

TTIC 31230 Fundamentals of Deep Learning

RL Problems.

Problem 1. REINFORCE for BLEU Translation Score. 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 autoregressive 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 very close to 1 for a particular translation \hat{y} so 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) 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 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

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

Problem 2. Rapid Mixing for Asymptotic Average Reward.

We consider a case where we are interested in asymptotic average reward.

$$R(\pi) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T r_t$$

For a given policy π we have a Markov process over states — a well defined state transition probability $P_\pi(s_{t+1}|s_t)$ defined by

$$P_\pi(s_{t+1}|s_t) = \sum_a \pi(a|s_t) P_\pi(s_{t+1}|s_t, a)$$

Under very mild conditions a Markov process has a well define stationary distribution on states which we will write $P_\pi(s)$. This distribution is “stationary” in the sense that

$$\sum_{s_1} P_\pi(s_1) P_\pi(s_2|s_1) = P_\pi(s_2)$$

(a) Write the asymptotic average reward $R(\pi)$ in terms of the stationary distribution P_π , the policy $\pi(a|s)$ and the reward function $R(s, a)$

Solution:

$$R(\pi) = E_{s \sim P_\pi(s), a \sim \pi(a|s)} R(s, a)$$

(b) Now for $\Delta t > 1$ we define $P_\pi(s_{t+\Delta t}|s_t)$ recursively as by

$$P_\pi(s_{t+\Delta t}|s_t) = \sum_{s_{t+\Delta t-1}} P_\pi(s_{t+\Delta t-1}|s_t) P_\pi(s_{t+\Delta t}|s_{t+\Delta t-1})$$

We now assume a “mixing parameter” $0 < \gamma < 1$ for π defined by the property

$$\sum_{s_{t+\Delta t}} |P_\pi(s_{t+\Delta t}|s_t) - P_\pi(s_{t+\Delta t})| \leq \gamma^{\Delta t}$$

We now define an advantage function on state-action pairs to be the “extra” reward we get by taking action a (rather than drawing from $\pi(a|s)$) summed over all time.

$$A(s, a) = E \sum_{t=0}^{\infty} (r_t - R(\pi)) \mid s_0 = s, a_0 = a$$

Assuming the reward is bounded by r_{\max} and that we have the above mixing parameter γ , give a (finite) upper bound on the infinite sum $A(s, a)$.

Solution:

$$\begin{aligned}
& E[r_t - R(\pi) \mid s_0 = s, a_0 = a, t > 0] \\
&= \left(\sum_{s_t} P_\pi(s_t | s_0) E_{a \sim \pi(a | s_t)} R(s_t, a) \right) - R(\pi) \\
&= \left(\sum_{s_t} (P_\pi(s_t) + P_\pi(s_t | s_0) - P_\pi(s_t)) E_{a \sim \pi(a | s_t)} R(s_t, a) \right) - R(\pi) \\
&= R(\pi) + \left(\sum_{s_t} (P_\pi(s_t | s_0) - P_\pi(s_t)) E_{a \sim \pi(a | s_t)} R(s_t, a) \right) - R(\pi) \\
&= \sum_{s_t} (P_\pi(s_t | s_0) - P_\pi(s_t)) r_{\max} \\
&\leq r_{\max} \gamma^t \\
A(s, a) &\leq r_{\max} \sum_{t=0}^{\infty} \gamma^t = \frac{r_{\max}}{1 - \gamma}
\end{aligned}$$

It can be shown that

$$\nabla_\Phi R(\pi_\Phi) = E_{s \sim P_\pi(s), a \sim \pi(a | s)} \nabla_\Phi \ln \pi_\Phi(a | s) A(s, a)$$

You do not have to prove this.