IV.E Actor-Critic Algo

Last time: Applied func approx to value fcn

Q: Can we apply fcn approx to the control policy: $u_k = \pi(x_k)$

← regression

Let's parameterize the control policy:

$$\underset{\text{scalar } \in \mathbb{R}}{u_k} = \pi(x_k) = \underbrace{\overset{\text{learn online}}{U}^T}_{1\times M} \cdot \underbrace{\sigma(x_k)}_{M\times 1}$$

truncated (dense) basis set.

where $\sigma(x) = [\sigma_1(x), \sigma_2(x), \ldots, \sigma_M(x)]^T$

Return to Step 2: Policy Improvement.

Execute following minimization, using data $\left(x_k, x_{k+1}, c(x_k, \pi(x_k))\right)$

$$\underset{U}{\text{minimize}} \quad \underbrace{c\left(x_k, U^T \sigma(x_k)\right) + \gamma \cdot W_m^T \phi(x_{k+1})}_{= T(U)} \leftarrow$$

where $x_{k+1} = f\left(x_k, U^T \sigma(x_k)\right)$

A classic approach to solve $\min T(U)$ is gradient descent...

$$U_{j+1} = U_j - \beta \cdot \frac{\partial T}{\partial U}(U_j) \quad \text{for } \beta > 0$$

It's instructive to derive gradient $\frac{\partial T}{\partial U}$ $\leftarrow^{m \times 1}$ $\in \mathbb{R}^{m \times 1}$

$$\frac{\partial T}{\partial U}(U_j) = \Big[\frac{\partial c}{\partial u}(x_k, U_j^T \sigma(x_k)) \cdot \underline{\sigma}(x_k)$$

$$+ \gamma \cdot W_m^T \nabla \phi(x_{k+1}) \cdot \frac{\partial f}{\partial u}(x_k, U_j^T \sigma(x_k)) \cdot \underline{\sigma}(x_k)$$

$$= \Big[\frac{\partial c}{\partial u}(x_k, U_j^T \sigma(x_k)) + \gamma \cdot W_m^T \cdot \nabla \phi(x_{k+1}) \cdot \frac{\partial f}{\partial u}(x_k, U_j^T \sigma(x_k))\Big] \cdot \sigma(x_k)$$

Consider LQR case:

$$\text{minimize } T(\bar{U}) = x_k^T Q x_k + R u_k^2 + W^T \phi(x_{k+1})$$

$$= x_k^T Q x + R(\bar{U}^T \sigma(x_k))^2$$

$$+ W^T \phi(A x_k + B U^T \sigma(x_k))$$

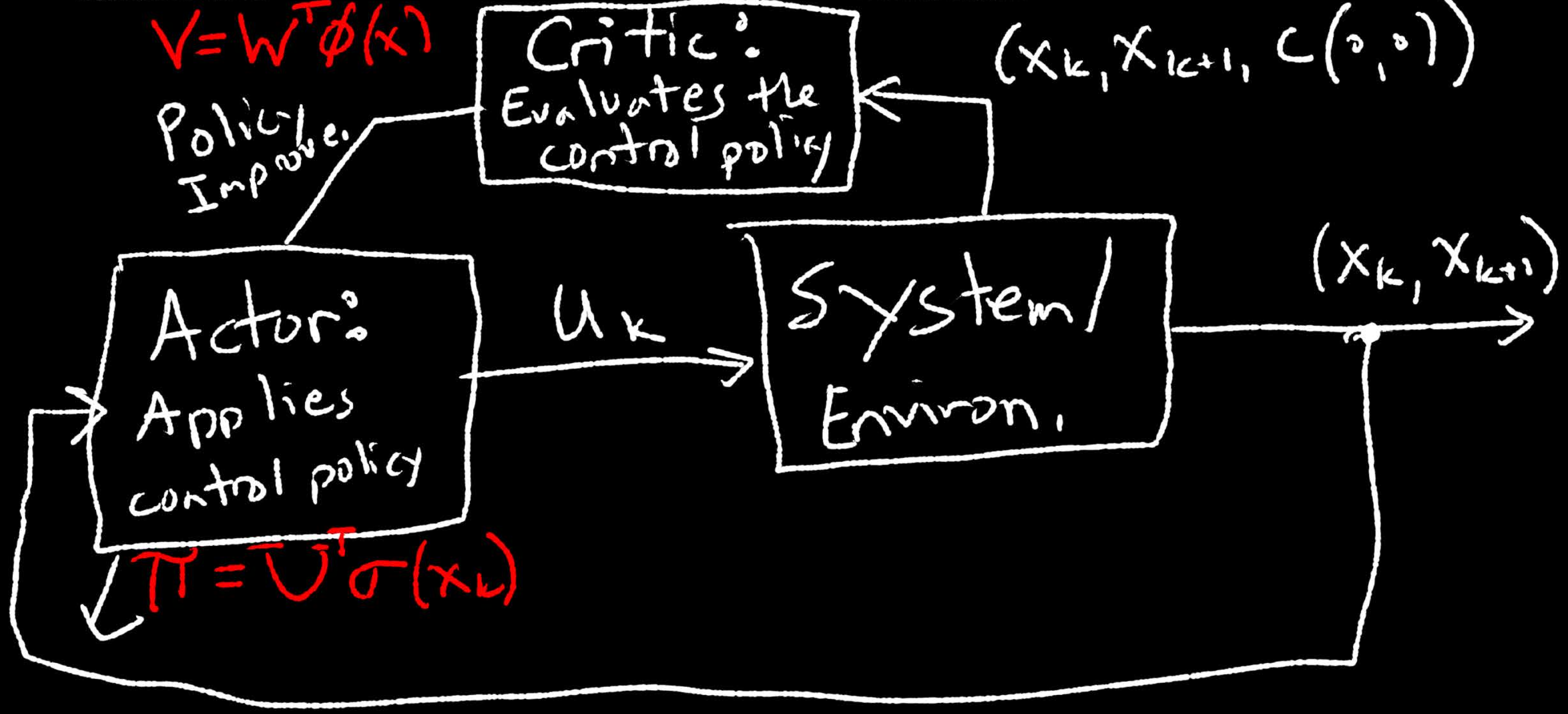
the gradient is

only info req'd from dyn.

$$\frac{\partial T}{\partial \bar{U}}(\bar{U}_j) = \left[2R(U_j^T \sigma(x_k)) + W_m^T \nabla \phi(x_{k+1}) B \right] \sigma(x_k)$$

Remark: Only info req'd from model dyn. is "B" in LQR, and $\frac{\partial f}{\partial u}(\cdot, u)$ in NL case.

Actor-Critic Algo



Summary of Actor-Critic Algo

1) Policy Eval

$$V = W^T \phi(x)$$

supervised learning

→ ←

2) Policy Improve,

$$\pi = U^T \sigma(x)$$

optimization