Robot Learning

Policy Gradient

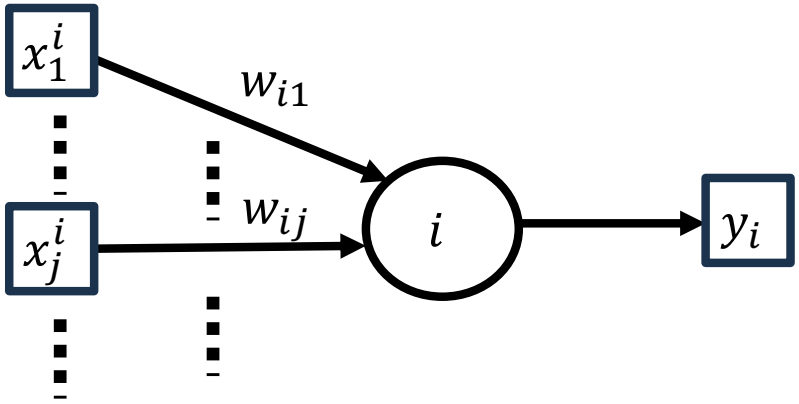
Prof. Songhwai Oh
ECE, SNU

R. J. Williams, "**Simple statistical gradient-following algorithms for connectionist reinforcement learning**," Machine Learning, 1992.

REINFORCE

Reinforcement Learning Connectionist Networks

Feedforward network



$$x^i = (x_1^i, x_2^i, \ldots) \quad : \text{input}$$
$$w^i = (w_{i1}, w_{i2}, \ldots) \quad : \text{weight}$$
$$W = (w^1, w^2, \ldots)$$

- $y_i = f(x^i, w^i)$, where f is an activation function (e.g., logistic function)
- $g_i(k, w^i, x^i) = P(y_i = k|W, x^i)$: policy function
- r: reinforcement signal (reward)
- **Goal**: Update the weights W (incrementally) to maximize the expected reward

REINFORCE Algorithm

$$\Delta w_{ij} = \alpha_{ij} \left(r - b_{ij}\right) e_{ij} = \alpha_{ij} \left(r - b_{ij}\right) \frac{\partial \log g_i}{\partial w_{ij}}$$

- α_{ij}: learning rate
- r: reward signal
- b_{ij}: reinforcement baseline (can be chosen to minimize estimate variance)
- $e_{ij} = \partial \log g_i / \partial w_{ij}$: characteristic eligibility
- REINFORCE: REward Increment = Nonnegative Factor $\times$ Offset Reinforcement $\times$ Characteristic Eligibility

REINFORCE Algorithm

$$\Delta w_{ij} = \alpha_{ij} \left(r - b_{ij}\right) e_{ij} = \alpha_{ij} \left(r - b_{ij}\right) \frac{\partial \log g_i}{\partial w_{ij}}$$

Theorem. If $\alpha_{ij} = \alpha$, then $\mathbb{E}\left(\Delta W | W\right) = \alpha \nabla_w \mathbb{E}\left(r | W\right)$.

- $\mathbb{E}\left(r | W\right)$: expected reward
- By updating weights by ΔW, we can increase the expected return
- The proof uses the following fact

$$e_{ij} = \frac{\partial \log g_i}{\partial w_{ij}} = \frac{1}{g_i} \frac{\partial g_i}{\partial w_{ij}} \qquad \text{general form in RL: } \frac{\partial \log \pi(\theta)}{\partial \theta} = \frac{1}{\pi(\theta)} \frac{\partial \pi(\theta)}{\partial \theta}$$

REINFORCE Algorithm

$$\Delta w_{ij} = \alpha_{ij} \left(r - b_{ij} \right) e_{ij} = \alpha_{ij} \left(r - b_{ij} \right) \frac{\partial \log g_i}{\partial w_{ij}}$$

Theorem. If $\alpha_{ij} = \alpha$, then $\mathbb{E}\left(\Delta W | W\right) = \alpha \nabla_w \mathbb{E}\left(r | W\right)$.

Proof

$$\begin{aligned}
\mathbb{E}(r|W) &= \int r(y) P(y|W,x) dy \\
\nabla \mathbb{E}(r|W) &= \int r(y) \nabla P(y|W,x) dy \\
&= \int r(y) \frac{\nabla P(y|W,x)}{P(y|W,x)} P(y|W,x) dy \\
&= \int \left(r(y) \nabla \log P(y|W,x) \right) P(y|W,x) dy \\
&= \mathbb{E}\left(r \nabla \log g_i(\cdot|W,x) \right)
\end{aligned}$$

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "**Policy gradient methods for reinforcement learning with function approximation**," Advances in Neural Information Processing Systems (NIPS), Nov. 2000.

