The Cross-Entropy Method

A variance reduction method for importance sampling

Hugo Berard

Reinforcement Learning (COMP-762), Winter 2017

Importance Sampling

We want to estimate:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi[G(X_t)|S_t = s] \\ &= \mathbb{E}_\mu[G(X_t)W(X_t, \pi, \mu)|S_t = s] \end{aligned}$$

with $W(X_t, \pi, \mu) = \prod_{k=t}^{T(t)} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} = \frac{f(X_t)}{g(X_t)}$ and $X_t = (A_t, S_{t+1}, A_{t+1}, ...S_{T(t)})$ sampled from μ

Using **importance sampling** we get:

$$\hat{V}^\pi(s) = \frac{1}{|\mathcal{T}(s)|} \sum_{t \in \mathcal{T}(s)} G(X_t)W(X_t, \pi, \mu)$$

The **variance** of this estimator is:

$$var(\hat{V}^\pi(s)) = \frac{1}{|\mathcal{T}(s)|} var(G(X_t)W(X_t, \pi, \mu))$$

Idea: Can we find a behavior policy that **minimize** this variance ?

Variance Minimization

We want to find:

$$g^* = \min_g var_g(G(X)\frac{f(X)}{g(X)})$$

We can easily show that the solution is:

$$g^*(x) = \frac{|G(x)|f(x)}{\int |G(x)|f(x)dx}$$

For convenience let's define some **parametric** family of distribution:

$$\mathcal{F} = \{f(.;v), v \in \mathcal{V}\}$$

with $f = f(.;u) \in \mathcal{F}$

The Cross Entropy Method

Let's define a measure between two distributions g and h:

$$\mathcal{D}_{\mathcal{KL}}(g||h) = \mathbb{E}_g[log(\frac{g(X)}{h(X)}] \tag{1}$$

$$= \int g(x)log(g(x))dx - \int g(x)log(h(x))dx \tag{2}$$

this measure is known as the **Kullback-Leibler Divergence** (Cross-Entropy).

we want to find:

$$\min_v \mathcal{D}_{\mathcal{KL}}(g^*, f(.;v))$$

this is minimum for $\mathcal{D}_{\mathcal{KL}} = 0$:

$$g^*(x) = f(x;v^*) = \frac{|G(x)|f(x;u)}{\int |G(x)|f(x;u)dx} \tag{3}$$

Cross-Entropy Method applied to Policy Prediction

Let $X_t = (A_t, S_{t+1}, A_{t+1}, ...S_{T(t)})$ be an episode sampled from a policy μ_u, where for each state the distribution is a **categorical** distribution over the actions :

$$f(x|s;u) = \sum_i u_i(s) I_{x=a_i}$$

We can show that:

$$v_i^*(s) = \frac{\mathbb{E}_u[G(X) I_{X=a_i}]}{\mathbb{E}_u[G(X)]} = \frac{\mathbb{E}_w[G(X)W(X,u,w) I_{X=a_i}]}{\mathbb{E}_w[G(X)W(X,u,w)]} \tag{4}$$

This can be estimated through importance sampling again:

$$\hat{v}_i(s) = \frac{\sum_{t \in \mathcal{T}(s)} G_t W_t(u,w) I_{X_t=a_i}}{\sum_{t \in \mathcal{T}(s)} G_t W_t(u,w)}$$

Algorithm for Policy Prediction

Initialize for all $s \in \mathcal{S}, a \in \mathcal{A}$:

$\pi_u \leftarrow$ the target policy
$\mu_w \leftarrow$ an arbitrary behavior policy (eg. uniform)
$V(s) \leftarrow$ an arbitrary state value function
$C \leftarrow 0$

Repeat:

Repeat:

Generate episode using μ_w
$G \leftarrow 0$
$W \leftarrow 1$
For $t = T-1, T-2, ...downto 0$:
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$v_i(S_t) \leftarrow v_i(S_t, A_t) + \frac{W}{C(S_t,A_t)}(GI(A_t = a_i) - v_i(S_t))$
$W \leftarrow W\frac{\pi_u(A_t|S_t)}{\mu_w(A_t|S_t)}$
if $W = 0$ then ExitForLoop

Algorithm for Policy Evaluation (Continue)

Generate episode using μ_v

$G \leftarrow 0$

$W \leftarrow 1$

For $t = T - 1, T - 2, ... downto 0$:

$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$V(S_t) \leftarrow V(S_t) + \frac{W}{C(S_t, A_t)}(G - v_i(S_t))$
$W \leftarrow W \frac{\pi_u(A_t|S_t)}{\mu_v(A_t|S_t)}$
if $W = 0$ then ExitForLoop

Appendix: Proof o f Eq.4

From Eq.3, we have:

$$\begin{aligned} f(x|s;v^*) &= \frac{|G(x)|f(x;u)}{\int |G(x)|f(x;u)dx} \\ &= \frac{\sum_i |G_t|u_i(s)I_{x=a_i}}{\mathbb{E}_u[|G_t|]} \\ &= \sum_i \frac{|G_t|u_i(s)}{\mathbb{E}_u[|G_t|]} I_{x=a_i} \\ &= \sum_i \end{aligned}$$