Sinkformers: Transformers with Doubly Stochastic Attention

M. Sander, P. Ablin, M. Blondel, G. Peyré

Medha Agarwal and Garrett Mulcahy

January 24, 2023

Contents

Introduction

Introduction

Plan for next two weeks

- Careful summary of Sander et al. (2022) and relevant background materia
- This week: transformers, sinkformers, and connections between gradient flows and attention
- Next week: connections to PDEs/diffusion and experimental results

Transformers

Setup

The Transformer architecture follows an encoder-decoder structure

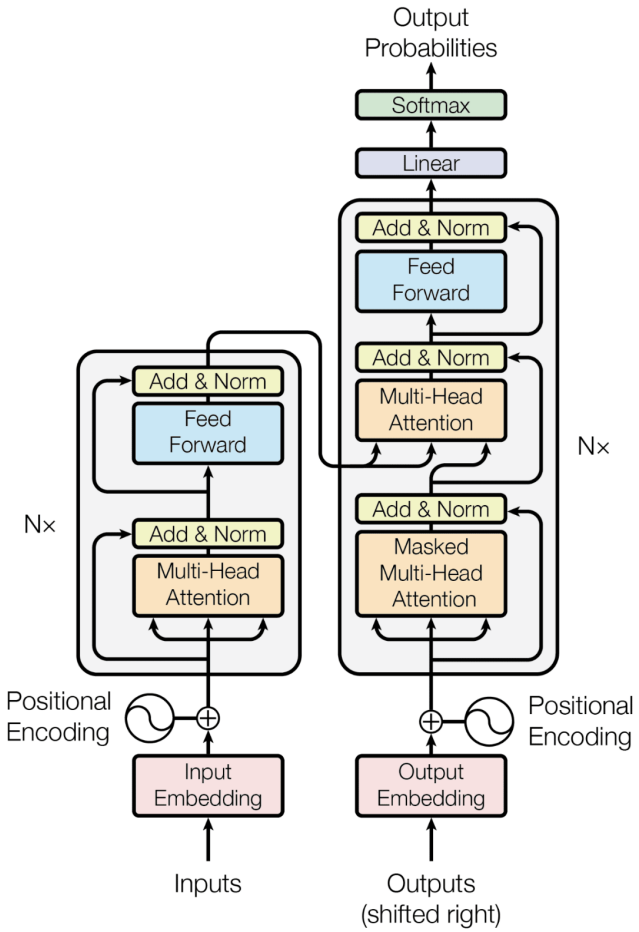

Figure: The encoder-decoder structure of the Transformer architecture.

General Attention Mechanism

The general attention mechanism makes use of three main components, namely the queries, Q, the keys, K, and the values, V. The main components used by the Transformer attention are the following:

- q and k denoting vectors containing the queries and keys, respectively.
- v denoting a vector containing the values.
- Q, K, and V denoting matrices packing sets of queries, keys, and values, respectively.
- W_Q, W_K, and W_V denoting projection matrices that are used in generating different subspace representations of the query, key, and value matrices denoting a projection matrix for the multi-head output.

General Attention Mechanism

- Suppose there are n queries and m keys. Each query vector, q_i, is matched against a database of keys to compute a score value. For key k_j, the score is

$$e_{q_i,k_j} = q_i^\top k_j \, .$$

- The scores are passed through a softmax operation to generate the weights:

$$\alpha_{q_i,k_j} = \frac{\exp(q_i^\top k_j)}{\sum_{l=1}^{m} \exp(q_i^\top k_l)} \, .$$

- The generalized attention is then computed by a weighted sum of the m value vectors, where each value vector v_j is paired with a corresponding key k_j:

$$\text{attention}(q_i, K, V) = \sum_{j=1}^{m} \alpha_{q_i,k_j} v_j \, .$$

Self-Attention Mechanism

Self-attention, sometimes called intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence.

