Notes on interpreting the estimation error

Daniel Strauss

December 12, 2024

1 Motivation

1.1 Introduction to the concept

In their study, [JLLVR24] aimed to find an information-theoretic explanation for why in-context learning works so effectively. To carry out their analysis, they made specific assumptions regarding the distribution of the text data. They assumed that all documents, which they wanted transformers to learn, were generated by a sparse mixture of transformers, with some uncertainty in their parameters.

They then proposed a loss function of how well a model could express the distribution of a given sequence of tokens, see Definition 2.13. To evaluate the difficulty of learning the distribution of sequential data, they used the loss of the optimal Bayesian estimator. The optimal Bayesian estimator is the probability distribution for the next token, given the Bayesian prior and the previous sequence of tokens. The loss of the optimal Bayesian estimator turned out to be equal to the entropy per token of the sequence, see Theorem 2.1.

The authors decomposed the optimal loss into two components: the irreducible error (see definition 2.16) and the estimation error (see definition 2.15). The estimation error represents the error of learning the parameters, scaled by their necessity for reproducing the training data distribution. The irreducible error was that part of the error that would persist even if all parameters were known. One could see the estimation error as a result of randomness caused by uncertainty about the data-generating model and the irreducible error as a result of randomness that is "wanted" by the model.

The rest of their work focused on the big part of making statements about the estimation error in different scenarios and interpreting the results about the estimation error to make statements about the distribution of the random sequence, e.g., how easy it is to learn this sequence given certain scenarios. This analysis led them to propose an explanation/point of view on why in-context learning works well (see section 4.6 of [JLLVR24]).

1.2 Limitation of the estimation error

In this work, we highlight a limitation of using the estimation error to make inferences about a random sequence. Specifically, we demonstrate that the same sequence can be modeled to exhibit any estimation error between 0 and the entropy of the sequence, depending solely on the choice of the Bayesian prior. This implies that the estimation error can be made arbitrarily large or small without changing the underlying distribution, suggesting it is not a reliable indicator of the sequence's learnability.

Despite this issue, we believe the work by [JLLVR24] might still hold value. If a sequence has a specific estimation error given that it was generated by a particular transformer architecture, with defined uncertainty about the tokens, further analysis could potentially offer insights—not necessarily into how well the optimal Bayesian estimator learns this distribution—but rather into how effectively a transformer of this specific or similar architecture might learn the distribution.

Another way in which this segregation might be usefull, would be by using a good argument on why a specific bayesian prior Θ was choosen. Basically the idea of the estiamtion error seems to be evolving around that in some way Θ can be learned by watching the sequence and $H_T|\Theta$ can not be learned any better by watching the sequence. If the definition of the estimation error, includes, that the bayiesian prior has to fullfill this criterion, then it can not be choosen arbitrarily anymore. This would result in that the proofs of this script not beeing valid anymore.

1.3 Summary of motivation

The estimation error got extra care in the analysis of [JLLVR24], as it conveys how hard it is to guess the model's parameters that generated this distribution. Therefore, it seems to give a good representation of how hard it is to learn the resulting distribution.

As we will argue in this script, the estimation error is fully dependent on the Bayesian prior model and independent of the resulting distribution. E.g. the same sequence could be generated by different priors, each leading to a different estimation error, despite the underlying distribution of the sequence remaining the same. This shows that the estimation error can be manipulated through the choice of prior and is not necessarily indicative of the complexity of the sequence.

In Section 1.2, we described how in our opinion, the estimation error still may hold value.

2 Initial Definitions

2.1 General definitions

Definition 2.1. An alphabet is a finite non-empty set of tokens.

Definition 2.2. A word over an alphabet is a tuple containing just tokens of that alphabet.

Definition 2.3. Σ^∞ is the set of infinite words over the alphabet Σ.

Definition 2.4. Σ^* is the set of finite words over the alphabet Σ.

Definition 2.5. Σ^n is the set of words over the alphabet Σ of length n.

Definition 2.6. By $X \sim \mathbb{R}^n$, we denote that X is a random vector in $\mathbb{R}^n$.

2.2 Definitions related to this script

Definition 2.7. $\mathcal{P}_\Sigma$ is the set of probability distribution functions of a random variable that takes a value of Σ. More formaly $\mathcal{P}_\Sigma = \{f \in \{\Sigma \to [0,1]\} | \sum_{i\in\Sigma} f(i) = 1\}$.

