

For ease of reading, and although this is already covered in prior works which are cited in the main body of the paper, we provide here a full derivation of the OP-MBRD algorithm.

## Optimal Policy

The decoding rule is defined as

$$\hat{y} = \arg \max_{y' \in V^+} \underbrace{\{\mathbb{E}_{p^*(y|x)} \{M(y, y', x)\}\}}_{Q(y', x)}$$

for a  $p^*(y | x)$  satisfying

$$p^*(y | x) = \arg \max_q \{\mathcal{L}(q)\}$$

with

$$\mathcal{L}(q) = \mathbb{E}_{q(y|x)} \{R(y, x)\} - \beta \cdot \text{KL}(q(y | x) || p_R(y | x))$$

where KL is the Kullback-Leibler (KL) divergence and  $p_R(y | x)$  is a reference policy we want to stay close to. For the purpose of optimization, we can multiply this objective by  $1/\beta$  and reformulate it by applying the definition of KL divergence

$$\begin{aligned} \mathcal{L}(q)/\beta &= \mathbb{E}_{q(y|x)} \left\{ \frac{R(y, x)}{\beta} \right\} - \text{KL}(q(y | x) || p_R(y | x)) \\ &= \mathbb{E}_{q(y|x)} \left\{ \frac{R(y, x)}{\beta} + \log \frac{p_R(y | x)}{q(y | x)} \right\} \\ &= \mathbb{E}_{q(y|x)} \left\{ \log \frac{p_R(y | x) \cdot \exp\left(\frac{R(y, x)}{\beta}\right)}{q(y | x)} \right\} \\ &= \mathbb{E}_{q(y|x)} \left\{ \log \frac{\frac{p_R(y|x) \cdot \exp\left(\frac{R(y, x)}{\beta}\right)}{Z}}{q(y | x)} \right\} + \log Z(x) \\ &= -\text{KL} \left( q(y | x) || \frac{p_R(y | x) \cdot \exp\left(\frac{R(y, x)}{\beta}\right)}{Z} \right) + \log Z(x) \end{aligned}$$

where we have summed and subtracted a partition function to obtain a distance between distributions, which crucially does not depend on  $q$

$$\log Z = \sum_{y \in V^+} p_R(y | x) \cdot \exp\left(\frac{R(y, x)}{\beta}\right)$$

due to the properties of the KL divergence,  $\mathcal{L}(q)$  is maximized ( $= 0$ ) when the two distributions in the KL match i.e.

$$p^*(y | x) = \arg \max_q \{\mathcal{L}(q)\} = \frac{p_R(y | x) \cdot \exp\left(\frac{R(y, x)}{\beta}\right)}{Z}$$

## Rao-Blackwellized Rejection Sampling

For the purposes of solving MBRD, we can approximate the expectation with the following estimator

$$\hat{Q}(y', x) = \frac{1}{N} \sum_{y_n \in S(N)} M(y_n, y', x)$$

where  $S(N) = \{y_1 \cdots y_n\} \sim p^*(y | x)$ . It rests to sample from this distribution to obtain the estimator. For this we use statistical rejection sampling, also known as accept-reject sampling. This uses the fact that if we sample from any proposal  $y \sim p(y | x)$  the probability of that sample belonging to  $p^*(y | x)$  is given by

$$p(\text{accept } y_n) = \frac{p^*(y | x)}{M \cdot p(y | x)}$$

where  $M$  is a scaling factor to ensure the ratio is a valid probability. The tightest scaling factor is given by

$$M = \max_y \left\{ \frac{p^*(y | x)}{p(y | x)} \right\}$$

replacing Eq. [appendixop] into Eq [acceptratio], this can be reformulated as

$$p(\text{accept } y_n) = \exp \left( \frac{R(y, x)}{\beta} + \log \frac{p_R(y | x)}{p(y | x)} - \log M \right)$$

with

$$\begin{aligned} M &= \max_y \left\{ \exp \left( \frac{R(y, x)}{\beta} + \log \frac{p_R(y | x)}{p(y | x)} \right) \right\} \\ &= \exp \left( \max_y \left\{ \frac{R(y, x)}{\beta} + \log \frac{p_R(y | x)}{p(y | x)} \right\} \right) \end{aligned}$$

renaming as

$$\tilde{R}(y, x) = \frac{R(y, x)}{\beta} + \log \frac{p_R(y | x)}{p(y | x)}$$

yields

$$p(\text{accept } y_n) = \exp \left( \tilde{R}(y_n, x) - \max_{y'} \tilde{R}(y', x) \right)$$

A naive estimator based on these probabilities can be built using an indicator function and samples from a uniform distribution  $u_n \sim U(0, 1)$

$$I_n = \delta_{p(\text{accept } y_n) \geq u_n}$$

which will select a set of samples from the proposal  $p(y \mid x)$  using the acceptance probabilities. This results in

$$\hat{Q}(y', x)^{\text{RS}} = \frac{1}{N} \sum_{y \in S(N)} I_n \cdot M(y_n, y', x)$$

