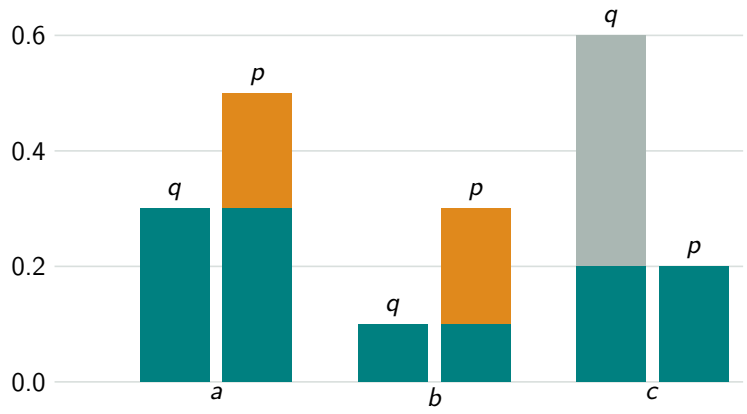
How Often Can We Keep the Draft Token?

Due: Wednesday, December 2 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

Generating a paragraph with a large language model takes one pass through the model per token. In *speculative decoding*, we let a smaller, faster model draft the next few tokens. With the draft in hand, the large model can compute its next-token probabilities at every position in one pass. If we keep several guesses in a row, that pass gives us several tokens.

Fix the text so far. The small model proposes a token Y from its next-token distribution q. We can keep Y or replace it. Either way, the token we return, X, should have exactly the large model's distribution p. In the figure, the small model proposes c three times as often as it should.



1. For tokens (a, b, c), we have $q = (0.3, 0.1, 0.6)$ and $p = (0.5, 0.3, 0.2)$. Suppose we keep every proposed a and b. When the draft proposes c, with what probabilities should we return a, b, and c?

2. Now suppose we have V tokens. We keep a proposed token j with probability $\min(1, p_j/q_j)$. Whenever we reject a token, we draw from the same replacement distribution. What probability should that distribution assign to each token?

3. We could use any randomized rule that takes $Y \sim q$ and returns $X \sim p$. Show that the rule in part 2 maximizes $\Pr(X = Y)$.

Hint: For each $j = 1, \ldots, V$, compare $\Pr(X = Y = j)$ with $\Pr(X = j)$ and $\Pr(Y = j)$.

Takeaway: *The teal overlap in the figure tells us how often we can keep the draft token and still sample from p. The more tokens we can keep in a row, the further we can get with one pass of the large model.*