POLICY GRADIENT

Notations

- $\rho(\pi) = \mathbb{E}\left(\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_0, \pi\right)$, value, expected reward-to-go
  - γ, discount factor
  - r_t, reward
  - s_o, initial state
  - π, policy
- $Q^{\pi}(s, a) = \mathbb{E}\left(\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, a_t = a, \pi\right)$
- $d^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s | s_0, \pi)$, discounted state occupancy probability

Policy Gradient

Theorem (Policy Gradient). For any MDP,

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \underbrace{\sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)}_{L(s)}.$$

- θ, parameters for policy π
- If $s \sim \pi$, sample $s_1, s_2, \ldots, s_M$. Then $\frac{\partial \rho}{\partial \theta}$ can be approximated by $\frac{1}{M} \sum_{m=1}^{M} L(s_m)$.

Connection to REINFORCE

- From the policy gradient theorem,

$$\Delta\theta \propto \frac{\partial \pi(s,a)}{\partial \theta} Q^{\pi}(s,a),$$

where

$$Q^{\pi}(s,a) \approx \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} =: R_t.$$

- Then we have the REINFORCE update rule

$$\Delta\theta \propto \frac{\partial \pi(s,a)}{\partial \theta} R_t \cdot \frac{1}{\pi(s,a)},$$

where $\frac{1}{\pi(s,a)}$ corrects for oversampling actions preferred by π.

Function Approximation

- f_w to approximate Q^π
- Update w by

$$\begin{aligned}\Delta w_t &\propto \frac{\partial}{\partial w}\left(\hat{Q}^\pi(s_t,a_t) - f_w(s_t,a_t)\right)^2 \\ &\propto \left(\hat{Q}^\pi(s_t,a_t) - f_w(s_t,a_t)\right)\frac{\partial f_w}{\partial w}\end{aligned}$$

  - $\hat{Q}^\pi$, estimate for Q^π, e.g., $R_t = \sum_{k=1}^{\infty}\gamma^{k-1}r_{t+k}$

- If $\frac{\partial f_w(s,a)}{\partial w} = \frac{\partial \pi(s,a)}{\partial \theta}\frac{1}{\pi(s,a)}$, then

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s)\sum_a \frac{\partial \pi(s,a)}{\partial \theta} f_w(s,a)$$

- due to the optimality condition (local)

$$\sum_s d^\pi(s)\sum_a \pi(s,a)\left(Q^\pi(s,a) - f_w(s,a)\right)\frac{\partial f_w(s,a)}{\partial w} = 0.$$

Convergence to a locally optimal policy

- $w_k = w$ such that $\sum_s d^\pi(s) \sum_a \pi_k(s,a) \left(Q^{\pi_k}(s,a) - f_w(s,a)\right) \frac{\partial f_w(s,a)}{\partial w} = 0$
- $\theta_{k+1} = \theta_k + \alpha_k \sum_s d^\pi(s) \sum_a \frac{\partial \pi_k(s,a)}{\partial \theta} f_{w_k}(s,a)$
- Conditions: $\lim_{k\to\infty} \alpha_k = 0$, $\sum_k \alpha_k = \infty$, $\pi_k = \pi(\cdot,\cdot;\theta_k)$
- Then with some additional technical conditions,

$$\lim_{k\to\infty} \frac{\partial \rho(\pi_k)}{\partial \theta} = 0$$

the local maximum.

- Consider a policy function that is a Gibbs distribution in a linear combination of features, $\pi(s,a) = \frac{e^{\theta^T \phi_{sa}}}{\sum_b e^{\theta^T \phi_{sb}}}$
  - ϕ_{sa}, feature for (s,a)
  - θ, parameters
- Then

$$\frac{\partial \pi(s,a)}{\partial \theta} = \frac{\phi_{sa} e^{\theta^T \phi_{sa}}}{\sum_b e^{\theta^T \phi_{sb}}} - \frac{e^{\theta^T \phi_{sa}}}{\left(\sum_b e^{\theta^T \phi_{sb}}\right)^2} \left(\sum_b \phi_{sb} e^{\theta^T \phi_{sb}} \right)$$

$$= \phi_{sa} \pi(s,a) - \pi(s,a) \sum_b \phi_{sb} \pi(s,b)$$

$$\frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)} = \phi_{sa} - \sum_b \phi_{sb} \pi(s,b)$$

