1 Introduction

¡++¿

Transformers are important.

Transformers can be interpreted probabilistically through exponential families. The discrete number of hidden states in the input mapping to the same number of discrete hidden states in the output seems arbitrary.

We provide an interpretation of the transformer attention sublayer as approximate message passing, similar to the general pseudo-Bayesian model used for inference for temporal tracking/filtering problems.

When combined with the exponential family interpretation of softmax attention, we can provide a further interpretation as an update to the natural parameter distribution along with an update to the hidden state variable distribution.

2 Transformers

¡++¿

We focus on the attention sublayer, which performs a weighted average of the values.

$$H_j = \sum_{i=1}^{M} A\left(Q_j, K_i\right) V_i, \quad \forall j \in \{1, \ldots, M\} \tag{1}$$

¡++¿

For each j, the attention weights are defined by a discrete probability distribution over the keys. For short, we can write $A_{i,j} = A\left(Q_j, K_i\right)$. Notably, $\sum_i A_{i,j} = 1$.

For softmax attention, the attention function is defined by a normalized exponentiated dot product. Furthermore, the queries, keys, and values are linear transformation of the input hidden states by weight matrices W_q, W_k, W_v, respectively.

$$A_{i,j} = \frac{e^{\langle H_j W_k, H_i W_q \rangle}}{\sum_{i'=1}^{M} e^{\langle H_j W_k, H_{i'} W_q \rangle}} \tag{2}$$

¡++¿

From previous work, this is exactly an exponential family distribution of the key random variable over the natural parameter defined by the query:

$$\tilde{P}\left(K|Q\right) = h(K) e^{\langle K, Q \rangle - \log G(Q)} \tag{3}$$

¡++¿ where the intrinsic measure, $h(K)$, is a uniform measure over the discrete set of points $\{H_i W_k\}_{i=1}^{M}$, and $G(Q)$ defines the normalization function, which is the denominator of the softmax.

3 General Pseudo-Bayesian

¡++¿

We consider the general pseudo-Bayesian algorithm, with a collapsing over $k = 2$ timesteps.

For exact inference,

$$\begin{aligned} \sigma^{(t)}\left(A^{(t)}, X^{(t)}\right) &\rightarrow A^{(t+1)}, X^{(t+1)} \rightarrow O^{(t+1)} \\ \sigma^{(t+1)}\left(A^{(t+1)}, X^{(t+1)}\right) &:= P\left(A^{(t+1)}, X^{(t+1)} | O^{(1:t+1)}\right) \end{aligned} \tag{4}$$

¡++¿

For the general pseudo-Bayesian algorithms, we collapse distributions with the same discrete outcomes for the last k timesteps into a single distribution. More specifically, we perform are sending approximate messages $\sigma^{(t)}$ instead of exact messages $\sigma^{(t)}$, and as in expect propagation the approximations are found by performing an M-projection on the $X^{(t)}$ variable, possibly conditioned over each discrete assignment over the last k-1 timesteps. Thus, for the joint distribution at each time step, we are conditioning $X^{(t+1)}$ on $A^{(t-k:t+1)}$, which in the case of a discrete distribution of M points for A, is an M^k mixture model over k timesteps

For example, for $k = 2$, we are breaking up the M^2 mixtures into M mixtures of M components, and performing an M-projection on each of the mixtures. The result is we are condensing M^2 mixtures of probability distributions of an exponential family on $X^{(t+1)}$ into M mixtures of an exponential family.

$$\begin{aligned} \tilde{\sigma}^{(t)}\left(A^{(t)}, X^{(t)}\right) &= \tilde{\sigma}^{(t)}\left(A^{(t)}\right) \tilde{\sigma}^{(t)}\left(X^{(t)} | A^{(t)}\right) \\ P\left(A^{(t:t+1)}\right) &= P\left(A^{(t+1)} | A^{(t)}\right) \tilde{\sigma}^{(t)}\left(A^{(t)}\right) \\ \tilde{\sigma}^{(t+1)}\left(A^{(t+1)}\right) &= \int P\left(A^{(t+1)} | A^{(t)}\right) \tilde{\sigma}^{(t)}\left(A^{(t)}\right) dA^{(t)} \\ \tilde{P}\left(X^{(t:t+1)} | A^{(t:t+1)}\right) &= P\left(X^{(t+1)} | A^{(t+1)}\right) \tilde{\sigma}^{(t)}\left(X^{(t)} | A^{(t)}\right) \\ \sigma^{(t+1)}\left(X^{(t+1)} | A^{(t)}\right) &= \int dA^{(t)} \int P\left(X^{(t+1)} | A^{(t+1)}\right) \tilde{\sigma}^{(t)}\left(X^{(t)} | A^{(t)}\right) P\left(O^{(t+1)} | X^{(t+1)}\right) dX^{(t)} \\ \tilde{\sigma}^{(t+1)}\left(X^{(t+1)} | A^{(t+1)}\right) &= \text{M-proj}\left(\sigma^{(t+1)}\left(X^{(t+1)} | A^{(t+1)}\right)\right) \end{aligned} \tag{5}$$

¡++¿

For the forward message for $A^{(t+1)}$, we use marginalize the conditional probability $A^{(t+1)} | A^{(t)}$ with the forward message $\tilde{\sigma}^{(t)}\left(A^{(t)}\right)$.

