that $X_{t+1}^{(m)} \perp \tilde{H}_{M,T}|(\psi,H_{m,t})$ and (b) follows from the fact that $\tilde{\psi} \notin \tilde{\Psi}_{\epsilon,M,T}$. Therefore, for all $\epsilon \geq 0$, $\mathbb{I}(H_{M,T};\psi) \geq \min\{\mathbb{H}_{\epsilon,M,T}(\psi),\epsilon MT\}$.

Suppose that $\mathbb{I}(H_T^{(m)};\theta_m|\psi) < \mathbb{H}_{\epsilon,T}(\theta_m|\psi)$. Let $\tilde{\theta}_m = \tilde{D}_m \notin \tilde{\Theta}_{\epsilon,T}$ where $\tilde{D}_m$ is another history sampled in the same manner as D_m.

$$
\begin{aligned}
\mathbb{I}(D_m;\theta_m|\psi) &= \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}^{(m)};\theta_m|X_1^{(m)},\ldots,X_t^{(m)},\psi) \\
&\overset{(a)}{\geq} \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}^{(m)};\theta_m|\tilde{D}_m,X_1^{(m)},\ldots,X_t^{(m)},\psi) \\
&= \sum_{t=0}^{T-1} \mathbb{I}(X_{t+1}^{(m)};\theta_m|\tilde{\theta}_m,H_{m,t},\psi) \\
&\overset{(b)}{\geq} \epsilon T,
\end{aligned}
$$

where (a) follows from the fact that conditioning reduces entropy and that $X_{t+1}^{(m)} \perp \tilde{D}_m|(\psi,X_1^{(m)},\ldots,X_t^{(m)})$ and (b) follows from the fact that $\tilde{\theta}_m \notin \tilde{\Theta}_{\epsilon,T}$. Therefore, for all $\epsilon \geq 0$, $\mathbb{I}(D_m;\theta_m|\psi) \geq \min\{\mathbb{H}_{\epsilon,M,T}(\theta_m),\epsilon T\}$. The lower bound follows as a result. $\qquad\square$

B.1. Linear Representation Learning Example

We introduce a simple linear representation learning problem as a concrete example of meta-learning to demonstrate our method of analysis. Just as in the logistic regression example, the documents in this example consist of iid data but we begin with such an example for simplicity and to demonstrate this as a special case of meta-learning from sequences under our framework.

For all $d, r \in \mathbb{Z}_{++}$, we let $\psi : \Omega \mapsto \Re^{d \times r}$ be distributed uniformly over the set of $d \times r$ matrices with orthonormal columns. We assume that $d \gg r$. For all m, let $\xi_m : \Omega \mapsto \Re^r$ be distributed iid $\mathcal{N}(0, I_r/r)$. We let $\theta_m = \psi \xi_m$ and hence ψ induces a distribution on θ_m. As for the observable data, for each (m, t), $X_t^{(m)}$ be drawn as according to the following probability law:

$$X_t^{(m)} = \begin{cases} 1 & \text{w.p. } \sigma(\theta_m)_1 \\ 2 & \text{w.p. } \sigma(\theta_m)_2 \\ \dots & \\ d & \text{w.p. } \sigma(\theta_m)_d \end{cases},$$

where $\sigma(\theta_m)_j = e^{\theta_{m,j}} / \sum_{k=1}^{d} e^{\theta_{m,k}}$ denotes softmax. Note that in this problem, the input X does not influence the output Y. For each task m, the algorithm is tasked with estimating a vector θ_m from noisy observations $(X_1^{(m)}, \dots, X_T^{(m)})$. By reasoning about data from previous tasks, the algorithm can estimate ψ which reduces the burden of estimating θ_m to just estimating ξ_m for each task. This is significant given the assumption that $d \gg r$. We now present the theoretical result.

Theorem B.1. (linear representation learning Bayesian error bound) *For all $d, r, M, T \in \mathbb{Z}_{++}$,*

$$\mathcal{L}_{M,T} \leq \frac{dr\left(1 + \ln\left(1 + \frac{M}{r}\right)\right)}{2MT} + \frac{r\left(1 + \ln(1 + \frac{2n}{r})\right)}{2T}.$$

The first term indicates the statistical error incurred in the process of estimating ψ. Since $\psi \in \Re^{d \times r}$ and there are $M \times T$ data points in total which contain information about ψ. The second term represents statistical error incurred in the process of estimating $\xi_1, \dots, \xi_T$. Since each $\xi_m \in \Re^r$ and there are T data points which contain information about each ξ_i the $\tilde{O}(r/T)$ follows standard statistical intuition.

We note that this tightens a result shown in (Tripuraneni et al., 2021) which studies an almost identical problem. Their proposed upper bound is $\tilde{\mathcal{O}}(\frac{dr^2}{MT} + \frac{r}{T})$ which contains an extra factor of r in the meta-estimation error.