Attention is All You Need (Vaswani et al. (2017))

- Given a n-sequence $(x_1, x_2, \ldots, x_n)$, embedded in dimension d, for instance a d-dimensional one-hot vector for n words sequence.
- Here $W_Q, W_K \in \mathbb{R}^{m \times d}$ and $W_V \in \mathbb{R}^{d \times d}$ are the query, key, and value matrices.
- The ith query, key, and value is respectively given by $W_Q x_i, W_K x_i$, and $W_V x_i$.
- The score matrix is given by $C \in \mathbb{R}^{n \times n}$ where $C_{i,j} = (W_Q x_i)^\top (W_K x_j)$.
- The SoftMax operator can be seen as a row-normalization of the matrix $K^0 = \exp(C)$ as follows $K^1_{i,j} = K^0_{i,j} / \sum_{l=1}^n K^0_{i,l}$.
- Therefore, the matrix K^1 is row-wise stochastic.
- The self attention mechanism is

$$x_i = x_i + \sum_{j=1}^{n} K^1_{i,j} (W_V x_j) \, .$$

Sinkformer

Sinkformer

- The main contribution of Sander et al. (2022) is towards **successively normalizing the rows and columns of** K^0, instead of only row normalization, based on empirical findings.
- This process is known to provably converge to a doubly stochastic matrix and is called **Sinkhorn's algorithm** .
- Given the score matrix $C \in \mathbb{R}^{n \times n}$ and $K^0 = \exp(C)$, Sinkhorm algorithm starts from K^0 and iterates as

$$K^{l+1} = \begin{cases} N_R(K^l) & \text{if } l \text{ is even} \\ N_C(K^l) & \text{if } l \text{ is odd}\,, \end{cases}$$

  where N_R and N_C correspond to row-wise and columnwise normalizations.
- The resulting matrix is denoted by $K^\infty = \texttt{Sinkhorn(C)}$.
- **Sinkformer** is the Transformer where SoftMax is replaced by Sinkhorn. It iterates as $x_i \leftarrow x_i + \sum_{j=1}^n K^\infty_{i,j}(W_V x_j)$.

Empirical Intuition behind Sinkformer

Sander et al. (2022) claim that on three different models and three different learning tasks, the learning process makes the attention matrices more and more doubly stochastic.

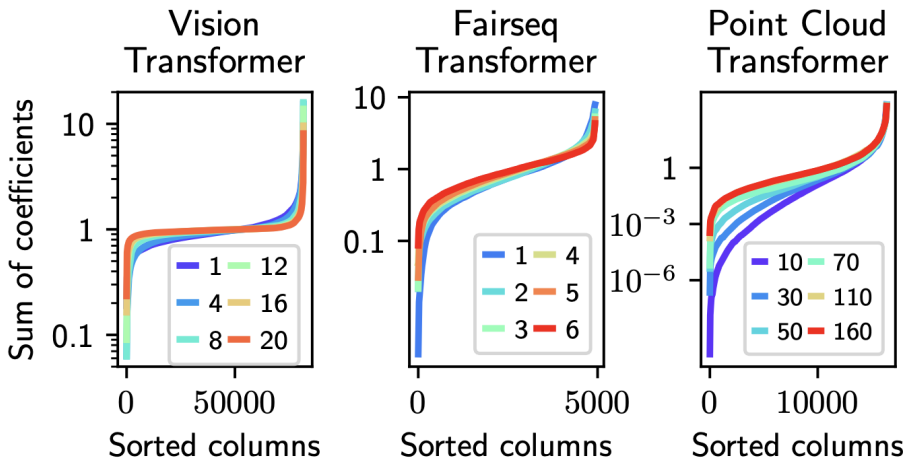

Figure: Sum over columns of attention matrices at different training epochs (color).

The **majority of columns** naturally sum closely to 1. Therefore, it seems natural to impose double stochasticity as a prior.

Wasserstein Metric Space

Consider particles at time $t = 0$ that are distributed according to density ρ_0. At time t, the velocity field v_t moves the particles around, i.e. $\dot{x}_t = v_t(x_t)$.
For smooth curves in $P^2(\mathbb{R}^d)$, the density of particles at time t, denoted by ρ_t, evolves as a solution to the continuity equation

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t v_t) = 0.$$

Suppose F is a function on $P^2(\mathbb{R}^d)$, then the Wasserstein gradient of F at ρ is given by

$$\nabla_W F(\rho) = \nabla_x \left(\frac{\delta F(\rho)}{\delta \rho} \right),$$

where $\delta F(\rho)/\delta \rho$ is the first variation.

Wasserstein Gradient Flows

For $F : P^2(\mathbb{R}^d) \to \mathbb{R}$, the Wasserstein gradient flow is given by the gradient flow with velocity

$$v_t = -\nabla_x \left(\frac{\delta F(\rho_t)}{\delta \rho_t} \right).$$

Plugging this velocity in the continuity equation, we get the flow equation

$$\dot{\rho}_t = \nabla \cdot \left(\nabla_x \left(\frac{\delta F(\rho_t)}{\delta \rho_t} \right) \cdot \rho_t \right).$$

A useful Wasserstein gradient flow for this paper is the heat flow.
Consider the entropy function $F(\rho) = \int \rho(x) \log \rho(x) dx$. The first variation is $(\delta F(\rho)/\delta \rho)(x) = 1 + \log \rho(x)$. And finally $\nabla_x (1 + \log \rho(x)) = \nabla_x \rho(x)/\rho(x)$. This implies

$$\dot{\rho}_t = \Delta \rho_t \,.$$

This is the **heat equation**. Therefore, heat equation is the gradient flow of the entropy function in the Wasserstein space.