Definition 2.8. $\mathcal{P}^*_\Sigma$ is the set of functions that, take in a word w over the alphabet Σ and return a probability distribution function of a random variable that takes a value of Σ. $\mathcal{P}^*_\Sigma := \{\Sigma^* \to \mathcal{P}_\Sigma\}$. An element of P^*_Σ is called a parametrized autoregressive model (PAM).

Remark 2.1. Why we call a parametrized autoregressive model such will become, more clear after defining autoregressive models.

Definition 2.9. Let $\{X_t\}_{t\in\mathbb{N}}$ be a sequence, then

- $X_{i:j} := (X_i, X_{i+1}, ..., X_j)$

- $H_T := X_{1:T}$
- $H_0 := ()$

Definition 2.10. An autoregressive sequence $\{X_t\}_{t\in\mathbb{N}}$ is called an autoregressive sequence on $\mathcal{P} \in \mathcal{P}^*_\Sigma$ if $\forall_{T\in\mathbb{N}}.\mathbb{P}[H_{T-1} = (x_1, ..., x_{T-1})] > 0 \implies \mathbb{P}[X_T = x_T|H_{T-1} = (x_1, ..., x_{T-1})] = \mathcal{P}(x_T)(x_1, ..., x_{T-1})$. We denote this by $\{X_t\}_{t\in\mathbb{N}} \sim \mathcal{P}$. We also say $\mathcal{P}$ represents $\{X_t\}_{t\in\mathbb{N}}$.

Remark 2.2. Note that for every random sequence $\{X_t\}_{t\in\mathbb{N}}$ over an alphabet, there is a PAM $\mathcal{P}$ representing $\{X_t\}_{t\in\mathbb{N}}$.

Furthermore if $f, g \in \mathcal{P}^*_\Sigma$ and $\{X_t\}_{t\in\mathbb{N}} \sim f$ and $\{Y_t\}_{t\in\mathbb{N}} \sim g$, have different distributions, then $f \neq g$.

The reverse scenario is not necessarily true. Consider a case where $\forall_T \mathbb{P}[H_T = w] = 0$. Then both PAMs with $f(w)(i) = 1$ and $g(w)(i) = .5$ could represent $\{X_t\}_{t\in\mathbb{N}}$.

Definition 2.11. An Autoregressive Model (AM) is an element of the set $\{\mathbb{R}^n \ \rightarrow \mathcal{P}^*_\Sigma\}$, where $n \in \mathbb{N}$.

Remark 2.3. Be aware: other people use the word Autoregressive Model for Parametrized Autoregressive Models and some other people use the word Autoregressive Model for something even more different.

Let's look at what is an AM A. I hope I didn't make it look more complicated than it is. An AM is a function that - by definition - maps parameters to PAMs. A PAM is nothing more than an expression of a probability distribution of a sequence of tokens. For example, a transformer architecture can be seen as an AM and if you give the parameters of this architecture values, then you obtain a transformer, which can be seen as a PAM.

To be more specific: Let A be an AM, then $A(\theta)$ is a PAM, $A(\theta)(x_1, ..., x_{t-1})$ represents a probability distribution function for a random variable that takes an element of the alphabet and $A(\theta)(x_1, ..., x_{t-1})(x_t)$ represents a probability.

2.3 Definitions and results from [JLLVR24]

Definition 2.12. Let π denote an estimator, for a probability distribution of the next variable X_{t+1} given the previous random sequence of tokens H_t. π can be dependent on the Bayesian prior distribution but not on the Bayesian prior itself.

Definition 2.13. The average cumulative expected log-loss: $\mathbb{L}_T(\pi) := \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[-\log(\pi(H_t, X_{t+1}))]$

Definition 2.14. $\mathbb{L}_T := \min_\pi \mathbb{L}_T(\pi)$

Theorem 2.1. *[JLLVR24]* $\mathbb{L}_T T = \mathbb{I}(H_T; \theta) + \mathbb{H}(H_T|\theta)$

Definition 2.15. $\frac{\mathbb{I}(H_T;\theta)}{T}$ is called the estimation error.

Definition 2.16. $\frac{\mathbb{H}(H_T|\theta)}{T}$ is called the irreducible error.