However, the Rao-Blackwell estimator based on the sufficient statistic  $p_1 \cdots p_N$ , which involves all samples rather than just the accepted ones, can be proven to be equal or lower variance.

$$\hat{Q}(y', x)^{\text{OP}} = \mathbb{E}\{Q(y', x) \mid p_1 \cdots p_N\}$$

For rejection sampling this can be approximated by self-normalized importance sampling using the relation on Eq [acceptratio]. For normal importance sampling we have that

$$\begin{aligned} Q(y', x) &= \mathbb{E}_{p^*(y|x)}\{M(y, y', x)\} \\ &= \sum_{y \in V^+} p^*(y \mid x) \cdot M(y, y', x) \\ &= \sum_{y \in V^+} p(y \mid x) \cdot M(y, y', x) \cdot \frac{p^*(y \mid x)}{p(y \mid x)} \\ &= \sum_{y \in V^+} p(y \mid x) \cdot M(y, y', x) \cdot p(\text{accept } y) \cdot M \end{aligned}$$

with the scaling factor  $M$  canceling after self-normalization leading to the OP-MBRD estimator expression

$$\begin{aligned} \hat{Q}(y', x)^{\text{OP}} &= \sum_{y_n \in S(N)} M(y_n, y', x) \cdot \frac{p(\text{accept } y_n)}{\sum_{y_n \in S(N)} p(\text{accept } y_n)} \\ &= \sum_{y_n \in S(N)} M(y_n, y', x) \cdot \frac{\exp(\tilde{R}(y_n, x) - \max_{y''} \tilde{R}(y'', x))}{\sum_{y_m \in S(N)} \exp(\tilde{R}(y_m, x) - \max_{y''} \tilde{R}(y'', x))} \end{aligned}$$

## Importance Sampling Equivalence and Expected Number of Optimal Policy Samples

One can also use importance sampling directly in the place of rejection sampling, for the non normalized version

$$\begin{aligned}
Q(y', x) &= \mathbb{E}_{p^*(y|x)} \{M(y, y', x)\} \\
&= \sum_{y \in V^+} p^*(y | x) \cdot M(y, y', x) \\
&= \sum_{y \in V^+} p(y | x) \cdot M(y, y', x) \cdot \frac{p_R(y | x) \cdot \exp\left(\frac{R(y, x)}{\beta}\right)}{Z \cdot p(y | x)} \\
&= \sum_{y \in V^+} p(y | x) \cdot M(y, y', x) \cdot \frac{1}{Z} \cdot \exp\left(\frac{R(y, x)}{\beta} + \log \frac{p_R(y | x)}{p(y | x)}\right)
\end{aligned}$$

after self-normalization this leads to

$$\hat{Q}(y', x) = \sum_{y_n \in S(N)} M(y_n, y', x) \cdot \frac{\exp(\tilde{R}(y_n, x))}{\sum_{y_m \in S(N)} \exp(\tilde{R}(y_m, x))}$$

which only differs from the previous estimator by a constant inside of the exponential (the scaling factor), which cancels out. Both estimator are therefore identical.

One important difference in practice is that we can derive the expected number of samples that are accepted, which is used in OPE-MBRD as the stopping condition.

$$\hat{N}^{\text{OP}} = \sum_{y_n \in S(N)} p_n = \sum_{y_n \in S(N)} \exp\left(\tilde{R}(y_n, x) - \max_{y'} \tilde{R}(y', x)\right).$$

## PRM Version

The estimator presented above could be used both with Outcome Reward Models (ORM), and Process Reward Models (PRM). We focused on the latter for our experimental set up and we also set  $p_R(y | x) = p(y | x)$ . PRMs compute one score per step of thought and thus we need to compute the probabilities for each

$$p(\text{accept } y_{<t+1}^n) = \exp\left(\frac{\text{PRM}(y_{<t+1}^n, x)}{\beta} - \max_{\tilde{y}_{<t+1}^n} \frac{\text{PRM}(\tilde{y}_{<t+1}^n, x)}{\beta}\right)$$

This leads to a special case detailed here.

$$\tilde{R}(y, x) = \frac{\text{PRM}(y_{<t+1}^n, x)}{\beta} + \underbrace{\log \frac{p(y | x)}{p(y | x)}}_{=0}$$

since PRMs are bounded by 1.0 we used this value as the maximum and normalized by the length, leading to

$$p(\text{accept } y^n) = \prod_{t=1}^T p(\text{accept } y_{<t+1}^n)^{1/T} = \exp \left( \frac{1}{\beta} \cdot \frac{1}{T} \cdot \sum_{t=1}^T (\text{PRM}(y_{<t+1}^n, x) - 1) \right)$$