LEARNING MINIMUM ENTROPY COUPLING

MILTON LIN

ABSTRACT. This is a short report on trying to learn the minimum entropy coupling problem via supervised learning. The experiments show that mindless use of transformers is not effective.

CONTENTS

1. INTRODUCTION

Minimum entropy couplings (MECs) are centrally required for a variety of applications. [Wit+23] showed that under an information-theoretic model of steganography, a procedure is secure if and only if it corresponds to a coupling, and a minimum entropy coupling corresponds to a maximally efficient secure procedure. They demonstrate strong performance for experiments using GPT-2 and WaveRNN as communication channels. In doing so, they used greedy coupling algorithm of [Koc+16] , [Koc+17] to couple pairs of distributions.

1.1. **Methodologies.**

(1) Set up a small transformer network to learn MECs between discrete uniform distributions of (variable) dimension smaller than 10, using supervised learning from exact MECs, see [Com+23].

(2) Can one define and implement suitable side-constraints that ensure that the learnt couplings are valid? [1]

Our goal is thus to get better transmission rates in perfectly secure steganography and shed light on whether faster/better MEC heuristics, which is still an open question. All key ideas are due to Christian Schroeder de Witt, and mistakes due to me.

Date: June 28, 2024.

[1] may consider a Lagrangian multiplier, or reinforcement learning to ensure that couplings are valid couplings - potentially this may be a two-step process where a given supervised learnt coupling is further refined using validity constraints

2. Results

The experiments focused from two directions: these are labelled as $(k, m, e, ?)$. k means that we trained 10^k datapoints per epoch. m is the dimension of the input distribution. e is number of epoch, and ? is the type of loss function. We generally focused on varying the dataset and loss function.

(1) Dataset: k, an optimal choice for a M2 Macbook Apple Pro, is $k = 6$, and we restrict between $e = 300 - 600$ epochs. One training takes approximately 5 to 6 hours. The generation of data is follows the algorithm provided in [Koc+16].

(2) Dimension of distribution: m, is restricted to 3: generally beyond $m = 4$, experiments are too time-consuming. Doing one 100 epoch dimension 6 experiment takes one day. Approximately 90% of the data exhibited an additive gap of less than 0.45, but focusing on this subset did not significantly improve results.

(3) ? would refer to the loss. There were three cases:

   (a) MSE (Mean Squared Error) has an additive gap of 1.15 but the marginal properties are far from close from inspection.

   (b) MSE+ marginal contstrained, is given by adding a a marginal loss:

$$MSE + \mu \left(\sum_{i=1}^{3} |c_i - p_i|^2 + |c_i - q_i|^2 \right)$$

   where $(p_i, q_i)_{i=1}^3$, are the two dimension 3 marginal distributions. This achieved an optimal additive gap of 1.23 with a weight parameter of $\mu = 1$.

   (c) MSE +M + entropy, is given by

$$MSE + M + H(c)$$

   where $H(c)$ is the entropy of joint coupling. Adding an entropy constraint improved the gap to 1.21 at 122 epochs, although MSE loss plateaued at 0.00205 in all constrained scenarios.

For each form of experiment $S := (k, m, e, ?)$ our resulting model M_S's performance is measured by additive gap, Definition A.10. From a theoretical standpoint, an additive gap difference of 1.21 is far from ideal, since from [Com+23], one is theoretically guarantee of 0.53 bits. Below, is a typical diagram of a training $(k, m, e, ?)$ for $k = 5, m = 3, e = 600, ? = MSE$, though, other experiments exhibit the same behavior, so we do not attach.