Lemma B.2. (sq error upper bounds softmax KL-divergence) *For all $d \in \mathbb{Z}_{++}$ and random vectors $\theta, \tilde{\theta} \in \Re^d$,*

$$\mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \ln \frac{\frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}}}{\frac{e^{\tilde{\theta}_l}}{\sum_{k=1}^{d} e^{\tilde{\theta}_k}}}\right] \leq \mathbb{E}\left[\|\tilde{\theta} - \theta\|_2^2\right].$$

Proof.

$$
\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot|\theta)\|\mathbf{d}_{\mathrm{KL}}(Y \in \cdot|\theta \leftarrow \tilde{\theta}))\right] = \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \ln \frac{\frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}}}{\frac{e^{\tilde{\theta}_l}}{\sum_{k=1}^{d} e^{\tilde{\theta}_k}}}\right]
$$

$$
= \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \left(\ln \frac{e^{\theta_l}}{e^{\tilde{\theta}_l}} + \ln \frac{\sum_{k=1}^{d} e^{\tilde{\theta}_k}}{\sum_{k=1}^{d} e^{\theta_k}}\right)\right]
$$

$$
= \mathbb{E}\left[\ln \frac{\sum_{k=1}^{d} e^{\tilde{\theta}_k}}{\sum_{k=1}^{d} e^{\theta_k}}\right] + \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \ln \frac{e^{\theta_l}}{e^{\tilde{\theta}_l}}\right]
$$

$$
= \mathbb{E}\left[\ln \frac{\sum_{k=1}^{d} e^{\tilde{\theta}_k}}{\sum_{k=1}^{d} e^{\theta_k}}\right] + \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \left(\theta_l - \tilde{\theta}_l\right)\right]
$$

$$
= \mathbb{E}\left[\ln \frac{\sum_{k=1}^{d} e^{\tilde{\theta}_k}}{\sum_{k=1}^{d} e^{\theta_k}}\right] + \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \left(\theta_l - \tilde{\theta}_l\right)\right]
$$

$$
= \mathbb{E}\left[\ln \frac{\sum_{k=1}^{d} e^{\tilde{\theta}_k}}{\sum_{k=1}^{d} e^{\theta_k}}\right] + \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \left(\theta_l - \tilde{\theta}_l\right)\right]
$$

$$
\overset{(a)}{\leq} \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\tilde{\theta}_l}}{\sum_{k=1}^{d} e^{\tilde{\theta}_k}} \ln \frac{e^{\tilde{\theta}_l}}{e^{\theta_l}}\right] + \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \left(\theta_l - \tilde{\theta}_l\right)\right]
$$

$$
= \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\tilde{\theta}_l}}{\sum_{k=1}^{d} e^{\tilde{\theta}_k}} \left(\tilde{\theta}_l - \theta_l\right)\right] + \mathbb{E}\left[\sum_{l=1}^{d} \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}} \left(\theta_l - \tilde{\theta}_l\right)\right]
$$

$$
= \mathbb{E}\left[\sum_{l=1}^{d} \left(\frac{e^{\tilde{\theta}_l}}{\sum_{k=1}^{d} e^{\tilde{\theta}_k}} - \frac{e^{\theta_l}}{\sum_{k=1}^{d} e^{\theta_k}}\right) \left(\tilde{\theta}_l - \theta_l\right)\right]
$$

$$
\overset{(b)}{\leq} \mathbb{E}\left[\sum_{l=1}^{d} \left(\tilde{\theta}_l - \theta_l\right)^2\right]
$$

$$
= \mathbb{E}\left[\|\tilde{\theta} - \theta\|_2^2\right],
$$

where (a) follows from the log-sum inequality and (b) follows from the fact that the softmax function is 1-Lipschitz. $\square$

Lemma B.3. (meta rate-distortion upper bound) *For all $d, r, M, T \in \mathbb{Z}_{++}$, if for all $(m, t) \in [M] \times [T]$, $X_t^{(m)}$ is generated according to the linear representation learning process, then*

$$
\mathbb{H}_{M,T,\epsilon}(\psi) \leq \frac{dr}{2MT} \ln \left(1 + \frac{d}{r \left(e^{\frac{2\epsilon T}{r}} - 1\right)}\right).
$$

Proof. Let $\tilde{\psi} = \psi + Z$ where $Z \in \Re^{d \times r}$ is independent of ψ and consists of elements which are distributed iid $\mathcal{N}(0, \epsilon')$

where $\epsilon' = (e^{\frac{2\epsilon T}{r}} - 1)/d$ We begin by upper bounding the rate.

$$
\begin{aligned}
\frac{\mathbb{I}(\psi; \tilde{\psi})}{MT} &= \frac{\mathbf{h}(\tilde{\psi}) - \mathbf{h}(\tilde{\psi}|\psi)}{MT} \\
&\overset{(a)}{\leq} \frac{\frac{dr}{2} \ln\left(2\pi e \left(\epsilon' + \frac{1}{r}\right)\right) - \frac{dr}{2} \ln\left(2\pi e \epsilon'\right)}{MT} \\
&= \frac{dr \ln\left(1 + \frac{1}{r\epsilon'}\right)}{2MT} \\
&= \frac{dr}{2MT} \ln\left(1 + \frac{d}{r\left(e^{\frac{2\epsilon T}{r}} - 1\right)}\right),
\end{aligned}
$$

where (a) follows from the maximum differential entropy of a random variable of fixed variance being upper bounded by a Gaussian random variable.

