L173 Exercise 1
The probabilistic modelling behind language diffusion

D.J. Wischik

This exercise asks you to explain Large Language Diffusion Models*, Nie et al. 2025, in the language used by this course — as a probabilistic model, trained so as to minimize a KL divergence.*

Your answer should be mathematically self-contained: you should define all the terms and concepts you need, so that your answer is meaningful to someone who hasn't read the paper or its surrounding literature. Maths isn't like poetry, relying on resonances and conventions; it is like code, where you have to declare all your symbols before you can use them. Your answer should be roughly 800 'prose words' (not including equations), which is roughly 2½ pages of typeset text. You won't lose marks for not hitting this number, but if you are writing much more or much less then you've probably missed the point of the exercise. LLM usage is permitted.

1. Expression (1) is bad maths notation. In the course's notation, it would be written

$$\text{maximizing } \mathbb{E}_{X\sim q_X} \log p^\theta(X) \text{ over } \theta \quad \equiv \quad \text{minimizing } \mathrm{KL}(q_X \,\|\, p^\theta_X) \text{ over } \theta.$$

   In this paper, what is the sample space for X?

2. The paper proposes a model in which X is generated sequentially, over the course of T steps where T is a hyperparameter. Define this generative model carefully, and give a 'wiring diagram' of the style used in the course. How does the model specify which tokens to unmask? How does it fill them in? Introduce whatever supplementary random variables you need, and whatever black boxes for generating conditional random variables.

3. Derive a bound for $\mathrm{KL}(q_X \,\|\, p^\theta_X)$ corresponding to equation (3). *[At the bottom of the handout for lecture 4, there is a calculation you may find useful. Your bound should depend on T, unlike (3).]*

4. If we could train the generator network perfectly, your KL bound should approach zero as $T \to \infty$, but for finite T it may not be possible to achieve zero. Explain why, and explain why we nonetheless don't want T too large.

5. The paper says in §2.4 that "the remasking strategy should be purely random". Consider a generator with a probabilistic unmasking rule, which might not be uniform: each step, the rule randomly selects a set of tokens to unmask, given the current string. Rederive the bound on $\mathrm{KL}(q_X \,\|\, p^\theta_X)$ for this generator.

6. What does your answer to part 4 suggest about this unmasking rule?