Mentored decoding

Vivien Tran-Thien

September 2023

1 Optimization problem

We want to adapt the speculative sampling scheme proposed in [2][1] to increase the acceptance rate of draft tokens while limiting the Kullback-Leibler divergence between the resulting distribution and the distribution of the target model. With the following notations:

- p_i, probability that the next token is i according to the draft model;
- q_i, probability that the next token is i according to the target model;
- r_i, probability to accept the draft token i as the next token;
- s_i, probability to select i as the next token if the draft token is rejected;
- π_i, the distribution resulting from the adapted rejection sampling scheme. As shown in [1], $\pi_i = p_i r_i + s_i(1 - \sum_j p_j r_j)$;

...we aim at solving the optimization problem below:

$$\max_{(r_i)_i,(s_i)_i} \sum_i p_i r_i \tag{1}$$

$$\text{s.t.} \quad \sum_i q_i \ln \frac{q_i}{\pi_i} \leq D \tag{2}$$

$$\sum_i s_i = 1 \tag{3}$$

$$\forall i \quad r_i \in [0,1], s_i \in [0,1], r_i + s_i > 0^{12} \tag{4}$$

[1]We assume $r_i + s_i > 0$ so that $\pi_i > 0$ and the Kullback-Leibler divergence is defined.

[2]Whenever a dummy variable, e.g. i, is used in this document without further information, e.g. in $\sum_i, \forall i$ or $(r_i)_i$, it should be understood as covering the whole vocabulary V of the models. For example, $\forall i$... is equivalent here to $\forall i \in V$...

2 A solution exists

The values $(r_i)_i, (s_i)_i$ defined in the speculative sampling scheme proposed in [2][1] satisfy the constraints (2), (3) and (4). If we call R_0 the corresponding value for $\sum_i p_i r_i$, we can then add the constraint $R_0 \leq \sum_i p_i r_i$ without changing the set of solutions.

Given (2), for any point in the feasible region, we have for all i:

$$q_i \ln \frac{1}{\pi_i} \leq \sum_j q_j \ln \frac{1}{\pi_j} \leq D - \sum_j q_j \ln q_j$$

... or:

$$\ln \pi_i \geq -\frac{1}{q_i}(D - \sum_j q_j \ln q_j)$$

Since $R_0 \leq \sum_i p_i r_i$, this leads to:

$$p_i r_i + s_i(1 - R_0) \geq p_i r_i + s_i(1 - \sum_j p_j r_j) \tag{5}$$

$$\geq e^{-\frac{1}{q_i}(D - \sum_j q_j \ln q_j)} \tag{6}$$

(6) is a more stringent constraint than $r_i + s_i > 0$ so this latter constraint can be ignored when defining the feasible region. The feasible region then becomes a compact space and, since the objective function is continuous, we conclude that the optimization problem admits at least one solution.

3 Some constraints can be simplified

Let's now show that we can always assume that $r_i > 0$. Suppose a solution of the optimization problem is found with $r_i = 0$. Since $r_i + s_i > 0$, we have $s_i > 0$. We select k so that $r_k > 0$[3]. For a sufficiently small ϵ and with the following substitutions:

- $r_i \to r_i + \epsilon / p_i$
- $r_k \to r_k - \epsilon / p_k$
- $s_i \to s_i - \epsilon / (\sum_j p_j (1 - r_j))$
- $s_k \to s_k + \epsilon / (\sum_j p_j (1 - r_j))$

$\ldots \pi$, $\sum_i p_i r_i$ and all the constraints would be left unchanged. We would then get a solution with $r_i > 0$ (and $r_k > 0$). This proves that all r_i can safely be assumed to be strictly positive and we can rewrite the constraints (4) as:

$$r_i \in]0, 1], s_i \in [0, 1]$$

[3]Not all r_k are equal to zero because $\sum_i p_i r_i \geq R_0 > 0$ (cf. previous section).