For the forward message for $X^{(t+1)}$, we calculate the joint conditioned on A for both time steps, $P\left(X^{(t)}, X^{(t+1)} | A^{(t)}, A^{(t+1)}\right)$.

Then we marginalize over $X^{(t)}$ and then $A^{(t)}$, and use that as a prior for our likelihood $P\left(O^{(t+1)} | X^{(t+1)}\right)$. After incorporating our observation, we obtain the posterior distribution.

Finally, we have to perform an M-projection of this mixture distribution to our family of distributions, to calculate our approximate forward message for the conditional $X^{(t+1)} | A^{(t+1)}$.

3.1 GPB2 for linear-Gaussian conditionals

$$\begin{aligned} \tilde{\sigma}^{(t)}\left(A^{(t)} = a_k^{(t)}\right) &= \pi_k^{(t)} \\ \tilde{\sigma}^{(t)}\left(X^{(t)}|A^{(t)}\right) &= \mathcal{N}\left(\mu_{X^{(t)}|A^{(t)}}; \Sigma_{X^{(t)}|A^{(t)}}\right) \end{aligned} \tag{6}$$

¡++¿

A general linear-Gaussian conditioned on a discrete variable would be

$$\begin{aligned} P\left(X^{(t+1)}|X^{(t)}, A^{(t+1)} = a_k^{(t+1)}\right) &= \mathcal{N}\left(W_k X^{(t)}; Q_k\right) \\ P\left(X^{(t+1)}|A^{(t+1)} = a_k^{(t+1)}\right) &= \mathcal{N}\left(W_k \mu_{X^{(t)}}; W_k \Sigma_{X^{(t)}} W_k^\intercal + Q_k\right) \end{aligned} \tag{7}$$

¡++¿

However, our forward message does not pass a marginal distribution for $X^{(t)}$; instead we pass a conditional distribution for $X^{(t)}|A^{(t)}$.

Hence, we have M conditional Gaussians applied to M conditional Gaussian forward messages, for a total of M^2 conditional Gaussians.

$$\begin{aligned} P\left(X^{(t+1)}|X^{(t)}, A^{(t)}, A^{(t+1)} = a_k^{(t+1)}\right) &= \mathcal{N}\left(W_k X^{(t)}|A^{(t)}; Q_k\right) \\ P\left(X^{(t+1)}|A^{(t)}, A^{(t+1)} = a_k^{(t+1)}\right) &= \mathcal{N}\left(W_k \mu_{X^{(t)}|A^{(t)}}; W_k \Sigma_{X^{(t)}|A^{(t)}} W_k^\intercal + Q_k\right) \\ P\left(X^{(t+1)}|A^{(t+1)} = a_k^{(t+1)}\right) &= \sum_{k'=1}^{M} \pi_{k'}^{(t)} \mathcal{N}\left(W_k \mu_{X^{(t)}|A^{(t)}=a_{k'}^{(t)}}; W_k \Sigma_{X^{(t)}|A^{(t)}=a_{k'}^{(t)}} W_k^\intercal + Q_k\right) \end{aligned} \tag{8}$$

¡++¿

Then we multiply by the likelihood $P\left(O^{(t+1)}|X^{(t+1)}\right)$, and somehow calculate the posterior distribution.

Finally, we calculate the M-projection to a Gaussian distribution, by calculating the first two moments.

That is our message, $\sigma^{(t+1)}\left(X^{(t+1)}|A^{(t+1)}\right)$.

3.2 GPB2 for exponential families

The previous example for GPB2 defined mixtures of Gaussians, but we can extend that to mixtures of exponential families.

We had exact solutions for linear Gaussian marginalizations, and we can extend this to linear exponential families through the change of variables of a deterministic transformation:

$$P\left(X^{(t+1)}|X^{(t)}, A^{(t)}, A^{(t+1)} = a_k^{(t+1)}\right) \propto e^{\left\langle W_k X^{(t)}|A^{(t)}\right\rangle} <++> = \frac{P\left(X^{(t)}|A^{(t)}\right)}{|W_k|} <++> \tag{9}$$

¡++¿ (???)

As with linear Gaussians, the mean is just a linear transformation of the mean of $X^{(t)}|A^{(t)}$:

$$\mu_{X^{(t+1)}|A^{(t)}, A^{(t+1)}=a_k^{(t+1)}} = W_k \mu_{X^{(t)}|A^{(t)}} \tag{10}$$

¡++¿

If this linear transformation is the same for every k, which would make it independent of $A^{(t+1)}$, we can define it as a single weight matrix, such as the linear transformation W_v for values.