We now upper bound the distortion.

$$
\begin{aligned}
\frac{\mathbb{I}(H_{M,T}; \psi|\tilde{\psi})}{MT} &= \frac{\mathbb{H}(H_{M,T}|\tilde{\psi}) - \mathbb{H}(H_{M,T}|\psi)}{MT} \\
&= \frac{\sum_{m=1}^{M} \mathbb{H}(X_{0:T}^{(m)}|\tilde{\psi}, H_{m-1,T}) - \mathbb{H}(X_{0:T}^{(m)}|\psi, H_{m-1,T})}{MT} \\
&\leq \frac{\mathbb{H}(X_{0:T}^{(1)}|\tilde{\psi}) - \mathbb{H}(X_{0:T}^{(1)}|\psi)}{T} \\
&= \frac{\mathbb{I}(X_{0:T}^{(1)}; \psi|\tilde{\psi})}{T} \\
&\leq \frac{\mathbb{I}(\theta_1; \psi|\tilde{\psi})}{T} \\
&= \frac{\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(\theta_1 \in \cdot|\psi)\|\mathbb{P}(\theta_1 \in \cdot|\tilde{\psi}))\right]}{n} \\
&\overset{(a)}{\leq} \frac{\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(\theta_1 \in \cdot|\psi)\|\mathbb{P}(\theta_\in \cdot|\psi \leftarrow \tilde{\psi}))\right]}{T} \\
&\leq \frac{\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\lim_{\delta \to 0}\mathbb{P}(\theta_\delta \in \cdot|\psi)\|\lim_{\delta \to 0}\mathbb{P}(\theta_\delta|\psi \leftarrow \tilde{\psi}))\right]}{T} \\
&\overset{(b)}{=} \frac{1}{T}\mathbb{E}\left[\lim_{\delta \to 0}\frac{1}{2}\ln\left(\frac{\left|\delta I_d + \frac{\tilde{\psi}\tilde{\psi}^\top}{r}\right|}{\left|\delta I_d + \frac{\psi\psi^\top}{r}\right|}\right) - d + \mathrm{Tr}\left(\left(\delta I_d + \frac{\tilde{\psi}\tilde{\psi}^\top}{r}\right)^{-1}\left(\delta I_d + \frac{\psi\psi^\top}{r}\right)\right)\right] \\
&\overset{(c)}{\leq} \frac{1}{T}\mathbb{E}\left[\lim_{\delta \to 0}\frac{1}{2}\ln\left(\frac{\left|\delta I_d + \frac{\tilde{\pi}\tilde{\psi}^\top}{r}\right|}{\left|\delta I_d + \frac{\psi\psi^\top}{r}\right|}\right)\right] \\
&\overset{(d)}{=} \frac{1}{T}\mathbb{E}\left[\lim_{\delta \to 0}\frac{1}{2}\ln\left(\frac{|\delta I_d| \cdot \left|I_r + \frac{\tilde{\psi}^\top\tilde{\psi}}{r\delta}\right|}{|\delta I_d| \cdot \left|I_r + \frac{\psi^\top\psi}{r\delta}\right|}\right)\right] \\
&= \frac{1}{T}\mathbb{E}\left[\lim_{\delta \to 0}\frac{1}{2}\ln\left(\frac{\left|I_r + \frac{\tilde{\psi}^\top\tilde{\psi}}{r\delta}\right|}{\left|I_r + \frac{I_r}{r\delta}\right|}\right)\right] \\
&\overset{(e)}{\leq} \lim_{\delta \to 0}\frac{1}{2T}\ln\left(\frac{\left|I_r + \frac{\mathbb{E}[\tilde{\psi}^\top\tilde{\psi}]}{r\delta}\right|}{\left|I_r + \frac{I_r}{r\delta}\right|}\right) \\
&= \lim_{\delta \to 0}\frac{1}{2T}\ln\left(\frac{\left|I_r + \frac{\mathbb{E}[I_r + d\epsilon' I_r]}{r\delta}\right|}{\left|I_r + \frac{I_r}{r\delta}\right|}\right) \\
&= \lim_{\delta \to 0}\frac{r}{2T}\ln\left(\frac{1 + \frac{1+d\epsilon'}{r\delta}}{1 + \frac{1}{r\delta}}\right) \\
&= \frac{r}{2T}\ln(1 + d\epsilon) \\
&= \epsilon,
\end{aligned}
$$

where (a), (b) follows from continuity of the KL-divergence between two multivariate normal distributions w.r.t the covariance matrix, (c) follows from the fact that the trace term is upper bounded by d, (d) follows from the matrix determinant lemma, $\epsilon = \frac{1}{m}$, and (e) follows from Jensen's inequality. The result follows. □

