1 Introduction

¡++¿

2 Related Work

¡++¿

2.1 Conditional Bayesian Networks

¡++¿

2.2 Kernel interpretation of transformers

¡++¿

2.3 Contrastive Divergence

¡++¿

3 Theory

¡++¿

3.1 Exponential Family

¡++¿ Members of an exponential family of distributions can be written in natural parameter form, through the Pitman-Koopman-Dermois Theorem:

$$p(x|\eta) = h(x)e^{\langle u(x),\eta\rangle - G(\eta)} < ++ > \tag{1}$$

¡++¿

where $h(x)$ is the intrinsic measure, η is the natural parameters, $u(x)$ is the sufficient statistic, and $G(\eta)$ is the log-normalizer:

$$G(\eta) = \log \int h(x)e^{\langle u(x),\eta\rangle} dx \tag{2}$$

¡++¿

When the sufficient statistic is simply the random variable of interest, a special exponential family can be generated using a more general form of the Laplace transform of the intrinsic measure:

$$\begin{aligned} G(y) &= \log \mathcal{L}\left\{h(x)\right\}(-y) \\ &= \log \int e^{\langle y,x\rangle} h(x)dx \end{aligned} \tag{3}$$

¡++¿

In particular, if we choose our intrinsic measure to be a distribution over a discrete set of points, such as from an empirical distribution, the induced exponential family distribution is given by a dot-product similarity softmax:

$$
\begin{aligned}
h(x) &= \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}(x) \\
p(x|y) &= h(x)\frac{e^{\langle y,x\rangle}}{\frac{1}{N}\sum_{n=1}^{N}e^{\langle ymx_n\rangle}} \\
&= \frac{e^{\langle y,x\rangle}}{\sum_{j=1}^{N}e^{\langle y,x_i\rangle}}, \text{if} \quad x \in \{x_1,\ldots,x_N\}
\end{aligned}
\tag{4}
$$

¡++¿

Thus, the exponential family induced from the empirical distribution is functional form of contrastive divergence, which is also used in attention. In particular, if we choose one of the points x_i as a our natural parameter, our distribution provides us similarity to the points of the empirical distribution:

$$
p(x_i|x_k) = \frac{e^{\langle x_k,x_i\rangle}}{\sum_{j=1}^{N}e^{\langle x_k,x_j\rangle}} \tag{5}
$$

¡++¿

Therefore, the functional form of transformers can be interpretted as an exponential family where the sample space of key variables is restricted to the input variables of the attention layer, and for self-attention, the natural parameters are also restricted the same discrete space.

One particularly useful property of exponential families relates the expected sufficient to the gradient of the log partition function.

$$
\nabla_\eta G(\eta) = E_{x\sim p(x|\eta)}\left[u(x)\right] < ++ > \tag{6}
$$

¡++¿

In the particular case of the exponential family induced from an empirical distribution, the expectation is exactly the output of the attention layer, which can also be shown to be Nadaraya-Watson kernel regression:

$$
\nabla_\eta G(x_q) = \sum_{i=1}^{N}\frac{e^{\langle x_q,x_i\rangle}}{\sum_{i'=1}^{N}e^{\langle x_q,x_{i'}\rangle}}x_i \tag{7}
$$

¡++¿

Ignoring the weight matrices, this is ith output of an attention head, so we can also view it as a gradient ascent step of the log normalizer, $G(\eta)$. In effect, we are taking parallel gradient step of log normalizer function, at each query point.

The DEQ interpretation of applying the same attention layer and adding it to a residual connectio is that we are doing a discrete update, possibly towards a fixed point. Iterating our gradient ascent for our log normalizer function would ultimately cause divergence to infinity. If we were to hope for convergence, we would need to compensate for the expansion of the gradient step with a contraction step.

In practice, it turns out models do include a contractive step. For contrastive divergence, they do not use dot-product similarity; rather, they use cosine similarity. This is effectively projecting the points on to a unit sphere, so it no longer matters if the points are being pushed towards infinity. Transformers use layer normalization, so the set of input hidden states become renormalized into a Gaussian distribution with a learned mean and (diagonal) covariance.

One interpretation is that there is a true probability distribution of the hidden states, and we have collected samples from that probability distribution as an empirical data distribution of the true probability distribution. Then we apply the Laplace transformation upon the discrete distribution to induce an exponential family distribution, and we are now applying gradient ascent upon the the log normalizer to generate a new empirical data distribution.

Intuitively, this would push the points outwards towards infinity. Contrastive divergence resolves this by using cosine similarity, which is a projection of the points back to the unit sphere. Transformers uses layer normalization, which renormalizes the values back to a learned mean and variance.

One question of interest is what is the fixed point of this attention layer? If we were to not use an empirical distribution, the fixed point should be a gaussian distribution, so we would assume that the attention layer spreads apart the points while the normalization shrinks it back down towards a Gaussian distribution.

Hence, we would expect

Another question is that if we replaced an empirical distribution of points with an non-discrete distribution, would we eventually converge to Gaussian fixed point distribution?

3.2 Laplace Transformation

¡++¿