Connection to Gradient Flows

Proposition 1

Proposition 1

Let $C \in \mathbb{R}^{n \times n}$. Consider for $(f, g) \in \mathbb{R}^n \times \mathbb{R}^n$ the modified cost function $\tilde{C}_{ij} := C_{ij} + f_i + g_j$. Then $\textbf{Sinkhorn}(\tilde{C}) = \textbf{Sinkhorn}(C)$.

Proof of Proposition 1

Proof

The variational formulation for Sinkhorn (Peyré and Cuturi, 2018) is

$$\texttt{Sinkhorn}(K^0) = \underset{K: K\mathbb{1}_n = K^\top \mathbb{1}_n = \mathbb{1}_n}{\arg\max} \mathcal{D}_{\mathrm{KL}}(K|K^0)$$

where

$$\mathcal{D}_{\mathrm{KL}}(K|K^0) = \sum_{i,j} K_{i,j} \log\left(\frac{K_{i,j}}{K^0_{i,j}}\right) .$$

If $K \in \{K : K\mathbb{1}_n = K^\top \mathbb{1}_n = \mathbb{1}_n\}$, then

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(K|\exp\tilde{C}) &= \sum_{i,j} K_{i,j} \log\left(\frac{K_{i,j}}{\exp(C_{i,j} + f_i + g_j)}\right) \\
&= \sum_{i,j} K_{i,j} \log\left(\frac{K_{i,j}}{\exp(C_{i,j})}\right) - \sum_i f_i - \sum_j g_j \\
&= \mathcal{D}_{\mathrm{KL}}(K|K^0) - \sum_i f_i - \sum_j g_j .
\end{aligned}$$

Attention and Gradient Flows

We have understood attention blocks as acting on a finite number of particles $(x_1, \ldots, x_n)$ irrespective of their ordering, so acting on $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$. We pass to the continuous case, i.e. acting on $\mu \in \mathcal{M}(\mathbb{R}^d)$, via defining

$$k^0(x, x') = \exp(c(x, x')) = \exp((W_Q x)^T W_K x')$$
$$k^1(x, x') = \frac{k^0(x, x')}{\int k^0(x, y) d\mu(y)}$$
$$k^\infty(x, x') = \texttt{Sinkhorn}(c),$$

where the Sinkhorn process here is the limiting function resulting from alternating normalizing with respect to the first and second argument.

Attention and Gradient Flows

At the particle level, we understand attention acting for $\ell \in \{0, 1, \infty\}$ as

$$x_i \mapsto x_i + \sum_{j=1}^{n} K_{i,j}^{\ell} W_V x_j$$

$$x \mapsto x + \int k^{\ell}(x, x') W_V x' d\mu(x')$$

That is, we are defining an Euler discretization (with step-size 1) of the ODE

$$\partial x_i / \partial t = T_{\mu}^{\ell}(x_i).$$

The distribution of the particles evolves with time according to the continuity equation

$$\partial_t \mu + \nabla(\mu T_{\mu}^{\ell}) = 0.$$

When do we have a Wasserstein gradient flow?

Proposition 2

Assumption 1

$$W_K^T W_Q = W_Q^T W_K = -W_V$$

Under this assumption, we now have $k^0(x, x') = \exp(-x^T W_V x')$.

Proposition 2 (PDEs associated to attention matrices)

Suppose assumption 1 holds. Let $\mathcal{F}^0, \mathcal{F}^\infty : \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}$ be defined by $\mathcal{F}^0(\mu) = \frac{1}{2}\int k^0 d(\mu \otimes \mu)$ and $\mathcal{F}^\infty(\mu) = -\frac{1}{2}\int k^\infty \log(k^\infty / k^0) d(\mu \otimes \mu)$. Then k^0, k^1, and k^∞ respectively generate the PDE

$$\partial\mu/\partial t + \nabla(\mu T_\mu^k) = 0$$

for $T_\mu^0 = -\nabla_W \mathcal{F}^0$, $T_\mu^1 = -\nabla\left(\log(\int k^0(., x')d\mu(x'))\right)$, and $T_\mu^\infty = -\nabla_W \mathcal{F}^\infty(\mu)$.