Lemma B.4. (intra-task rate-distortion upper bound) *For all $r, T \in \mathbb{Z}_{++}$,*

$$\mathbb{H}_{T,\epsilon}(\theta_m|\psi) \leq \frac{r}{2}\ln\left(1 + \frac{1}{\epsilon}\right).$$

Proof. Let $\tilde{\xi} = \xi^{(m)} + Z$ where $Z \perp \xi^{(m)}$ and $Z \sim \mathcal{N}(0, \epsilon I_r/r)$. Let $\tilde{\theta} = \psi\tilde{\xi}$. We begin by upper bounding the rate.

$$
\begin{aligned}
\mathbb{I}(\theta_m; \tilde{\theta}|\psi) &\leq \mathbb{I}(\xi; \tilde{\xi}|\psi) \\
&= \mathbf{h}(\tilde{\xi}|\psi) - \mathbf{h}(\tilde{\xi}|\psi, \xi) \\
&= \mathbf{h}(\tilde{\xi}) - \mathbf{h}(\tilde{\xi}|\xi) \\
&= \mathbf{h}(\tilde{\xi}) - \mathbf{h}(Z) \\
&= \frac{r}{2}\ln\left(2\pi e(\frac{\epsilon}{r} + \frac{1}{r})\right) - \frac{r}{2}\ln\left(\frac{2\pi e\epsilon}{r}\right) \\
&= \frac{r}{2}\ln\left(1 + \frac{1}{\epsilon}\right).
\end{aligned}
$$

We now upper bound the distortion:

$$
\begin{aligned}
\frac{\mathbb{I}(X_{0:T}^{(m)}; \theta_m|\tilde{\theta}, \psi)}{T} &\leq \frac{\mathbb{I}(X_{0:T}^{(m)}; \theta_m|\tilde{\theta}, \psi)}{T} \\
&\leq \mathbb{I}(X_1^{(m)}; \theta_m|\tilde{\theta}) \\
&= \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(X_1^{(m)} \in \cdot|\theta_m\right) \| \mathbb{P}\left(X_1^{(m)} \in \cdot|\tilde{\theta}\right)\right)\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(X_1^{(m)} \in \cdot|\theta_m\right) \| \mathbb{P}\left(X_1^{(m)} \in \cdot|\theta_m \leftarrow \tilde{\theta}\right)\right)\right] \\
&\overset{(b}{\leq} \mathbb{E}\left[\|\tilde{\theta} - \theta_m\|_2^2\right] \\
&= \mathbb{E}\left[\left(\xi^{(m)} - \tilde{\xi}\right)^{\top} \psi^{\top}\psi\left(\xi^{(m)} - \tilde{\xi}\right)\right] \\
&= \mathbb{E}\left[\left(\xi^{(m)} - \tilde{\xi}\right)^{\top}\left(\xi^{(m)} - \tilde{\xi}\right)\right] \\
&= \mathbb{E}\left[Z^{\top}Z\right] \\
&= \epsilon
\end{aligned}
$$

where (a) follows from Lemma A.1, and (b) follows from Lemma B.2. $\square$

Theorem B.1. (linear representation learning Bayesian error bound) *For all $d, r, M, T \in \mathbb{Z}_{++}$,*

$$\mathcal{L}_{M,T} \leq \frac{dr\left(1 + \ln\left(1 + \frac{M}{r}\right)\right)}{2MT} + \frac{r\left(1 + \ln(1 + \frac{2n}{r})\right)}{2T}.$$

Proof.

$$
\begin{aligned}
\mathcal{L}_{M,T} &\overset{(a)}{\leq} \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{M,T\epsilon}(\psi)}{MT} + \epsilon + \inf_{\epsilon' \geq 0} \frac{\mathbb{H}_{T,\epsilon'}(\theta_m|\psi)}{T} + \epsilon' \\
&\overset{(b)}{\leq} \inf_{\epsilon \geq 0} \frac{dr}{2MT} \ln\left(1 + \frac{d}{r\left(e^{\frac{2\epsilon T}{r}} - 1\right)}\right) + \epsilon + \inf_{\epsilon' \geq 0} \frac{r}{2n} \ln\left(1 + \frac{1}{r\epsilon'}\right) + \epsilon' \\
&\overset{(c)}{\leq} \frac{dr}{2MT} \ln\left(1 + \frac{d}{r\left(e^{\frac{d\epsilon}{M}} - 1\right)}\right) + \frac{dr}{2MT} + \frac{r\ln(1 + \frac{2T}{r})}{2T} + \frac{r}{2T} \\
&\leq \frac{dr \ln\left(1 + \frac{M}{r}\right)}{2MT} + \frac{dr}{2MT} + \frac{r\ln(1 + \frac{2T}{r})}{2T} + \frac{r}{2T} \\
&= \frac{dr \ln\left(e\left(1 + \frac{M}{r}\right)\right)}{2MT} + \frac{r\ln\left(e\left(1 + \frac{2T}{r}\right)\right)}{2T},
\end{aligned}
$$

where (a) follows from Theorem 4.3, (b) follows directly from Lemmas B.3 and B.4, and (c) follows from setting $\epsilon = \frac{dr}{2MT}$ and $\epsilon' = \frac{r}{2T}$. $\qquad\square$