3 Results

Lemma 3.1. *There is an AM A_0 over Σ, such that for every random sequence $\{X_t\}_{t\in\mathbb{N}}$ on Σ, there is $\Theta_0 \in R_n$, such that $\{X_t\}_{t\in\mathbb{N}} \sim A_0(\Theta_0)$, with the estimation error $\mathbb{I}(H_T;\Theta_1)/T = 0$.*

Idea of the proof: We can easily show the existence of an AM A that can represent any random sequence on Σ, with deterministic parameters. The lemma follows immediately, as the determenistic parameters have the entropy of zero.

Proof.

1. To show the existence of a surjective AM it suffices to show that $|\mathbb{R}^n| \geq |\mathcal{P}^*_\Sigma|$.

$$|\mathcal{P}^*_\Sigma| \leq |\{\Sigma^* \to \{\Sigma \to [0,1]\}\}| = |\{\Sigma^* \to \{\Sigma \to \mathbb{R}\}\}| = |\{\Sigma^* \to \mathbb{R}\}| \leq |\{\mathbb{N} \to \mathbb{R}\}| = |\mathbb{R}|$$

2. Let A_0 be a surjective AM.

3. Consider this function $f : \Sigma^* \to \mathcal{P}_\Sigma$:

$$f : w \mapsto (x \mapsto \mathbb{P}[X_{|w|} = x | H_{|w|-1} = w]) \tag{1}$$

4. $f \in \mathcal{P}^*_\Sigma$ (f is a PAM). (By definition of f)

5. $\{X_t\}_{t\in\mathbb{N}} \sim f$. (By construction of f and by definition 2.10)

6. $\exists_{\Theta_0 \in \mathbb{R}^n} . A_0(\Theta_0) = f$. (From A_0 surjective)

7. $\{X_t\}_{t\in\mathbb{N}} \sim A_0(\Theta_0)$

8. If we model $\{X_t\}_{t\in\mathbb{N}} \sim A_0(\Theta_0)$, then $\mathbb{H}(\Theta_0) = 0$ (since $\Theta_0 \in \mathbb{R}^n$) and therefore we obtain an estimation error of 0: $\mathbb{I}(H_T;\Theta_0) = 0$.

□

Definition 3.1. We call an AM A_0, that satisfies the condition of lemma 3.1, a distributionaly full AM.

Lemma 3.2. *There is an AM A_1 over Σ, such that for every random sequence $\{X_t\}_{t\in\mathbb{N}}$ on Σ, there is $\Theta_1 \sim R_n$, such that $\{X_t\}_{t\in\mathbb{N}} \sim A_1(\Theta_1)$, with the estimation error $\mathbb{I}(H_T;\Theta_1)/T = \mathbb{H}(H_T)/T$.*

Idea of the proof: Suppose you have an AM that always represents a deterministic sequence if parametrized with constant parameters. Note that this AM parameterized by $\Theta \sim \mathbb{R}^n$ will always have the estimation error of $\mathbb{H}(H_T)/T$. Since it creates only deterministic probability distributions, all randomness is contained in Θ.

All that is left is to construct such an AM with the additional feature that it can be parametrized (with random parameters) to represent any random sequence.

Proof.

1. Let $f : \Sigma^\infty \to \mathbb{R}^n$ be a bijection.
2. We define [1]: $A_1(\theta) := h \mapsto (x \mapsto \mathbb{P}[f^{-1}\ (\theta)_{|h|+1} = x | f^{-1}(\theta)_{1:|h|} = h])$
    3. Note that A_1 is a (not well-defined) AM
4. Now for a given autoregressive sequence $\{X_t\}_{t\in\mathbb{N}}$, we define the random vector $\Theta_1 = f(\{X_t\}_{t\in\mathbb{N}})$.
    5. Then

    $$A_1(\Theta_1)(h_t)(x) \overset{a)}{=} \mathbb{P}(f^{-1}(\Theta_1)_{t+1} = x | f^{-1}(\Theta_1)_{1:t} = h_t) \overset{b)}{=} \mathbb{P}[X_{t+1} = x | H_t = h_t]$$

    . (a) from Def. A_1, b) from Def. Θ_1)