Moreover, let's show that the constraint on the Kullback-Leibler divergence is always saturated as soon as $\sum_i p_i r_i < 1$. Indeed, in this case, there is at least one i with $r_i < 1$. If $\sum_i q_i \ln \frac{q_i}{\pi_i} \in]0, D[$, it is possible to increase r_i by $\epsilon > 0$ (and $\sum_i p_i r_i$ by $p_i \epsilon$) while still keeping the Kullback-Leibler divergence in $]0, D[$ (because the Kullback-Leibler divergence is continuous). Moreover, $\sum_i q_i \ln \frac{q_i}{\pi_i}$ cannot be equal to 0 because increasing any $r_i < 1$ would lead to a higher objective function and a strictly positive Kullback-Leibler divergence. Constraint (2) can then be rewritten as:

$$\sum_i q_i \ln \frac{q_i}{\pi_i} = D \tag{7}$$

4 A closely related optimization problem

Instead of directly tackling the optimization problem presented above, we focus on the following one:

$$\min_{(r_i)_i, (s_i)_i} \sum_i q_i \ln \frac{q_i}{\pi_i} \tag{8}$$

$$\text{s.t.} \quad \sum_i p_i r_i \geq R \tag{9}$$

$$\sum_i s_i = 1 \tag{10}$$

$$\forall i \quad r_i \in]0,1], s_i \in [0,1] \tag{11}$$

Like its predecessor, we can show that this optimization problem admits a global minimum. Since all equality and inequality constraints are affine, this global minimum satisfies the Karush–Kuhn–Tucker conditions. In particular, with the Lagrangian written as:

$$\begin{aligned}\mathcal{L}(p_i, r_i, \lambda, \mu, \nu_i, \chi_i) = &\sum_i q_i \ln \frac{q_i}{\pi_i} + \lambda(R - \sum_i p_i r_i) \\ &+ \mu(\sum_i s_i - 1) + \sum_i \nu_i (r_i - 1) - \sum_i \chi_i s_i\end{aligned}$$

... there exist constants λ, μ, ν_i and χ_i such that:

Stationarity:

$$\frac{\partial \mathcal{L}}{\partial r_i} = p_i (\sum_j \frac{q_j}{\pi_j} s_j - \frac{q_i}{\pi_i} - \lambda) + \nu_i = 0 \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial s_i} = \mu - \frac{q_i}{\pi_i}(1 - \sum_j p_j r_j) - \chi_i = 0 \tag{13}$$

Dual feasibility:

$$\lambda \geq 0, \quad \nu_i \geq 0, \quad \chi_i \geq 0 \tag{14}$$

Complementary slackness:

$$\lambda = 0 \text{ or } \sum_i p_i r_i = R \tag{15}$$

$$\nu_i = 0 \text{ or } r_i = 1 \tag{16}$$

$$\chi_i = 0 \text{ or } s_i = 0 \tag{17}$$

Using (13), (14) and (17), we can see that $\frac{q_i}{\pi_i}$ is constant if $s_i > 0$ and takes lower values if $s_i = 0$:

$$\text{If } s_i > 0, \quad \frac{q_i}{\pi_i} = \frac{\mu}{1 - \sum_j p_j r_j} \equiv \beta \tag{18}$$

$$\text{If } s_i = 0, \quad \frac{q_i}{\pi_i} = \frac{q_i}{p_i r_i} = \frac{\mu - \chi_i}{1 - \sum_j p_j r_j} \leq \beta \tag{19}$$

Besides, given (12), (14) and (16):

$$\begin{aligned} \text{If } r_i < 1, \quad \frac{q_i}{\pi_i} &= \sum_j \frac{q_j}{\pi_j} s_j - \lambda \\ &= \beta - \lambda \quad \text{(given (10) and (18))} \\ &\equiv \alpha \end{aligned} \tag{20}$$

$$\text{If } r_i = 1, \quad \frac{q_i}{\pi_i} = \alpha + \frac{\nu_i}{p_i} \geq \alpha \tag{21}$$

In summary, $\frac{q_i}{\pi_i}$ is always between α and $\beta \equiv \alpha + \lambda$. Could α and β be equal? If so, they would be equal to 1 because q and π are both probability distributions. We would then be brought back to the case of lossless speculative decoding[2][1]. Therefore, if we want to achieve a strictly higher acceptance probability R, we need $\beta > \alpha$. Given (20) and (15), this leads to $\lambda > 0$ and:

$$\sum_i p_i r_i = R \tag{22}$$

With $\beta > \alpha$, (18) and (20) imply that:

$$s_i = 0 \text{ if } r_i < 1$$
$$r_i = 1 \text{ if } s_i > 0$$

Therefore, $\frac{q_i}{\pi_i} = \frac{q_i}{p_i r_i}$ if $r_i < 1$ and, given (20) and (21):

$$r_i = \min(\frac{q_i}{\alpha p_i}, 1) \tag{23}$$

This leads to:

$$
\begin{aligned}
1 - R &= 1 - \sum_i p_i \min(\frac{q_i}{\alpha p_i}, 1) \\
&= -\frac{1}{\alpha} \sum_{q_i/p_i \le \alpha} q_i + (1 - \sum_{q_i/p_i > \alpha} p_i) \\
&= \sum_{q_i/p_i \le \alpha} p_i - \frac{1}{\alpha} \sum_{q_i/p_i \le \alpha} q_i \qquad (24) \\
&= \sum_i Relu(p_i - \frac{q_i}{\alpha}) \qquad (25)
\end{aligned}
$$

