

TTIC 31230 Fundamentals of Deep Learning

RL Problems.

Problem 1. Consider training machine translation on a corpus of translation pairs (x, y) where x is, say, an English sentence x_1, \dots, EOS and y is a French sentence y_1, \dots, EOS where EOS is the “end of sentence” tag.

Suppose that we have a parameterized model defining $P_\Phi(y_t|x, y_1, \dots, y_{t-1})$ so that $P_\Phi(y_1, \dots, y_T|x) = \prod_{t=1}^{T'} P_\Phi(y_t|x, y_1, \dots, y_{t-1})$ where y_T is EOS.

For a sample \hat{y} from $P_\Phi(y|x)$ we have a non-differentiable BLEU score $\text{BLEU}(\hat{y}, y) \geq 0$ that is not computed until the entire output y is complete and which we would like to maximize.

(a) Give an SGD update equation for the parameters Φ for the REINFORCE algorithm for maximizing $E_{\hat{y} \sim P_\Phi(y|x)}$ for this problem.

Solution: For $\langle x, y \rangle$ samples form the training corpus of translation pairs, and for $\hat{y}_1, \dots, \hat{y}_T$ sampled from $P_\Phi(\hat{y}|x)$ we update Φ by

$$\Phi \leftarrow \Phi + \eta \text{BLEU}(\hat{y}, y) \sum_{t=1}^T \nabla_\Phi \ln P_\Phi(\hat{y}_t|x, \hat{y}_1, \dots, \hat{y}_{t-1})$$

Samples with higher BLEU scores have their probabilities increased.

(b) Suppose that somehow we reach a parameter setting Φ where $P_\Phi(y|x)$ assigns probability close enough to 1 for a particular translation \hat{y} that in practice we will always sample the same \hat{y} . Suppose that this translation \hat{y} has less than optimal BLEU score. Can the REINFORCE algorithm recover from this situation and consider other translations? Explain your answer.

Solution: No. The REINFORCE algorithm will not recover. The update will only increase the probability of the single translation which it always selects. A deterministic policy has zero gradient and is stuck.

(c) Show that for any function $V(x)$ we have

$$E_{\hat{y} \sim P_\Phi(\hat{y}|x)} V(x) \nabla_\Phi \ln P_\Phi(\hat{y}_t|x, y_1, \dots, y_{t-1}) = 0$$

Solution:

$$\begin{aligned}
& E_{\hat{y}} V(x) \nabla_{\Phi} \ln P_{\Phi}(\hat{y}_t | x, y_1, \dots, y_{t-1}) \\
&= V(x) E_{\hat{y}_1, \dots, \hat{y}_{t-1}} \sum_{\hat{y}_t} P_{\Phi}(\hat{y}_t | x, \hat{y}_1, \dots, \hat{y}_{t-1}) \frac{\nabla_{\Phi} P_{\Phi}(\hat{y}_t | x, \hat{y}_1, \dots, \hat{y}_{t-1})}{P_{\Phi}(\hat{y}_t | x, \hat{y}_1, \dots, \hat{y}_{t-1})} \\
&= V(x) E_{\hat{y}_1, \dots, \hat{y}_{t-1}} \sum_{\hat{y}_t} P_{\Phi}(\hat{y}_t | x, \hat{y}_1, \dots, \hat{y}_{t-1}) \frac{\nabla_{\Phi} P_{\Phi}(\hat{y}_t | x, \hat{y}_1, \dots, \hat{y}_{t-1})}{P_{\Phi}(\hat{y}_t | x, \hat{y}_1, \dots, \hat{y}_{t-1})} \\
&= V(x) E_{\hat{y}_1, \dots, \hat{y}_{t-1}} \nabla_{\Phi} \sum_{\hat{y}_t} P_{\Phi}(\hat{y}_t | x, \hat{y}_1, \dots, \hat{y}_{t-1}) \\
&= 0
\end{aligned}$$

(d) Modify the REINFORCE update equations to use a value function approximation $V_{\Phi}(x)$ to reduce the variance in the gradient samples and where V_{Φ} is trained by Bellman Error. Your equations should include updates to train $V_{\Phi}(x)$ to predict $E_{\hat{y} \sim P(y|x)} \text{BLEU}(\hat{y}, y)$. (Replace the reward by the “advantage” of the particular translation).

Solution: For $\langle x, y \rangle$ sampled from the training corpus of translation pairs, and for $\hat{y}_1, \dots, \hat{y}_T$ sampled from $P_{\Phi}(\hat{y} | x)$ we update Φ by

$$\begin{aligned}
\Phi & += \eta (\text{BLEU}(\hat{y}, y) - V_{\Phi}(x)) \sum_{t=1}^T \nabla_{\Phi} \ln P_{\Phi}(\hat{y}_t | x, \hat{y}_1, \dots, \hat{y}_{t-1}) \\
\Phi & -= 2\eta (V_{\Phi}(x) - \text{BLEU}(\hat{y}, y))
\end{aligned}$$