B.2. Sparse Mixture of Transformers

Lemma B.5. (Dirichlet-multinomial concentration) *For all $K, m \in \mathbb{Z}_{++}$, let $(A, \bar{\theta}) \sim \mathcal{DP}(K, \mathrm{Unif}(\mathbb{S}^{d-1}))$. If for all $i \in [m]$, $c_i \overset{iid}{\sim} \mathrm{Categorical}(\bar{\theta})$ and N_m denotes the random variable which represents the number of* unique *categories draw in $(c_1, \ldots, c_m)$, then*

$$
\mathbb{E}[N_m] \leq K \ln\left(1 + \frac{n}{K}\right).
$$

To obtain the tightest bounds for this problem instance, instead of consider two separate rate-distortion functions as in Theorem 4.3, we must instead analyze the combined rate-distortion function defined below:

$$
\mathbb{H}_{M,T,\epsilon}(\psi, \theta_{1:M}) = \inf_{\tilde{\theta} \in \tilde{\Theta}_{M,T,\epsilon}} \mathbb{I}(\psi, \theta_{1:M}; \tilde{\theta}),
$$

where

$$
\tilde{\Theta}_{M,T,\epsilon} = \left\{\tilde{\theta} : \tilde{\theta} \perp H_{M,T}|\theta_{1:M}; \quad \frac{\mathbb{I}(H_{M,T}; \psi, \theta_{1:M}|\tilde{\theta})}{MT} \leq \epsilon\right\}.
$$

Under this rate-distortion function, we obtain the following bounds on estimation error:

Theorem B.6. (rate-distortion estimation error bound 2) *For all $M, T \in \mathbb{Z}_+$,*

$$
\mathcal{L}_{M,T} \leq \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{M,T,\epsilon}(\psi, \theta_{1:M})}{MT} + \epsilon,
$$

and

$$
\mathcal{L}_{M,T} \geq \sup_{\epsilon \geq 0} \min\left\{\frac{\mathbb{H}_{M,T,\epsilon}(\psi, \theta_{1:M})}{MT}, \epsilon\right\}.
$$

Proof. We begin by establishing the upper bound.

$$
\begin{aligned}
\mathcal{L}_{M,T} &= \frac{\mathbb{I}(H_{M,T}; \psi, \theta_{1:M})}{MT} \\
&= \inf_{\tilde{\theta} \in \tilde{\Theta}_{M,T,\epsilon}} \frac{\mathbb{I}(H_{M,T}; \psi, \theta_{1:M}, \tilde{\theta})}{MT} \\
&= \inf_{\tilde{\theta} \in \tilde{\Theta}_{M,T,\epsilon}} \frac{\mathbb{I}(H_{M,T}; \tilde{\theta})}{MT} + \frac{\mathbb{I}(H_{M,T}; \theta_{1:M}|\tilde{\theta})}{MT} \\
&\leq \frac{\mathbb{H}_{M,T,\epsilon}(\psi, \theta_{1:M})}{MT} + \epsilon.
\end{aligned}
$$

We now establish the lower bound. Suppose that $\mathbb{I}(H_{M,T}; \psi, \theta_{1:M}) < \mathbb{H}_{M,T,\epsilon}(\psi, \theta_{1:M})$. Let $\tilde{\theta} = \tilde{H}_{M,T} \notin \tilde{\Theta}_{M,T,\epsilon}$ where $\tilde{H}_{M,T}$ is another independent history sampled in the same manner as $H_{M,T}$.

$$
\begin{aligned}
\mathbb{I}(H_{M,T}; \psi, \theta_{1:M}) &= \mathbb{H}(H_{M,T}) - \mathbb{H}(H_{M,T}|\psi, \theta_{1:M}) \\
&\overset{(a)}{=} \mathbb{H}(H_{M,T}) - \mathbb{H}(H_{M,T}|\psi, \theta_{1:M}, \tilde{\theta}) \\
&\overset{(b)}{\geq} \mathbb{H}(H_{M,T}|\tilde{\theta}) - \mathbb{H}(H_{M,T}|\psi, \theta_{1:M}, \tilde{\theta}) \\
&= \mathbb{I}(H_{M,T}; \psi, \theta_{1:M}|\tilde{\theta}) \\
&\overset{(c)}{\geq} \epsilon MT
\end{aligned}
$$

where (a) follows from conditional independence assumptions, (b) follows from the fact that conditioning reduces entropy, and c) follows from the fact that $\tilde{\theta} \notin \tilde{\Theta}_{M,T,\epsilon}$. Therefore, for all $\epsilon \geq 0$, $\mathbb{I}(H_{M,T}; \psi, \theta_{1:M}) \geq \min\{\mathbb{H}_{M,T,\epsilon}(\psi, \theta_{1:M}), \epsilon MT\}$. The result follows. $\qquad\square$

With this rate-distortion bound in place, we established that in order to bound the estimation error of ICL, we simply must bound the rate-distortion function. In the following result, we establish an upper bound on the rate-distortion function for the ICL problem.