Let's now focus on β. Following (18) and (19):

$$
\text{If } s_i > 0 \text{ or } q_i = \beta p_i, \quad q_i = \beta \pi_i = \beta(p_i + s_i(1 - R)) \qquad (26)
$$

Summing on i when $q_i \ge \beta p_i$:

$$
\begin{aligned}
\sum_{q_i/p_i \ge \beta} q_i &= \beta(\sum_{q_i/p_i \ge \beta} p_i + (1 - R) \sum_{q_i/p_i \ge \beta} s_i) \\
&= \beta(\sum_{q_i/p_i \ge \beta} p_i + (1 - R))
\end{aligned}
$$

This implies that:

$$
\begin{aligned}
1 - R &= \frac{1}{\beta} \sum_{q_i/p_i \ge \beta} q_i - \sum_{q_i/p_i \ge \beta} p_i \qquad (27) \\
&= \sum_i Relu(\frac{q_i}{\beta} - p_i) \qquad (28)
\end{aligned}
$$

We can deduce from (25) that R is a strictly decreasing continuous function of α over $[\min_i(\frac{q_i}{p_i}), 1]$, reaching 1 at $\min_i(\frac{q_i}{p_i})$ and R_0 at 1. Similarly, given (28), R is a strictly increasing continuous function of β over $[1, \max_i(\frac{q_i}{p_i})]$, reaching R_0 at 1 and 1 at $\max_i(\frac{q_i}{p_i})$. Moreover, $R = 1$ for $\alpha \le \min_i(\frac{q_i}{p_i})$ and $\beta \ge \max_i(\frac{q_i}{p_i})$.

This means that there is a single α and a single β for each R in $[R_0, 1[$, these α and β are given by (24) and (27) and these α and β entirely determine the solution to problem (8) in the following way:

$$
\begin{array}{llll}
\text{If } \dfrac{q_i}{p_i} \le \alpha, & \dfrac{q_i}{\pi_i} = \alpha, & r_i = \dfrac{q_i}{\alpha p_i}, & s_i = 0 \\
\text{If } \alpha < \dfrac{q_i}{p_i} < \beta, & \dfrac{q_i}{\pi_i} = \dfrac{q_i}{p_i}, & r_i = 1, & s_i = 0 \\
\text{If } \dfrac{q_i}{p_i} \ge \beta, & \dfrac{q_i}{\pi_i} = \beta, & r_i = 1, & s_i = \dfrac{1}{1 - R}(\dfrac{q_i}{\beta} - p_i)
\end{array}
$$

5 Both problems share some solutions

For the sake of simplicity, we rewrite in this section our two optimization problems as follows:

$$\text{Primary optimization problem (1):} \qquad \max_{x\in\Omega, D(x)\leq d} R(x)$$

$$\text{Auxiliary optimization problem (8):} \qquad \min_{x\in\Omega, R(x)\geq r} D(x)$$

... with D, the Kullback-Leibler divergence between q and π, and R, the acceptance rate.

We established in the previous sections (7)(22) that the first inequality constraints of these problems are always saturated:

$$\forall d \in [0, D_{KL}(q,p)], \max_{x\in\Omega, D(x)\leq d} R(x) = \max_{x\in\Omega, D(x)=d} R(x)$$

$$\forall r \in [R_0, 1], \min_{x\in\Omega, R(x)\geq r} D(x) = \min_{x\in\Omega, R(x)=r} D(x)$$

For any d, we now show that any $x_1 \in \Omega$, solution of (1), is also a solution of (8) for a certain r. Let x_2 be a solution of (8) for $r = R(x_1)$.

Since the inequality constraints of both problems are saturated, we know that $D(x_1) = d$ and $R(x_2) = R(x_1)$. Moreover, $D(x_2) \leq D(x_1)$ because x_1 is in the feasible region of (8) for $r = R(x_1)$.

Could $D(x_2)$ be strictly less than $D(x_1)$? If it were the case, it would be possible to modify x_2 (by increasing one of the $r_i < 1$) to strictly increase $R(x_2)$ while keeping $D(x_2) \leq D(x_1) = d$. This would invalidate x_1 as a solution of (1). Therefore, $D(x_2) = D(x_1)$ and x_1 is indeed a solution of (8) for $r = R(x_1)$. This shows that all solutions of (1) are solutions of (8).