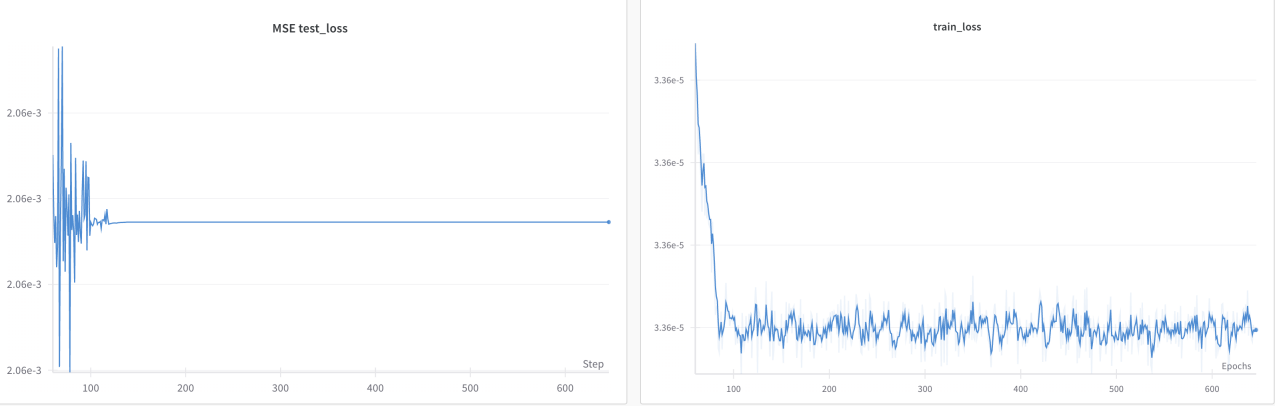

FIGURE 1. A typical training regime of test loss vs train loss for 500 epochs

2.1. **Conclusion and future direction.** These findings highlight the limitations of supervised learning. Given the computational constraints, we should begin by testing addition and multiplication as an easier test case. Then, on more basic architectures, such as CNNs, and new loss functions - such as an augmented Lagrnagian function.

I believe it would be practical to investigate the algorithmic capabilities of language models more systematically, particularly through the lens of program synthesis, [Mic+24], building upon the work of [Cha22a], [Cha22b], and combine this with the program synthesis pipeline. As a next step, we can combine work reinforcement learning and LLM methods for algorithm discovery, [Rom+23]. This gives a necessary guarantee on better interpretability of the algorithmic method than a supervised method.

Appendix A. Minimum entropy coupling algorithms

Definition A.1. Let X be a discrete random variable with distribution p.

$$\begin{aligned} H(X) &:= -\sum p(X=x)\log p(X=x) \\ &= -\mathbb{E}_p \log p(X) \end{aligned}$$

Definition A.2. Let X, Y be two random variables with joint distribution $p(x,y)$.

$$H(X,Y) := -\sum_{x,y} p(x,y)\log\frac{1}{p(x)p(y)}$$

The joint entropy is related to mutual information via the identity

$$I(X,Y) := H(X) + H(Y) - H(X,Y)$$

A.3. **Lattice structure on $\Delta_n^{\mathbf{dec}}$.** In the following discussion, $1 \le m \le n$.

- $\Delta_n^{\text{dec}} \hookrightarrow \Delta_n$, consisting of ordered tuples $x_0 \ge x_1 \ge \cdots x_n \ge 0$.
- $\Delta_n := \{(x_i)_{i=0}^n \in \mathbb{R}^{n+1} \,:\, x_i \ge 0\}$ is the positive n dimensional simplex.

We will begin by making Δ_n a poset.

Definition A.4. $x \le y$ if $\sum_{i=1}^k x_i \le \sum_{i=1}^k y_i$ for all $1 \le k \le n$.

Indeed this is important to us as :

Proposition A.5. *The Shanon entropy function is Schur-concave, i.e. it is a function $H : \Delta_n \to \mathbb{R}$, such that whenever $x \le y$, $H(x) \ge H(y)$.*

Example A.6. $m := (\frac{1}{n+1}, \ldots, \frac{1}{n+1}) \le n := (1, 0, \ldots, 0)$. In fact, m is the minimal element in Δ_n^{dec}, and n is the maximal element.

Proposition A.7. *$(\Delta_n^{dec}, \le)$ is a lattice.*

(1) Let $p, q \in \Delta_n$ and that both meets and joins (greatest lower bound). In fact, there is an explicit algorithm: $z : x \wedge y$ can be given as follows

- $z_0 = \min\{p_0, q_0\}$.
- *for $i = 2, \ldots, n$, it holds that*

$$z_i := \min\left\{\sum_{j=1}^i p_j, \sum_{j=1}^i q_j\right\}$$

Equivalently, we set

$$z_0 := \min\left\{\sum_{j=0}^i p_j, \sum_{j=0}^i q_j\right\} - \sum_{j=0}^{i-1} z_j$$

- *Using* $\sum_{k=0}^{n} z_k = \sum_{k=0}^{n} p_k = \sum_{k=0} q_k = 1$, *we have that for* $i = 0, \ldots, n$,

$$\sum_{k=i}^{n} z_k = \max\left\{\sum_{k=i}^{n} p_k, \sum_{k=i}^{n} q_k\right\}$$

A convenient definition would be that of aggregation, [CGV16], used in finding a D-measure,