The one concern is that if we want to calculate the posterior distribution analytically, we must be defining mixture distributions over conjugate priors, where the the distributions must be conjugate to the observed data conditional distributions, $P\left(O^{(t+1)}|X_{t+1}\right)$.

Explicitly we would define our exponential family distributions over the random variables $O^{(t)}$, with natural parameters $\eta_O := X$. $P\left(X^{(t+1)}|A^{(t+1),A^{(t)}}\right)$ would be the conjugate priors, which would make A natural parameters of the conjugate priors, which are the expected sufficient statistics of the exponential family distributions.

3.3 The Interacting Multiple Model (IMM) algorithm

4 Attention as a variant of the GPB2/IMM

¡++¿

For the attention sublayer, there is no analog to observation, which means we avoid having to calculate a posterior distribution.

As in GPB2/IMM, we can assume we are sending an approximate message:

$$\tilde{\sigma}^{(t)}\left(A^{(t)}, X^{(t)}\right) = \tilde{\sigma}^{(t)}\left(A^{(t)}\right)\tilde{\sigma}^{(t)}\left(X^{(t)}|A^{(t)}\right) \tag{11}$$

¡++¿

For exponential families, we can describe distribution by the natural parameter, η, or it's dual, the expected sufficient statistic, which in this case is the mean, μ. Hence, we send our message about $X^{(t)}|A^{(t)}$ by sending the expected sufficient statistic, which is the hidden state value, H_i in a transformer.

Most importantly, in GPB2 the discrete variables $A^{(t)}$ are updated independently of the continuous variables $X^{(t)}$, whereas the attention probabilities are dependent on the probability distribution for X.

$$P\left(A^{(t+1)} = j|A^{(t)} = i, \tilde{\sigma}^{(t)}\left(X^{(t)}|A^{(t)}\right)\right) = \frac{e^{\langle \mu_i, \eta_j \rangle}}{\sum_{i'=1}^{M} e^{\langle \mu_{i'}, \eta_j \rangle}}\tilde{\sigma}^{(t)}\left(A^{(t)}\right) \tag{12}$$

¡++¿

Note that under this interpretation, $\tilde{\sigma}^{(t)}\left(A^{(t)}\right) = \frac{1}{M}$.

As in IMM, we can define the transition $X^{(t+1)}|X^{(t)}, A^{(t)}$ through the deterministic linear transformation of our value matrix, which allows us to marginalize out $X^{(t)}$

$$E_{X\sim P(X^{(t+1)}|A^{(t)}=j)}\left[X\right] = H_j W_v \tag{13}$$

¡++¿

Finally, the only step left is to marginalize out $A^{(t)}$, which is the definition of a mixture distribution.

$$
\begin{aligned}
E_{X\sim X^{(t+1)}|A^{(t+1)}}\left[X\right] &= \sum_{i=1}^{M} \int dX^{(t)} P\left(A^{(t+1)} = j|A^{(t)} = i\right) X^{(t+1)} P\left(X^{(t+1)}|X^{(t)}, A^{(t)} = i\right) P\left(X^{(t+1)}|A^{(t)} = i\right) \\
&= \sum_{i=1}^{M} A_{i,j} H_i W_v
\end{aligned}
\tag{14}
$$

¡++¿

Notably, we note that by sending the expectation of this mixture distribution as a message, we are effectively collapsing information about it into a single sufficient statistic, and in an analogy to GPB algorithms, we are collapsing it into a single distribution as an approximate message.

Moreover, through the canonical link function / activation function, we can define a one to one mapping between the expected sufficient statistic and the natural parameters. This means that for each expected sufficient statistic $H_j = \mu_j$, we are determining the corresponding natural parameter η_j, and for our attention we are calculating a conjugate prior probability $P\left(\eta_i|\mu_j\right)$

Importantly, the natural parameters of this exponential family is not the expected sufficient statistic. So we conjecture that the weight and key matrices are used to provide a linear approximation to the canonical link function:

$$
\begin{aligned}
P\left(\eta_i|\mu_j\right) &\propto e^{\langle f(\eta_i), \mu_j \rangle} \\
&\approx e^{\langle \mu_i W_q, \mu_j W_k \rangle}
\end{aligned}
\tag{15}
$$

¡++¿

4.1 Approximate Gibbs Sampling

¡++¿

By using a discrete approximation for the distribution $P(\eta|\mu)$ to create a mixture distribution, we note that we performing a discrete approximation of a marginalization over η.

Under this interpretation, we are doing alternating updates to η and μ, reminiscent of Gibbs sampling.

Under this interpretation, repeated application of the transformer layers would seem to imply sequential convergence to a stationary distribution. However, from previous work we noted that when combined with the skip connection, the update is also equivalent to a gradient ascent step on the natural parameter, which suggests the linear approximation to the link function causes blowup of the normalization function.