    6. $\{X_t\}_{t\in\mathbb{N}} \sim A_1(\Theta_1)$. (from 5)
7. $\mathbb{H}[H_T|\Theta_1] \overset{a)}{=} \mathbb{H}[f^{-1}(\Theta_1)|\Theta_1] \overset{b)}{=} \mathbb{H}[\Theta_1|\Theta_1] = 0.$ (a) Def. Θ_1, b) f is deterministic and bijective)
8. $\mathbb{H}(H_T) = \mathbb{H}[H_T|\Theta_1] + \mathbb{I}(H_T;\Theta_1) = \mathbb{I}(H_T;\Theta_1)$

□

Definition 3.2. We call an AM A_1, which satisfies the condition of Lemma 3.2, a deterministically full AM.

Lemma 3.3.

- *Let X be a discrete random variable, Y be a random variable and A an random event, (such that $\mathbb{P}[A] \in (0,1)$.*
- *(I:) Let for all $x \in \mathcal{X}, y \in \mathcal{Y}$ hold that $f_{Y|A}(y) > 0 \implies \mathbb{P}[X = x|Y = y] = \mathbb{P}[X = x|Y = y, A]$.*
- *(II:) Let for all $x \in \mathcal{X}, y \in \mathcal{Y}$ hold that $f_{Y|\overline{A}}(y) > 0 \implies \mathbb{P}[X = x|Y = y] = \mathbb{P}[X = x|Y = y, \overline{A}]$.*

Then $\mathbb{H}[X|Y] = \mathbb{P}[A]\mathbb{H}[X|Y, A] + (1 - \mathbb{P}[A])\mathbb{H}[X|Y, \overline{A}]$

Proof.

1. $\mathbb{H}[X|Y] = \mathbb{E}[y \mapsto \mathbb{H}[X, Y = y](Y)]$
2. $$= \mathbb{P}[A]\mathbb{E}[y \mapsto \mathbb{H}[X, Y = y](Y)|A] + (1 - \mathbb{P}[A])\mathbb{E}[y \mapsto \mathbb{H}[X, Y = y](Y)|\overline{A}]$$

    (from law of total expectation)
3. We examine $\mathbb{E}[y \mapsto \mathbb{H}[X, Y = y](Y)|J]$ for the cases $J = A$ and $J = \overline{A}$
    4. $\mathbb{E}[y \mapsto \mathbb{H}[X, Y = y](Y)|J]$ (value is defined since $\mathbb{P}[J] > 0$)
    5. $$= \int_{\mathcal{Y}} f_{Y|J}(y)\mathbb{E}[x \mapsto -\log(\mathbb{P}[X = x|Y = y]|Y)(X)]\, dy$$

    (by writing out the expectation and the entropy)

[1] I am sorry that this expression looks confusing, but I will try to explain: $f^{-1}(\theta)$ is just an infinite long word. $A(\theta)(h)(x)$ returns 1 iff the word $h \circ x$ is a prefix of $f^{-1}(\theta)$. If h is a prefix of $f^{-1}(\theta)$, but $h \circ x$, then $A(\theta)(h)(x) = 0$ and otherwise $A(\theta)$ is undefined. A will start to make more sense once θ is a random variable.

6.
$$= \int_{\mathcal{Y}} f_{Y|J}(y)\mathbb{E}[x \to -\log(\mathbb{P}[X = x|Y = y, J]|Y)(X)]\, dy$$
(by I if $J = A$ and by II if $J = \overline{A}$)

7.
$$= \int_{\mathcal{Y}} f_{Y|J}(y) \sum_{x \in \mathcal{X}} [-\log(\mathbb{P}[X = x|Y = y, J])\mathbb{P}[X = x|Y = y]]\, dy$$

8.
$$= \int_{\mathcal{Y}} f_{Y|J}(y) \sum_{x \in \mathcal{X}} [-\log(\mathbb{P}[X = x|Y = y, A])\mathbb{P}[X = x|Y = y, J]]\, dy$$
(by I if $J = A$ and by II if $J = \overline{A}$)

9.
$$= \int_{\mathcal{Y}} f_{Y|J}(y)\mathbb{H}[X = x|Y = y, J]\, dy$$

10. $= \mathbb{H}[X|Y, J]$

11. We can now insert the reult from 4 to 10 into 2

12. $= \mathbb{P}[A] \cdot \mathbb{H}[X = x|Y, A] + (1 - \mathbb{P}[A]) \cdot \mathbb{H}[X = x|Y, \overline{A}]$

