Lecture 22: Reinforcement Learning I

Hallmark of RL is a collection of algorithms for "black-box optimization" of stochastic optimal control problems

$$\sum_1^\infty g(x,u) \to \sum_{n=0}^\infty E[g(x,u)]$$

$$\underbrace{E\left[\sum_{n=0}^\infty g(x,u)\right]}_{\text{What we optimize.}}$$

$\longrightarrow$ $\Sigma\, g(x,u)$ ...

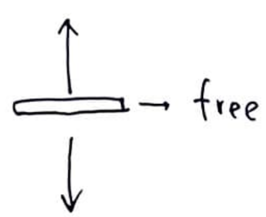
only have access to the costs

$\underbrace{\text{don't have } f(x,u) \ldots}$

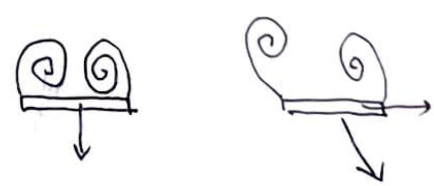
How do you optimize?

Example. Fluid dynamics of flapping flight

The heaving foil (Jun Zhang, NYU)



symmetric flat plate



unstable flow

dynamics governered by P.D.E.

difficult to compute dynamics

The Policy Gradient "trick" (REINFORCE)

$$\min_\alpha E[g(x)] \quad \text{with} \quad x \sim P_\alpha(x)$$

take gradient

$$\frac{d}{d\alpha} E[g(x)] = \frac{d}{dX} \int dx \; g(x) P_\alpha(x)$$

$$= \int dx \; g(x) \frac{d}{d\alpha} P_\alpha(x)$$

$$\cancel{\frac{d}{dx}} \frac{d}{d\alpha} \log P_\alpha(x) = \frac{1}{P_\alpha(x)} \frac{d}{d\alpha} P_\alpha(x)$$

$$= \int dx \; g(x) \frac{d}{d\alpha} \log P_\alpha(x) \cdot P_\alpha(x)$$

$$= E\left[g(x) \frac{d}{d\alpha} \log P_\alpha(x) \right]$$

approx with Monte-carlo

$$\approx \frac{1}{N} \sum_{n=1}^{N} g(x_n) \frac{d}{d\alpha} \log P_\alpha(x_n)$$

in optimal control case,

$$\frac{d}{d\alpha} E\left[\sum_{n=0}^{N} g(x[n], u[n]) \right] = \sum_{n=0}^{N} \int dx[n] \, du[n] \; g(x[n], u[n]) \frac{d}{d\alpha} P_\alpha(x[n], u[n])$$

$$= E\left[\sum_{n=0}^{N} g(x[n], u[n]) \frac{d}{d\alpha} \log P_\alpha(x[n], u[n]) \right]$$

$$P_\alpha(x[n], u[n]) = P_0(x[n]) \prod_{k=1}^{n} P(x[k] \mid x[k-1], u[k-1]) \prod_{k=0}^{n} P_\alpha(u[k] \mid x[k])$$

Taking the log we have

$$\log P_\alpha(x[n], u[n]) = \log P_0(x[n]) + \sum_{k=1}^{n} \log P(x[k] \mid x[k-1], u[k-1]) + \sum_{k=0}^{n} \log P_\alpha(u[k] \mid x[k])$$

Only the last term depends on α, which yields

$$\frac{d}{d\alpha} E\left[\sum_{n=0}^{N} g(x[n], u[n]) \right] = E\left[\sum_{n=0}^{N} g(x[n], u[n]) \sum_{k=0}^{n} \frac{d}{d\alpha} \log P_\alpha(u[k] \mid x[k]) \right]$$

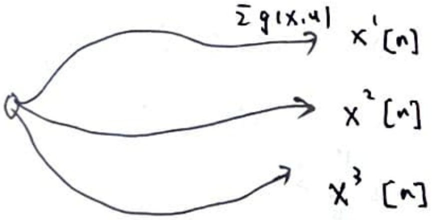
gradient on controller, not dependent on $P(x[k] \mid u[k-1], x[k-1])$

surprise ?

but a weak statement, because optimize expect?

Intuition

N random trajectories



$u = -Kx + w$ ← random number

make trajectory with better (lower) cost higher probability

~~Black-b~~ Black-box optimization

$$\min_\alpha g(\alpha)$$

How do you do gradient-free optimization?

Idea: finite differences

$$\underbrace{\left.\frac{\partial g}{\partial \alpha}\right|_\alpha}_{n \text{ parameters}} \approx \underbrace{\begin{bmatrix} \vdots \\ \dfrac{g(\alpha + \varepsilon_i) - g(\alpha)}{\varepsilon} \\ \vdots \end{bmatrix}}_{n+1 \text{ evaluations}} \qquad \varepsilon_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \varepsilon \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{index } i$$

expensive evaluation

Then gradient descent: $\alpha[k+1] = \alpha[k] - \eta \left.\frac{\partial g}{\partial \alpha}^T\right|_{\alpha[k]}$

Stochastic gradient descent

key idea: as long as I'm going "downhill", can get away w/ less evaluations of $g(\alpha)$

Simpler idea: "Weight perturbation"

$$\Delta \alpha = -\eta \left[g(\alpha+\beta) - g(\alpha) \right] \beta$$

small random vector

$\beta \sim N(0, \sigma)$

2 evaluations

$g(\alpha+\beta) > g(\alpha)$, in decrease direction

$g(\alpha+\beta) < g(\alpha)$, the same decrease direction

$$E[\Delta \alpha] = -\eta \sigma^2 \frac{\partial g}{\partial x}^T$$

$$\Delta \alpha = -\frac{\eta}{\cancel{\alpha^2}\,\sigma^2} \left[g(\alpha+\beta) - b \right] \beta$$

baseline

"expected performance"

still $E[\Delta\alpha] \propto -\frac{\partial g}{\partial \alpha}^T$

proportional

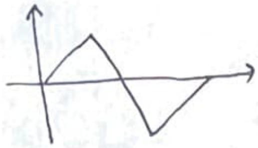
Key limitation of RL today "Sample Complexity"

Policy parameters for the heaving foil

Fourier base, not a good idea

different changes to the cost.

Cost function: max inverse cost of transport



converged reliably to a triangle wave

very repeatable experiments

"extrememum - seeking control"

" iterative learning control