Lemma B.7. (in-context learning rate-distortion upper bound) *For all d, r, K, L, M, T, if for all $(m, t) \in [M] \times [T]$, $X_t^{(m)}$ is generated by the ICL process, then*

$$
\mathbb{H}_{M,T,\epsilon}(\psi, \theta_{1:M}) \leq M \ln(N) + R \ln\left(1 + \frac{M}{R}\right) \cdot r \cdot \max\{r, d\} L \ln\left(1 + \frac{r \cdot \max\{r, d\} KLT(8K(1 + 16K))^L}{\epsilon}\right).
$$

Proof. For the ICL data generating process, we will construct a compression which incrementally builds up elements of the mixture. Concretely, let $\tilde{\theta} = (i_1, i_2, \ldots, i_M, \tilde{\psi})$ where $\psi = \{\tilde{\psi}_j : j \in \{i_1, i_2, \ldots, i_M\}\}$ and $\tilde{\psi}_j$ represents a noisy version of ψ_j. We will use $\mathcal{I}_M$ to denote the set of *unique* indices from $\{i_1, i_2, \ldots, i_M\}$.

Let $\tilde{\psi}_j = (\tilde{V}_{j,\ell}, \tilde{A}_{j,\ell} : \ell \in [L])$ where

$$
\left(\tilde{V}_{j,\ell}, \tilde{A}_{j,\ell}\right) = (V_{j,\ell} + Z_{j,\ell}, \ A_{j,\ell} + B_{j,\ell})
$$

$$
(Z_{j,\ell}, B_{j,\ell}) \perp (V_{j,\ell}, A_{j,\ell}), \quad Z_{j,\ell} \overset{iid}{\sim} \begin{cases} \mathcal{N}(0, \epsilon'/r^2) & \text{if } \ell < L \\ \mathcal{N}(0, \epsilon'/(rd)) & \text{if } \ell = L \end{cases}; \quad B_{j,\ell} \overset{iid}{\sim} \mathcal{N}(0, \epsilon'/r^2),
$$

and $\epsilon' = \frac{\epsilon}{KLT(8K(1+16K))^L}$.

We begin by upper bounding the rate.

$$
\begin{aligned}
\mathbb{I}(\psi, \theta_{1:M}; \tilde{\theta}) &= \mathbb{I}(\psi, \theta_{1:M}; (i_1, \ldots, i_M), \tilde{\psi}) \\
&= \mathbb{I}(\psi, (i_1, \ldots, i_M); (i_1, \ldots, i_M), \tilde{\psi}) \\
&= \mathbb{H}(i_1, \ldots, i_M) + \mathbb{I}(\psi; \tilde{\psi}|(i_1, \ldots, i_M)) \\
&\leq M \ln(N) + \mathbb{E}\left[\sum_{i=1}^{N} \mathbb{1}_{[i \in \mathcal{I}_M]} \cdot \mathbb{I}(\psi_i; \tilde{\psi}_i)\right] \\
&\overset{(a)}{\leq} M \ln(N) + R \ln\left(1 + \frac{M}{R}\right) \cdot r \cdot \max\{r, d\} L \ln\left(1 + \frac{r \cdot \max\{r, d\} KLT(8K(1 + 16K))^L}{\epsilon}\right),
\end{aligned}
$$

where (a) follows from Lemmas B.5 and A.8.

We now upper bound the distortion:

$$
\begin{aligned}
\frac{\mathbb{I}(H_{M,T};\psi,\theta_{1:M}|\tilde{\theta})}{MT} &= \frac{\mathbb{I}(H_{M,T};\theta_{1:M}|\tilde{\theta})}{MT} \\
&= \sum_{m=1}^{M} \frac{\mathbb{I}(X_{0:T}^{(m)};\theta_{1:M}|\tilde{\theta},H_{m-1,T})}{MT} \\
&\leq \frac{\mathbb{I}(X_{0:T}^{(1)};\theta_{1:M}|\tilde{\theta})}{T} \\
&= \frac{\mathbb{I}(X_{0:T}^{(1)};\theta_1|\tilde{\psi}_{i_1})}{T} \\
&\overset{(a)}{\leq} \epsilon,
\end{aligned}
$$

where (a) follows from Lemma A.8. $\qquad\square$

Theorem 4.5. (estimation error bound) *For all $d,r,K,L,M,N,R,T \in \mathbb{Z}_{++}$, if for all $(m,t) \in [M] \times [T]$, $X_t^{(m)}$ is generated according to the sparse mixture of transformers environment, then*

$$
\begin{aligned}
\mathcal{L}_{M,T} \leq\ & \frac{pRL\ln\left(1+\frac{M}{R}\right)\ln(136eK^2)}{MT} \\
& + \frac{pR\ln\left(1+\frac{M}{R}\right)\ln\left(\frac{2KMT^2}{L}\right)}{MT} \\
& + \frac{\ln(N)}{T},
\end{aligned}
$$

where $p = 2r^2(L-1) + (dr+r^2)$ denotes the parameter count of each transformer in the mixture.