□

Comments on this Lemma: Possibly the condition in the lemma could me much shorter and more intuitive. This lemma could require less text to apply if one shows $I \implies II$ (I attached a proof sketch to the end to the document (dont read it, it is messy). On request I can formalize it to see if it is actually true). Also it might be possible that I and II are almost equivalent with X and A are independent given Y. I put the condition of the lemma like this, because I think in total we require less text by doing so. Also where we are going to use the lemma, the condition will be easy to check. But it might also be possible, that everything looks cleaner if we use independence.

Theorem 3.4. *There is an AM A_c, such that given any $\alpha \in [0, 1]$ and any random sequence $\{X_t\}_{t\in\mathbb{N}}$ there is $\Theta_\alpha \sim \mathbb{R}^n$, such that $\{X_t\}_{t\in\mathbb{N}} \sim A_c(\Theta_\alpha)$, with the estiamtion error $\mathbb{I}(H_T; \Theta_1)/T = \alpha \cdot \mathbb{H}(H_T)/T$.*

Idea of the proof: We know there is A_0 and A_1, such that of any sequence $\{X_t\}_{t\in\mathbb{N}}$ on the same alphabet, there is Θ_0 and Θ_1, such that $X \sim A_0(\Theta_0)$ with minimal estimation error and $X \sim A_1(\Theta_1)$ with maximal estimation error. Let us suppose a random variable Θ_c that is with some probability a Θ_1 and otherwise Θ_0 and, which also "stores" which of the two variables it is. We combine A_0 and A_1 into an AM A_c, such that A_c is an AM that "chooses" whether to use A_0 or A_1, depending on $\Theta_\alpha = \Theta_0$ or $\Theta_\alpha = \Theta_1$. Then $X \sim A_c(\Theta_c)$. All that is left is to see whether we can choose a such that $\mathbb{I}(H_T; \Theta_c) = \alpha \cdot \mathbb{H}(H_T)$. (Spoiler alert: $a = \alpha$)

Proof.

1. Let A_0 be a distributionaly full AM and let A_1 a deterministicaly full AM.

2. Let $\Theta_0 \in \mathbb{R}^n$, such that $X \sim A_0(\Theta_0)$ and $\mathbb{I}(H_T; \Theta_0) = 0$.

    3. Θ_0 exists (from Lemma 3.1)

4. Let $\Theta_1 \sim \mathbb{R}^n$, such that $X \sim A_1(\Theta_1)$ and $\mathbb{I}(H_T; \Theta_1) = \mathbb{H}(H_T)$.

    5. Θ_1 exists (from Lemma 3.2)

6. From the existance of Θ_0, A_0 and Θ_1, A_1, we can assume w.l.o.g. that $\alpha \in (0, 1)$.

7. Let B be a binary random variable with $\mathbb{P}(B=1)=\alpha$ and $\mathbb{P}(B=0)=1-\alpha$.
8. Let $\Theta_\alpha \sim \mathbb{R}^{n+1}$, with $(\Theta_\alpha)_{n+1} := B$ and $(\Theta_\alpha)_{1:n} := (1-B)\cdot\Theta_0 + B\cdot\Theta_1$.
9. We define AM $A_c : \mathbb{R}^{n+1} \to \mathcal{P}^*_\Sigma$:

$$A_c(\Theta) := A_0((\Theta)_{0:n})\cdot(1-(\Theta)_{n+1}) + A_1((\Theta)_{0:n})\cdot(\Theta)_{n+1}$$