6 Algorithm

We know from the preceding sections that we can build all solutions of our optimization problem with α varying from 0 to 1. $\alpha = 0$ corresponds to $\sum_i q_i \ln \frac{q_i}{\pi_i} = D_{KL}(q,p)$ and $\sum_i p_i r_i = 1$ while $\alpha = 1$ corresponds to $\sum_i q_i \ln \frac{q_i}{\pi_i} = 0$ and $\sum_i p_i r_i = R_0$. Moreover, $\sum_i q_i \ln \frac{q_i}{\pi_i}$ and $\sum_i p_i r_i$ are decreasing functions of α.

This means that we can compute the solution (r_i, s_i) of the optimization problem through a binary search for α over $]0, 1]$, as illustrated in Algorithm 1.

In practice, we know that $r_i \geq \min(\frac{q_i}{p_i}, 1)$ so when deciding whether to accept a draft token, we can sample $u \sim \mathcal{U}_{[0,1]}$ and only compute the corresponding r_i if $u > \min(\frac{q_i}{p_i}, 1)$ and the $(s_j)_j$ if $u > r_i$. We also do not need to compute these quantities when $D_{KL}(q,p) \leq D$ for the current token. Moreover, all the values for $(\sum_{i\leq j} q_i)_j$, $(\sum_{i\leq j} p_i)_j$, $(\sum_{i\geq j} q_i)_j$, $(\sum_{i\geq j} p_i)_j$, $(\sum_{i\leq j} q_i \ln \frac{q_i}{p_i})_j$ and $(\frac{p_j}{q_j} \sum_{i\geq j} q_i - \sum_{i\geq j} p_i)_j$ can be computed in a vectorized manner before the loop to save time.

Algorithm 1 Mentored decoding

Require:
$(q_i)_i, (p_i)_i$ (reordered so that $\frac{q_i}{p_i}$ increases with i)
$D \geq 0$ (upper bound for the Kullback-Leibler divergence)
$\gamma \in]0,1[$ (tolerance for D)

$\alpha_{\min}, \alpha, \alpha_{\max} \leftarrow 0, \frac{1}{2}, 1$

while True **do**
$P \leftarrow -\frac{1}{\alpha}\sum_{q_i/p_i \leq \alpha} q_i + \sum_{q_i/p_i \leq \alpha} p_i$ ▷ Cf. (24)
$n \leftarrow \min\{j | \frac{p_j}{q_j}\sum_{i \geq j} q_i - \sum_{i \geq j} p_i \leq P\}$ ▷ Cf. (28)*
$\beta \leftarrow (\sum_{i \geq n} q_i)/(P + \sum_{i \geq n} p_i)$ ▷ Cf. (27)
$\delta \leftarrow \ln\alpha \sum_{q_i/p_i \leq \alpha} q_i + \sum_{q_i/p_i \in]\alpha,\beta[} q_i \ln\frac{q_i}{p_i} + \ln\beta \sum_{q_i/p_i \geq \beta} q_i$
▷ Cf. the summary at the end of Section 4
if $\delta < (1-\gamma)D$ **then**
$\alpha \leftarrow (\alpha_{\min} + \alpha)/2$
$\alpha_{\max} \leftarrow \alpha$
else if $\delta > (1+\gamma)D$ **then**
$\alpha \leftarrow (\alpha + \alpha_{\max})/2$
$\alpha_{\min} \leftarrow \alpha$
else Break
end if
end while

$\forall i \quad r_i \leftarrow \min(\frac{q_i}{\alpha p_i}, 1)$ ▷ Cf. (23)
$\forall i \quad s_i \leftarrow \max(\frac{1}{P}(\frac{q_i}{\beta} - p_i), 0)$ ▷ Cf. (26)
Return $(r_i)_i, (s_i)_i$

*The function compared to P is equivalent to the right hand side of (28) with $\beta = \frac{q_j}{p_j}$. This function decreases with j because $\frac{q_j}{p_j}$ increases with j. With n defined this way, $\frac{q_i}{p_i} \geq \beta$ is equivalent to $i \geq n$.

References

[1] Charlie Chen et al. *Accelerating Large Language Model Decoding with Speculative Sampling.* 2023. arXiv: `2302.01318 [cs.CL]`.

[2] Yaniv Leviathan, Matan Kalman, and Yossi Matias. "Fast inference from transformers via speculative decoding". In: *International Conference on Machine Learning.* PMLR. 2023, pp. 19274–19286.