CME 241 Kevin Monogue (kmonogue@stanford.edu)

Score Function

Score is defined as the gradient of the log likelihood function, $\nabla log(L(X;\theta))$ or in our case with regards to a parametric policy $\nabla log(\pi(s,a;\theta))$. We will derive the score function for a couple common functions.

Softmax

The Softmax function is defined in the setting of a policy as

$$\pi(s,a;\theta) = \frac{e^{\theta\cdot\phi(s,a)}}{\sum_{a'\in A} e^{\theta\cdot\phi(s,a')}}$$

We solve for the score by taking the log of the function and differentiating.

$$\begin{aligned}
log(\pi(s,a;\theta)) &= log(\frac{(e^{\theta\cdot\phi(s,a)}}{\sum_{a'\in A} e^{\theta\cdot\phi(s,a')}}) \\
&= log(e^{\theta\cdot\phi(s,a)}) - log(\sum_{a'\in A} e^{\theta\cdot\phi(s,a')}) \\
&= \theta\cdot\phi(s,a) - log(\sum_{a'\in A} e^{\theta\cdot\phi(s,a')}) \\
\nabla log(\pi(s,a;\theta)) &= \phi(s,a) - \nabla_\theta log(\sum_{a'\in A} e^{\theta\cdot\phi(s,a')}) \\
&= \phi(s,a) - \frac{\sum_{a'\in A}\phi(s,a')\cdot e^{\theta\cdot\phi(s,a')}}{\sum_{a'\in A} e^{\theta\cdot\phi(s,a')}} \\
&= \phi(s,a) - \sum_{a'\in A}\pi(s,a';\theta)\cdot\phi(s,a') \\
&= \phi(s,a) - E_\pi[\phi(s,)]
\end{aligned}$$

Note the second to last line simply uses the definition of the policy function from above.

Gaussian Normal

Our policy can also be defined by a continuous distrubition, such as a normal distribution. In this case, the probability, $a \sim N(\theta\cdot\phi(s),\sigma^2)$. The policy is therefore just the pdf and we can solve as normal.

$$\begin{aligned}
\pi(a,s;\theta) &= \frac{1}{\sigma\sqrt(2\pi)}\cdot e^{-\frac{1}{2}\cdot(\frac{a-\phi(s)\cdot\theta}{\sigma})^2} \\
log(\pi) &= log(\frac{1}{\sigma\sqrt(2\pi)}) + log(e^{-\frac{1}{2}\cdot(\frac{a-\phi(s)\cdot\theta}{\sigma})^2}) \\
&= log(\frac{1}{\sigma\sqrt(2\pi)}) - \frac{1}{2}\cdot(\frac{a-\phi(s)\cdot\theta}{\sigma})^2 \\
\nabla_\theta log(\pi(s,a;\theta) &= 0 + -\frac{1}{2}\cdot 2\cdot -\phi(s)\cdot\frac{a-\theta\cdot\phi(s)}{\sigma^2} \\
&= \phi(s)\cdot\frac{a-\theta\cdot\phi(s)}{\sigma^2}
\end{aligned}$$