   10. given $B=1$: $A_c(\Theta) = A_1((\Theta)_{0:n}) = A_1(\Theta_1) \implies X \sim A_c(\Theta_\alpha)$
   11. given $B=0$: $A_c(\Theta) = A_0((\Theta)_{0:n}) = A_0(\Theta_0) \implies X \sim A_c(\Theta_\alpha)$
   12. $\mathbb{P}[B \notin \{0,1\}] = 0$
   13. $X \sim A_c(\Theta_\alpha)$ (from 10, 11, 12)
14. In the next steps we evaluate: $\mathbb{I}(H_T;\Theta_\alpha) = \mathbb{H}[H_T] - \mathbb{H}[H_T|\Theta_\alpha]$
   15. We simplify $\mathbb{H}[H_T|\Theta_\alpha]$ using lemma 3.3.
   16. We first show that we can apply this lemma, i.e. that for all $x \in \Sigma^T, y \in \mathbb{R}^{n+1}$ holds that
      17. I: $f_{\Theta_\alpha|B=1}(y) > 0 \implies \mathbb{P}[H_T = x|\Theta_\alpha = y] = \mathbb{P}[H_T = x|\Theta_\alpha = y, B = 1]$.
      18. II: $f_{\Theta_\alpha|B=0}(y) > 0 \implies \mathbb{P}[H_T = x|\Theta_\alpha = y] = \mathbb{P}[H_T = x|\Theta_\alpha = y, B = 0]$.
   19. I and II are true, since $B = (\Theta_\alpha)_{n+1}$
   20. Now we apply lemma 3.3:

$$\mathbb{H}[H_T|\Theta_\alpha] = \alpha\mathbb{H}[H_T|\Theta_\alpha, B=1] + (1-\alpha)\mathbb{H}[H_T|\Theta_\alpha, B=0]$$

   21. $= \alpha\mathbb{H}[H_T|\Theta_1] + (1-\alpha)\mathbb{H}[H_T|\Theta_0]$ (Def. Θ_α)
   22. $= \alpha\cdot 0 + (1-\alpha)\mathbb{H}[H_T] = (1-\alpha)\mathbb{H}[H_T]$ (Def. Θ_0, Θ_1)
23. $\mathbb{I}(H_T;\Theta_\alpha) = \mathbb{H}[H_T] - \mathbb{H}[H_T|\Theta_\alpha] = \alpha\mathbb{H}[H_T]$ (From 15 - 22)
24. $\mathbb{I}(H_T;\Theta_\alpha)/T = \alpha\mathbb{H}[H_T]/T$

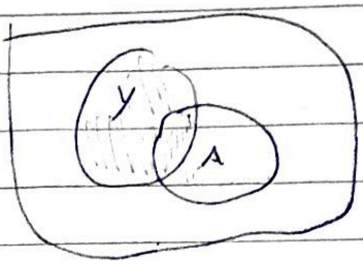
□

References

[JLLVR24] Hong Jun Jeon, Jason D Lee, Qi Lei, and Benjamin Van Roy. An information-theoretic analysis of in-context learning. *arXiv preprint arXiv:2401.15530*, 2024.

4 Below Proof Sketch, that I imples II in lemma 3.3

$f_{Y|A}(y) > 0$



- Case $f_{Y|A}(y), f_{Y|\bar{A}}(y) > 0$:

$$P[X=x \mid Y=y] = P(X=x \mid Y=y, A]$$

$$\underset{\text{bayes.}}{\overset{\uparrow}{\Rightarrow}} P(X=x \mid Y=y, \bar{A}] = P(X=y \mid Y=y)$$

- case $f_{Y|A}(y) = 0,\ f_{Y|\bar{A}}(y) > 0$:

Assume

$$P[X=x \mid Y=y] \neq P(X=x \mid Y=y, \bar{A}]$$

since $f_{Y|A}(y) = 0$; $f_Y(y) > 0$ $(f_{Y|\bar{A}}(y) \geq 0 \Rightarrow f_Y(y) > 0)$

$\hookrightarrow \mathbb{P}(\bar{A} \mid Y=y) = 1 \Rightarrow \mathbb{P}(\bar{A} \mid X=x, Y=y) = 1$

~~• by the law of bayes def. cond. f~~

~~$P(X=x \mid Y=y, \bar{A}) = \mathbb{P}(Y=$~~

~~$P[X=x \mid Y=y] = \mathbb{P}[\bar{A}]\, \mathbb{P}(X=x \mid Y=y, \bar{A})\, \mathbb{P}(\bar{A} \mid Y=y)$~~

- bayes :

$$P(X=x \mid Y=y, \bar{A}] = \overbrace{P(\cancel{x}\, \bar{A} \mid Y=y\cancel{t}, X=x]}^{1} \frac{P(X=x \mid Y=x)}{\underbrace{P(\bar{A} \mid Y=y)}_{1}}$$

$\Rightarrow$ contradict assumption □