Proof of Proposition 2

Case $k = 0$

Consider the functional $\mathcal{F} : \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}$ defined by

$$\mathcal{F}(\mu) = \int h(x, y) d(\mu(x) \otimes \mu(y))$$

where $h \in C(\mathbb{R}^d \times \mathbb{R}^d)$. Let $\mu, \tilde{\mu} \in \mathcal{M}(\mathbb{R}^d)$, then for all $\epsilon > 0$ we have

$$\begin{aligned}\mathcal{F}(\mu + \epsilon\tilde{\mu}) &= \int h(x, y) d((\mu + \epsilon\tilde{\mu})(x) \otimes (\mu + \epsilon\tilde{\mu})(y)) \\ &= \mathcal{F}(\mu) + \epsilon \int h(x, y) d(\mu \otimes \tilde{\mu}) + \epsilon \int h(x, y) d(\tilde{\mu} \otimes \mu) + \epsilon^2 \mathcal{F}(\tilde{\mu}).\end{aligned}$$

This now gives

$$\frac{d}{d\epsilon} \mathcal{F}(\mu + \epsilon\tilde{\mu})\big|_{\epsilon=0} = \int h(x, y) d(\mu \otimes \tilde{\mu}) + \int h(x, y) d(\tilde{\mu} \otimes \mu).$$

The first variation is then

$$\frac{\delta \mathcal{F}}{\delta \mu}(\mu)(x) = \int h(x', x) d\mu(x') + \int h(x, x') d\mu(x').$$

Case $k = 0$ continued

When $h(x, y) = h(y, x)$ for all $x, y \in \mathbb{R}^d$ we have

$$\frac{\delta \mathcal{F}}{\delta \mu}(\mu)(x) = 2 \int h(x, x') d\mu(x')$$

and thus

$$\nabla_x \frac{\delta \mathcal{F}}{\delta \mu}(\mu)(x) = 2 \int \nabla_x h(x, x') d\mu(x').$$

Recall that $\mathcal{F}^0(\mu) = \frac{1}{2} \int k^0 d(\mu \otimes \mu)$ where $k^0(x, x') = \exp(-x^T W_V x')$ by assumption 1. This gives

$$\begin{aligned}
\nabla_x \frac{\delta \mathcal{F}^0}{\delta \mu}(\mu)(x) &= \int \nabla_x \exp(-x^T W_V x') d\mu(x') \\
&= -\int \exp(-x^T W_V x') W_V x' d\mu(x') \\
&= -\int k^0(x, x') W_V x' d\mu(x').
\end{aligned}$$

Case $k = 0$ continued

Hence, our probability distribution will evolve according to the continuity equation

$$\frac{\partial \mu}{\partial t} + \nabla(\mu T_\mu^0) = 0$$

where

$$T_\mu^0 = \int k^0(x, x') W_V x' d\mu(x') = -\nabla_W \mathcal{F}^0(\mu).$$

Case $k = \infty$

We recall that $\mathcal{F}^{\infty}(\mu) = -\frac{1}{2}\int k^{\infty}\log(k^{\infty}/k^{0})d(\mu\otimes\mu)$. Appealing to a duality result from Peyré and Cuturi (2018), we have

$$2\mathcal{F}^{\infty}(\mu) = -\max_{f}\int f + f^{c}d\mu,$$

where

$$f^{c}(x') = -\log\left(\int e^{f(x)+c(x,x')}d\mu(x)\right).$$

Moreover, the optimal pair is given by (f, f^{c}) with f unique up to a constant. We can also write

$$k^{\infty}(x,x') = e^{c(x,x')+f(x)+f(x')}.$$

For our optimal $f = f^{c}$ we have that

$$\mathcal{F}^{\infty}(\mu) = -\int f d\mu.$$

We then compute

$$\frac{d}{d\epsilon}\,\mathcal{F}^{\infty}(\mu+\epsilon\tilde{\mu})\big|_{\epsilon=0} = \frac{d}{d\epsilon}\left(-\int f d\mu - \epsilon\int f d\tilde{\mu}\right)\Bigg|_{\epsilon=0} = -\int f d\tilde{\mu}.$$