Proof.

$$
\begin{aligned}
\mathcal{L}_{M,T} \leq\ & \inf_{\epsilon \geq 0} \frac{\mathbb{H}_{M,T,\epsilon}(\psi,\theta_{1:M})}{MT} + \epsilon \\
\leq\ & \inf_{\epsilon \geq 0} \frac{\ln(N)}{T} + \frac{r\max\{r,d\}RL\ln\left(1+\frac{M}{R}\right)\ln\left(1+\frac{r\max\{r,d\}KLT(8K(1+16K))^L}{\epsilon}\right)}{MT} + \epsilon \\
\leq\ & \frac{\ln(N)}{T} + \frac{r\max\{r,d\}RL\ln\left(1+\frac{M}{R}\right)\ln\left(1+\frac{KMT^2(8K(1+16K))^L}{L}\right)}{MT} + \frac{r\max\{r,d\}RL^2}{MT} \\
\leq\ & \frac{\ln(N)}{T} + \frac{r\max\{r,d\}RL\ln\left(1+\frac{M}{R}\right)\ln\left(\frac{2KMT^2(8K(1+16K))^L}{L}\right)}{MT} + \frac{r\max\{r,d\}RL^2}{MT} \\
\leq\ & \frac{\ln(N)}{T} + \frac{r\max\{r,d\}RL^2\ln\left(1+\frac{M}{R}\right)\ln(8Ke(1+16K))}{MT} + \frac{r\max\{r,d\}RL\ln\left(1+\frac{M}{R}\right)\ln\left(\frac{2KMT^2}{L}\right)}{MT}
\end{aligned}
$$

$\qquad\square$

B.3. In-context Learning

Theorem 4.7. (in context learning error bound) *For all $M,T,\tau \in \mathbb{Z}_{++}$, if $\tau \leq T$, then*

$$
\mathcal{L}_{M,T,\tau} \leq \underbrace{\frac{\mathbb{I}(H_{M,T};\psi)}{M\tau}}_{\substack{meta \\ estimation \\ error}} + \underbrace{\frac{\mathbb{I}(D_{M+1};\theta_{M+1}|\psi)}{\tau}}_{\substack{in\text{-}context \\ estimation \\ error}}.
$$

Proof.

$$\mathbb{L}_{M,T,\tau} = \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}\left[-\log \mathbb{P}(X_{t+1}^{(M+1)}|H_{M+1,t}) \right]$$

$$= \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}\left[\log \frac{1}{\mathbb{P}(X_{t+1}^{(M+1)}|\theta_{M+1}, H_{M+1,t})} + \mathbf{d}_{\mathrm{KL}}(\mathbb{P}(X_{t+1}^{(M+1)} \in \cdot|\theta_{M+1}, H_{M+1,t})\|\mathbb{P}(X_{t+1}^{(M+1)} \in \cdot|H_{M+1,t})) \right]$$

$$= \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{H}(X_{t+1}^{(M+1)}|\theta_{M+1}, H_{M+1,t}) + \mathbb{I}(X_{t+1}^{(M+1)}; \theta_{M+1}|H_{M+1,t})$$

$$= \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{H}(X_{t+1}^{(M+1)}|\theta_{M+1}, X_1^{(M+1)}, \ldots, X_t^{(M+1)}) + \mathbb{I}(X_{t+1}^{(M+1)}; \theta_{M+1}, \psi|H_{M+1,t})$$

$$\overset{(a)}{=} \frac{\mathbb{H}(D_{M+1}|\theta_{M+1})}{\tau} + \frac{\mathbb{I}(H_{M+1,\tau}; \theta_{M+1}, \psi|H_{M+1,0})}{\tau}$$

$$\overset{(b)}{=} \frac{\mathbb{H}(D_{M+1}|\theta_{M+1})}{\tau} + \frac{\mathbb{I}(H_{M+1,\tau}; \psi|H_{M+1,0})}{\tau} + \frac{\mathbb{I}(H_{M+1,\tau}; \theta_{M+1}|\psi, H_{M+1,0})}{\tau}$$

$$\overset{(c)}{\leq} \frac{\mathbb{H}(D_{M+1}|\theta_{M+1})}{\tau} + \frac{\mathbb{I}(H_{M+1,T}; \psi|H_{M+1,0})}{\tau} + \frac{\mathbb{I}(D_{M+1}; \theta_{M+1}|\psi)}{\tau}$$

$$\overset{(d)}{\leq} \frac{\mathbb{H}(D_{M+1}|\theta_{M+1})}{\tau} + \frac{\mathbb{I}(H_{M+1,T}; \psi)}{(M+1)\tau} + \frac{\mathbb{I}(D_{M+1}; \theta_{M+1}|\psi)}{\tau},$$

where (a) and (b) follow from the chain rule of mutual information, (c) follows from the fact that $\psi \perp H_{M+1,\tau}|H_{M+1,T}$ for $\tau \leq T$ and the data processing inequality, and (d) follows from the fact that for all m, $\mathbb{I}(H_{m+1,T}; \psi|H_{m,T}) \leq \mathbb{I}(H_{m,T}; \psi|H_{m-1,T})$ and the chain rule of mutual information. $\square$