Definition A.8. Let $p \in \Delta_n^{\text{dec}}$, $q \in \Delta_m^{\text{dec}}$. q is an *aggregation* of p if there is a partition of $\{1, \ldots, n\}$, into sets $\mathcal{P} := \{I_1, \ldots, I_m\}$, so that

$$q_j = \sum_{i \in I_j} p_i \quad j \in \{1, \ldots, m\}$$

Note that the choice of partition also defines a map

$$\mathcal{P} : \Delta_n^{\text{dec}} \to \Delta_m^{\text{dec}}$$

Note that for $m \leq n$ we always have a natural inclusion $(\Delta_m^{\text{dec}}, \leq) \hookrightarrow \left(\Delta_n^{\text{dec}}, \leq\right)$.

Corollary A.9. *Let* $q \in \Delta_m$, *which is an aggretaion of* $p \in \Delta_n$. *Then* $p \leq \mathcal{P}(q)$.

In otherwords, the aggregated distribution always has smaller entropy, by Schur-concavity, Proposition A.5.

Definition A.10. Any joint coupling $c \in \Delta_{mn+m+n}$, of $p \in \Delta_m$, $q \in \Delta_n$, is an aggregation of both p and q. Thus $c \leq p$ and $c \leq q$, so that $c \leq p \wedge q$, thus

$$H(c) \geq H(p \wedge q)$$

We refer the difference

$$H(c) - H(p \wedge q)$$

as the *additive gap* of the joint coupling c with repsect to p and q.

This gives a heuristic that to comute minimum entropy of joint coupling we would potentially want to compute $p \wedge q$, this is the approach of [CGV19]. We briefly describe it here: We say a matrix M is

- q-i-satisfied, if entries on columns $i, i+1, \ldots, n$ sums to $\sum_{k=i}^{n} q_i$.
- p-i-satisfied, if entries of rows $i, i+1, \ldots, n$, sums to $\sum_{k=i}^{n} p_i$

Definition A.11.

$$H(Y|X) := \mathbb{E}_X\left[H(Y|X = x)\right]$$

where $H(Y|X = x)$ is the entropy of conditional probability distribution $p_{Y|X=x}$.

Corollary A.12. $H(X, Y) = H(X) + H(X|Y)$

References

[CGV16] Cicalese, Ferdinando, Gargano, Luisa, and Vaccaro, Ugo. "Approximating probability distributions with short vectors, via information theoretic distance measures". In: *2016 IEEE International Symposium on Information Theory (ISIT)* (2016), pp. 1138–1142. URL: https://api.semanticscholar.org/CorpusID:17884614 (cit. on p. 5).

[CGV19] Cicalese, Ferdinando, Gargano, Luisa, and Vaccaro, Ugo. *Minimum–Entropy Couplings and their Applications*. 2019. arXiv: 1901.07530 [cs.IT] (cit. on p. 5).

[Cha22a] Charton, François. *Linear algebra with transformers*. 2022. arXiv: 2112.01898 [cs.LG] (cit. on p. 3).

[Cha22b] Charton, François. *What is my math transformer doing? – Three results on interpretability and generalization*. 2022. arXiv: 2211.00170 [cs.LG] (cit. on p. 3).

[Com+23] Compton, Spencer et al. "Minimum-Entropy Coupling Approximation Guarantees Beyond the Majorization Barrier". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, 25–27 Apr 2023 (cit. on pp. 1, 2).

[Koc+16] Kocaoglu, Murat et al. *Entropic Causal Inference*. 2016. arXiv: 1611.04035 [cs.AI] (cit. on pp. 1, 2).

[Koc+17] Kocaoglu, Murat et al. *Entropic Causality and Greedy Minimum Entropy Coupling*. 2017. arXiv: 1701.08254 [cs.IT] (cit. on p. 1).

[Mic+24] Michaud, Eric J. et al. *Opening the AI black box: program synthesis via mechanistic interpretability*. 2024. arXiv: 2402.05110 [cs.LG] (cit. on p. 3).

[Rom+23] Romera-Paredes, Bernardino et al. "Mathematical discoveries from program search with large language models". In: *Nature* 625 (2023), pp. 468–475. URL: https://api.semanticscholar.org/CorpusID:266223700 (cit. on p. 3).

[Wit+23] Witt, Christian Schroeder de et al. *Perfectly Secure Steganography Using Minimum Entropy Coupling*. 2023. arXiv: 2210.14889 [cs.CR] (cit. on p. 1).