- We have $\frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)} = \phi_{sa} - \sum_b \phi_{sb} \pi(s,b)$
- Since it requires

$$\frac{\partial f_w(s,a)}{\partial w} = \frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)}$$
$$= \phi_{sa} - \sum_b \phi_{sb} \pi(s,b)$$
$$f_w(s,a) = w^T \left(\phi_{sa} - \sum_b \phi_{sb} \pi(s,b) \right)$$

- Which implies $\sum_a \pi(s,a) f_w(s,a) = 0$ for all s
  - $f_w \approx A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$
  - $A^\pi(s,a)$, advantage function
  - Justifies the advantage function as the target for value function approximation

Baxter, Jonathan, and Peter L. Bartlett. "**Reinforcement learning in POMDP's via direct gradient ascent**." ICML. 2000.

J. Baxter and P. L. Bartlett. **Infinite-horizon policy-gradient estimation**. Journal of Artificial Intelligence Research, 15:319--350, 2001.

GPOMDP

Finite POMDP

- $\mathcal{S} = \{1, 2, \ldots, n\}$, states
- $\mathcal{U} = \{1, 2, \ldots, N\}$, actions or controls
- $\mathcal{Y} = \{1, 2, \ldots, M\}$, observations
- $r(i)$, reward at state $i \in \mathcal{S}$
- $P(u) = [P_{ij}(u)]$, where $P_{ij}(u)$ is the probability of moving from i to j given control u
- $\nu(i)$, distribution for observations $\nu_y(i) = P(\text{obs} = y | \text{state} = i)$
- $\mu(y)$, policy, i.e., distribution of actions for observation y, $\mu_u(y) = P(\text{action} = u | \text{obs} = y)$, where y can be the entire observation history
- μ and ν forms a Markov chain. Transition from i to j: $y \sim \nu(i)$, $u \sim \mu(y)$, and $j \sim P_{ij}(u)$
- $\mu(\theta, y)$, parameterized policy
  - Forms a Markov chain with transition matrix $P(\theta) = [P_{ij}(\theta)]$
  - $P_{ij}(\theta) = \mathbb{E}_{y \sim \nu(i)} \mathbb{E}_{u \sim \mu(\theta, y)} P_{ij}(u)$

- Assumption
  - $P(\theta)$ has a unique stationary distribution, $\pi(\theta)$, such that $\pi(\theta)^T P(\theta) = \pi(\theta)^T$
  - $|r(i)| < R < \infty$ for all i
- RL goal: Find $\theta \in \mathbb{R}^k$ which maximizes

$$\eta(\theta) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_\theta \left(\sum_{t=1}^{T} r(i_t) \right),$$

  where $\eta(\theta)$ is a long-term average reward and independent of starting state
- At stationarity,

$$\eta(\theta) = \sum_{i=1}^{n} \pi(\theta, i) r(i) = \pi(\theta)^T r$$

- Discounted version $(\eta_\beta(\theta) = \eta(\theta)/(1-\beta))$

$$\eta_\beta(\theta) = \sum_{i=1}^{n} \pi(\theta, i) J_\beta(\theta, i)$$
$$= \sum_{i=1}^{n} \pi(\theta, i) \mathbb{E}_\theta \left(\sum_{t=0}^{\infty} \beta^t r(i_t) \middle| i_0 = i \right)$$

Policy Gradient

- Gradient ascent rule: $\theta \leftarrow \theta + \gamma \nabla \eta(\theta)$
- Stationary distribution assumption: $\pi^T P = \pi^T$

$$
\begin{aligned}
\nabla \pi^T P + \pi^T \nabla P &= \nabla \pi^T \\
\nabla \pi^T (I - P) &= \pi^T \nabla P \\
\nabla \pi^T e = \nabla (\pi^T e) = \nabla (1) &= 0 \qquad (e = \text{vector of 1's}) \\
\nabla \pi^T (I - P) + \nabla \pi^T e \pi^T &= \pi^T \nabla P \\
\nabla \pi^T \left(I - P + e \pi^T\right) &= \pi^T \nabla P \\
\nabla \pi^T &= \pi^T \nabla P \left(I - P + e \pi^T\right)^{-1}
\end{aligned}
$$

- From $\eta = \pi^T r$, the policy gradient becomes

$$\nabla \eta = \nabla \pi^T r = \pi^T \nabla P \left(I - P + e \pi^T\right)^{-1} r$$

- It is difficult to compute $\left(I - P + e \pi^T\right)^{-1}$.
- The paper shows that $\nabla_\beta \eta = \pi^T \nabla P J_\beta$ is a good approximation for $\nabla \eta$ and a method for estimating $\nabla_\beta \eta$