C. Analysis of Suboptimal Meta-Learning Algorithms

All of the prior results bound the error incurred by the optimal algorithm which produces a prediction of the next token conditioned on the entire past sequence. In this section, we will derive some simple results which pertain to *suboptimal* algorithms.

The following result quantifies the shortfall incurred by an algorithm which produces an arbitrary prediction $\tilde{P}_{m,t}$ which may depend on the history $H_{m,t}$.

Lemma C.1. (loss of an arbitrary predictor) *For all $M, T \in \mathbb{Z}_{++}$, if for all $(m,t) \in [M] \times [T]$, $\tilde{P}_{m,t}$ is a predictive distribution which may depend on the previous data $H_{m,t}$ and $\tilde{\mathcal{L}}_{M,T}$ denotes its cumulative average log-loss, then*

$$\tilde{\mathbb{L}}_{M,T} = \mathbb{L}_{M,T} + \underbrace{\frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{P}_{m,t} \,\|\, \tilde{P}_{m,t} \right) \right]}_{\text{misspecification error}}.$$

Note that because KL divergence is always non-negative and $\mathcal{L}_{m,n}$ is the loss of the Bayesian posterior estimator $\hat{P}$, any prediction other than $\hat{P}$ will incur nonzero misspecification error.

For a particular class of predictors $\tilde{P}$, we can retrieve the following upper bound on the *misspecification error*. We consider predictors which perform Bayesian inference with respect to an incorrectly specified prior distribution $\tilde{P}_0$.

Theorem C.2. (misspecified prior error bound) *For all $M, T \in \mathbb{Z}_{++}$ and $m, t \in [M] \times [T]$, if $\tilde{P}_{m,t}$ is the Bayesian posterior under the prior $\tilde{P}_0(\psi)$, then*

$$\frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{P}_{m,t} \,\|\, \tilde{P}_{m,t} \right) \right] \leq \frac{\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(\psi \in \cdot)\|\tilde{P}_0(\psi \in \cdot) \right) \right]}{MT}.$$

Proof.

$$\frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{P}_{m,t} \,\|\, \tilde{P}_{m,t}\right)\right]$$

$$\stackrel{(a)}{=} \frac{1}{MT} \sum_{m=1}^{M} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(H_T^{(m)} \in \cdot) \,\|\, \tilde{P}_m\left(H_T^{(m)} \in \cdot\right)\right)\right]$$

$$\stackrel{(b)}{=} \frac{\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\mathbb{P}(H_{M,T} \in \cdot) \,\|\, \tilde{P}(H_{M,T} \in \cdot))\right]}{MT}$$

$$\stackrel{(c)}{\leq} \frac{\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(\psi \in \cdot) \,\|\, \tilde{P}_0(\psi \in \cdot)\right)\right]}{MT},$$

where (a) and (b) follow from the chain rule of KL divergence and (c) follows from the data processing inequality of KL Divergence. $\qquad \square$

Theorem C.2 suggests that so long as the KL divergence between prior distributions is finite, the misspecification error should decrease to 0 as M and $T \to \infty$. This can be ensured so long as the algorithm's prior $\tilde{P}_0(\psi)$ does not assign 0 probability mass to any set for which the environment prior $\mathbb{P}(\psi)$ assigns non-zero probability.

With these results in place, we provide the following Corollary which exactly characterizes the loss of a predictor $\tilde{P}$ which produces predictions via Bayesian inference with respect to a arbitrary prior distribution $\tilde{P}_0(\psi \in \cdot)$.

Corollary C.3. *For all $M, T \in \mathbb{Z}_{++}$ and $m, t \in [M] \times [T]$, if $\tilde{P}_{m,t}$ computes probabilities under an arbitrary prior distribution $\tilde{P}_0(\psi \in \cdot)$ and $\tilde{\mathcal{L}}_{M,T}$ denotes its cumulative average log-loss, then*

$$\tilde{\mathbb{L}}_{M,T} = \frac{\mathbb{H}(H_{M,T}|\theta_{1:M})}{MT} + \frac{\mathbb{I}(H_{M,T}; \psi)}{MT} + \frac{\mathbb{I}(D_m; \theta_m|\psi)}{T}$$
$$+ \frac{\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}(H_{M,T} \in \cdot) \,\|\, \tilde{P}(H_{M,T} \in \cdot)\right)\right]}{MT}.$$