Case $k = \infty$ continued

We then have

$$\frac{\delta \mathcal{F}^{\infty}}{\delta \mu}(\mu)(x) = -f(x).$$

We solve for $\nabla_x f(x)$ by recalling that

$$f(x) = f^c(x) = -\log\left(\int e^{f(x')+c(x',x)} d\mu(x')\right).$$

This gives

$$e^{-f(x)} = \int e^{f(x')+c(x',x)} d\mu(x').$$

Taking the gradient with respect to x on both sides gives

$$\begin{aligned} e^{-f(x)}\nabla f(x) &= -\int e^{f(x')+c(x',x)} W_V x' d\mu(x') \\ \nabla f(x) &= -\int e^{f(x)+f'(x)+c(x',x)} W_V x' d\mu(x) \\ \nabla f(x) &= -\int k^{\infty}(x,x') W_V x' d\mu(x). \end{aligned}$$

Case $k = \infty$ continued

This gives that

$$\begin{aligned}\nabla_W \mathcal{F}^\infty(\mu) &= \nabla_x \frac{\delta \mathcal{F}^\infty}{\delta \mu}(\mu)(x) \\ &= -\nabla_x f(x) \\ &= -\int k^\infty(x, x') W_V x' d\mu(x).\end{aligned}$$

Hence, our probability distribution will evolve according to the continuity equation

$$\frac{\partial \mu}{\partial t} + \nabla(\mu T_\mu^\infty) = 0$$

where

$$T_\mu^\infty = \int k^\infty(x, x') W_V x' d\mu(x') = -\nabla_W \mathcal{F}^0(\mu).$$

Case $k = 1$

We recall that $T^1_\mu = -\nabla\left(\log(\int k^0(.,x')d\mu(x'))\right)$ and using that $k^0(x,x') = \exp(-x^T W_V x')$ we compute

$$\begin{aligned}
-\nabla\left(\log(\int k^0(.,x')d\mu(x'))\right)(x) &= -\int \frac{\nabla_x k^0(x,x')}{\int k^0(x,y)d\mu(y)}d\mu(x') \\
&= \int \frac{k^0(x,x')W_V x'}{\int k^0(x,y)d\mu(y)}d\mu(x') \\
&= \int k^1(x,x')W_V x' d\mu(x')
\end{aligned}$$

Can we recover T^1_μ as a Wasserstein gradient flow?

Proposition 3

Proposition 3

One has that $T^1_\mu = -\nabla[\log(\int k^0(.,x')d\mu(x'))]$ is not a Wasserstein gradient.

Suppose there exists $\mathcal{F} : \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}$ with first variation given by

$$\frac{\delta \mathcal{F}}{\delta \mu}(\mu)(x) = \log\left(\int k^0(x,x')d\mu(x')\right).$$

We compute the second variation to be

$$\frac{\delta^2 \mathcal{F}}{\delta \mu^2}(\mu)(x,x') = \frac{k^0(x,x')}{\int k^0(x,y)d\mu(y)}.$$

The second variation is symmetric with respect to (x,x') for all μ. Hence, for all $\mu \in \mathcal{M}(\mathbb{R}^d)$ and $x, x' \in \mathbb{R}^d$ we have

$$\frac{\delta^2 \mathcal{F}}{\delta \mu^2}(\mu)(x,x') = \frac{\delta^2 \mathcal{F}}{\delta \mu^2}(\mu)(x',x).$$

Proposition 3 continued

This symmetry implies for all $\mu \in \mathcal{M}(\mathbb{R}^d)$ and $x, x' \in \mathbb{R}^d$ that

$$\frac{k^0(x,x')}{\int k^0(x,y)d\mu(y)} = \frac{\delta^2 \mathcal{F}}{\delta \mu^2}(\mu)(x,x') = \frac{\delta^2 \mathcal{F}}{\delta \mu^2}(\mu)(x',x) = \frac{k^0(x',x)}{\int k^0(x',y)d\mu(y)}.$$

As $k^0(x,x') = \exp(-x^T W_V x')$ is nonzero and symmetric, we have for all μ, x, x' that

$$\int k^0(x,y)d\mu(y) = \int k^0(x',y)d\mu(y).$$

Setting $\mu = \delta_z$ gives that

$$k^0(x,z) = k^0(x',z).$$

By applying the symmetry of k^0 once again we have that $k^0(x,x') = \exp(-x^T W_V x')$ is constant, a contradiction.

References

Peyré, G. and Cuturi, M. (2018). Computational optimal transport.

Sander, M. E., Ablin, P., Blondel, M., and Peyré, G. (2022). Sinkformers: Transformers with doubly stochastic attention. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3515–3530. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.