Lecture 28 — December 2

Lecturer: Benjamin Recht *Scribe: Kevin Chang*

28.1 Reinforcement Learning Overview

A generic reinforcement learning procedure:

1. Generate candidate solutions (policies or actions).
2. Receive a reward or score for each candidate.
3. Update the parameters of the policy based on the received scores.
4. Repeat from Step 1.

28.2 Policy Gradient

We aim to solve:

$$\max_{p \in \mathcal{Q}} \mathbb{E}_{x \sim p}[r(x)].$$

A general stochastic gradient procedure is:

1. Sample $x_1, \ldots, x_n \sim p_t$.
2. Evaluate rewards $r(x_1), \ldots, r(x_n)$.
3. Update the distribution parameters to obtain p_{t+1}.

Let the policy be parameterized by w, and define:

$$\Phi(w) = \mathbb{E}_{x \sim p(x|w)}[r(x)].$$

28.2.1 Policy Gradient Derivation

$$\nabla_w \Phi(w) = \nabla_w \int r(x)\, p(x \mid w)\, dx = \int r(x)\, \nabla_w p(x \mid w)\, dx$$

$$= \int r(x)\, p(x \mid w)\, \nabla_w \log p(x \mid w)\, dx = \mathbb{E}_{x \sim p(x|w)}[\, r(x) \nabla_w \log p(x \mid w)\,].$$

Thus,

$$g(x) = r(x) \nabla_w \log p(x \mid w)$$

is an unbiased estimator of the true gradient.

Policy gradient = SGA using this estimator.

28.2.2 Stochastic Update

Given samples $x_1, \ldots, x_B$,

$$w_{t+1} = w_t + \frac{1}{B} \sum_{i=1}^{B} r(x_i) \nabla_w \log p(x_i \mid w_t).$$

For a discrete policy with probabilities π_i:

$$\pi_i^{t+1} = \pi_i^t + \eta \frac{r_i}{\pi_i^t}.$$

Two update interpretations:

- **Option 1:** Euclidean projection.
- **Option 2:** Mirror descent / exponentiated gradient:

$$\pi_i^{t+1} = \pi_i^t \exp\left(\eta \, \frac{r_i}{\pi_i^t}\right), \quad \text{then renormalize.}$$

28.3 Example: Gaussian Policy

Suppose $x \sim \mathcal{N}(z, \sigma^2 I)$, i.e.,

$$x = z + \sigma v, \qquad v \sim \mathcal{N}(0, I).$$

Compute:

$$\nabla_z \log p(x) = \frac{x - z}{\sigma^2} = \frac{v}{\sigma}.$$

Two useful gradient estimators:

$$G_1 = r(z + \sigma v) \, \frac{v}{\sigma},$$

$$G_2 = (r(z + \sigma v) - r(z)) \, \frac{v}{\sigma},$$

where G_2 behaves similarly to a directional derivative and reduces variance.