Preliminary Research Idea (6): Correlated Sampling from Multiple Softmax Distributions Using Gumbel-Max Trick and Copulas

Professor Richard Yi Da Xu
HKBU and TadReamk Limited

The Gumbel-Max trick is a very ingenious method that, while mathematically equivalent to standard Softmax sampling, differs structurally in its sampling path. It eliminates the need to compute concrete probability vectors (i.e., Softmax vectors). Instead, it directly adds independent noise drawn from a Gumbel distribution to each unnormalized logit. As shown in online notes, the index k chosen by this method follows the same categorical distribution as performing Softmax and then standard sampling. The main advantage is that the $\arg\max$ operation does not require computing the denominator (normalization constant), making the sampling process extremely simplified and efficient.

One of my current research directions involves applying Copula-based methods to novel scenarios in the LLM era, specifically for sampling tasks involving multiple Softmax functions. An additional advantage of the Gumbel distribution is that its inverse cumulative distribution function (CDF) is easy to compute (tractable). This tractability means Copulas can be used to sample from the joint density of two (or more) Softmax functions (e.g., in multi-modal applications). For example, when sampling two different tokens (e.g., from two different attention heads, or from text and image modalities), Copula can be used to generate correlated Gumbel noise. This introduces a dependency structure between the choices from two different Softmax distributions without changing their respective marginal probabilities.

Consider the scenario of sampling discrete indices k_1 and k_2 from two different Softmax distributions (with logits composed of two N-dimensional vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively) such that their randomness has a correlation coefficient ρ. The procedure proceeds as follows:

Step 1: Define Dependency Relationship $\mathbf{\Sigma}$

Construct a covariance matrix to represent the expected correlation between sampling events:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{1}$$

Step 2: Sample Latent Gaussian Variables (Z)

Sample N pairs of latent Gaussian variables:

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{2}$$

Step 3: Probability Integral Transform (Gaussian → Uniform)

Apply the cumulative distribution function (CDF) Φ of the standard normal distribution. This maps $\mathbf{Z}$ to unit hypercube space while preserving rank correlation (i.e., Gaussian Copula):

$$\begin{aligned} U_1 &= \Phi(Z_1) \\ U_2 &= \Phi(Z_2) \end{aligned} \tag{3}$$

Step 4: Inverse Transform Sampling (Uniform → Gumbel)

Apply the inverse CDF (quantile function) of the Gumbel distribution to transform the correlated uniform variables into correlated Gumbel variables:

$$\begin{aligned} G_1 &= -\ln(-\ln(U_1)) \\ G_2 &= -\ln(-\ln(U_2)) \end{aligned} \tag{4}$$

Note that G_1 and G_2 still follow Gumbel$(0, 1)$ marginal distributions, but they are no longer independent.

Step 5: Apply Gumbel-Max Trick

Finally, we apply the Gumbel-Max trick:

$$\begin{aligned} k_1 &= \arg\max(\boldsymbol{\alpha} + \mathbf{G}_1) \\ k_2 &= \arg\max(\boldsymbol{\beta} + \mathbf{G}_2) \end{aligned} \tag{5}$$

Currently, the use of the Gumbel distribution within the copula framework, although obvious, is merely a prelude. I believe this modification opens up many new possibilities in the era of Large Language Models (LLMs) and provides ample material for subsequent model evolution.