Policy Gradient Methods

- Policy depends on some parameters Θ
 - Action preferences
 - Mean and variance
 - Weights of a neural network
- Modify policy parameters directly instead of estimating the action values
- Maximize: $\eta(\Theta) = E(\mathbf{r})$ $= \sum_{a} Q^{*}(a) \cdot \pi(\Theta, a)$ $\Theta \leftarrow \Theta + \alpha \cdot \nabla \eta(\Theta)$

• Taking gradients:

$$\nabla \eta(\Theta) = \sum_{a} Q^{*}(a) \cdot \nabla \pi(\Theta, a)$$

• Rewriting:

$$\nabla \eta(\Theta) = \sum_{a} Q^{*}(a) \cdot \frac{\nabla \pi(\Theta, a)}{\pi(\Theta, a)} \pi(\Theta, a)$$

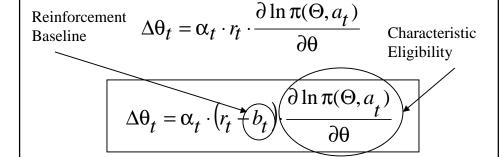
• Estimate gradient given *N* samples:

$$\hat{\nabla}\eta(\Theta) = \frac{1}{N} \sum_{t=1}^{N} r_t \cdot \frac{\nabla \pi(\Theta, a_t)}{\pi(\Theta, a_t)}$$

REINFORCE (Williams '92)

• Incremental version:

$$\Delta \theta_t = \alpha_t \cdot r_t \cdot \frac{\nabla \pi(\Theta, a_t)}{\pi(\Theta, a_t)}$$



Special case – Generalized L_{R-I}

• Consider binary bandit problems with arbitrary rewards

$$\pi(\theta, a) = \begin{cases} \theta & \text{if } a = 1 \\ 1 - \theta & \text{if } a = 0 \end{cases} \frac{\partial \ln \pi}{\partial \theta} = \frac{a - \theta}{\theta(1 - \theta)}$$

$$b = 0$$
 and $\alpha = \rho \cdot \theta(1 - \theta)$

$$\Delta\theta = \rho \cdot r \cdot (a - \theta)$$

Reinforcement Comparison

• Set baseline to average of observed rewards

$$b_t = \overline{r}_t = \overline{r}_{t-1} + \beta \cdot (r_t - \overline{r}_{t-1})$$

• Softmax action selection

$$\Delta \theta_i = \alpha \cdot (r - \overline{r})(1 - \pi(\Theta, a_i))$$

$$\pi (\Theta, a_i) = \frac{e^{\theta_i}}{\sum_{j=1}^n e^{\theta_j}}$$
Computation of characteristic eligibility for softmax action selection
$$\frac{\partial \ln \pi(\Theta, a_i)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \ln \frac{e^{\theta_i}}{\sum_{j=1}^n e^{\theta_j}}$$

$$= \frac{\partial}{\partial \theta_i} (\theta_i - \ln(\sum_{j=1}^n e^{\theta_j}))$$

$$= 1 - \frac{e^{\theta_i}}{\sum_{j=1}^n e^{\theta_j}}$$

$$= 1 - \pi(\Theta, a_i)$$

Continuous Actions

• Use a Gaussian distribution to select actions

$$\pi(a,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(a-\mu)^2}{2\sigma^2}}$$

• For suitable choice of parameters:

$$\Delta \mu = \alpha \cdot (r - \overline{r})(a - \mu)$$
$$\Delta \sigma = (\alpha / \sigma) \cdot (r - \overline{r})((a - \mu)^2 - \sigma^2)$$