Policy Parameterizations

Objectives

- Understand Actor-Critic with Softmax policies
- Understand Gaussian policies for continuous actions

Actor-Critic with Softmax Policies

- Actor-Scritic algorithm
  - The critic uses semi-gradient TD

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S, \mathbf{w})$$

  - The actor uses the TDR from the critic to update the policy parameters

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \delta \nabla \ln \pi(A \mid S, \boldsymbol{\theta})$$

Actor-Critic with Softmax Policies

- Policy update with a softmax policy

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \delta \nabla \ln \pi(A \mid S, \boldsymbol{\theta})$$

  - We use a Softmax policy that exponentiates the preferences and divides by the sum

$$\pi(a \mid s, \boldsymbol{\theta}) \doteq \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_{b \in \mathcal{A}} e^{h(s,b,\boldsymbol{\theta})}}$$

FPT Education
FPT UNIVERSITY

Actor-Critic with Softmax Policies

- Features of the action preference function:

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^T \mathbf{x}(s)$$

- The actor's action preferences depend on the state and action

$$h(s, a, \boldsymbol{\theta}) \doteq \boldsymbol{\theta}^T \mathbf{x}_h(s, a)$$

Actor-Critic with Softmax Policies

- Using stacked state features

Features of the Action Preference Function

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^T \mathbf{x}(s)$$

$$h(s, a, \boldsymbol{\theta}) \doteq \boldsymbol{\theta}^T \mathbf{x}_h(s, a)$$

$$\mathbf{x}_h(s, a) = \begin{bmatrix} x_0(s) \\ x_1(s) \\ x_2(s) \\ x_3(s) \\ x_0(s) \\ x_1(s) \\ x_2(s) \\ x_3(s) \\ x_0(s) \\ x_1(s) \\ x_2(s) \\ x_3(s) \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} a_0 \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} a_1 \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} a_2 \end{matrix}$$

FPT Education
FPT UNIVERSITY

Actor-Critic with Softmax Policies

- Actor-Scritic algorithm
  - The critic

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \, \mathbf{x}(s)$$

  - The actor

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \delta \nabla \ln \pi(A \mid S, \boldsymbol{\theta})$$

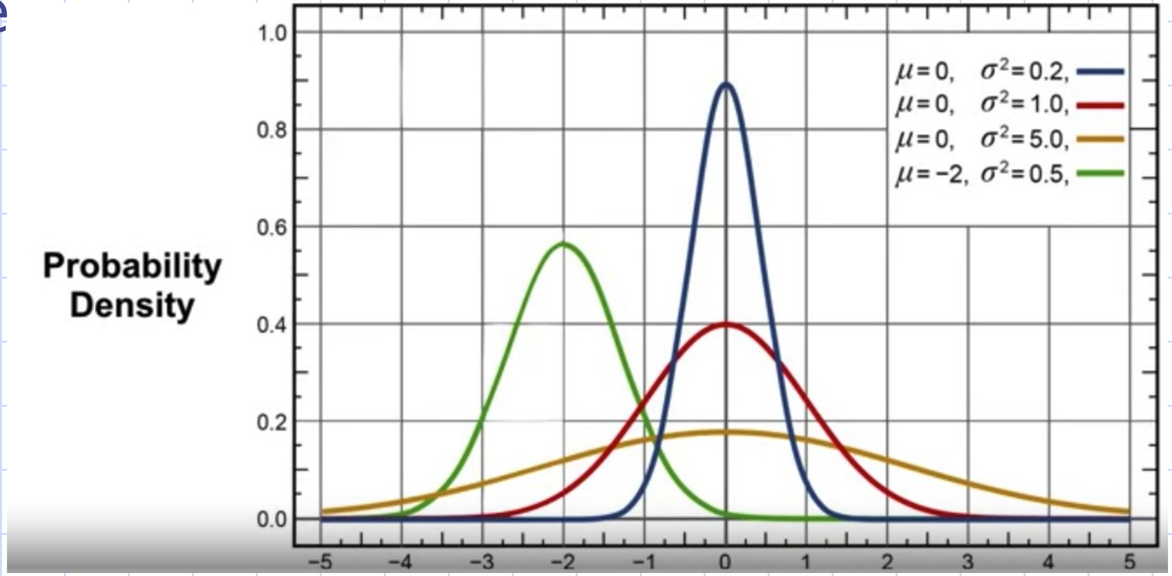
$$\nabla \ln \pi(a \mid s, \boldsymbol{\theta}) = \mathbf{x}_h(s, a) - \sum_b \pi(b \mid s, \boldsymbol{\theta}) \mathbf{x}_h(s, b)$$

Gaussian policies for continuous actions

- Gaussian Distribution

$$p(x) \doteq \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Example



Gaussian policies for continuous actions

- Define our policy using a Gaussian over actions.

$$\pi(a \mid s, \boldsymbol{\theta}) \doteq \frac{1}{\sigma(s, \boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{(a - \mu(s, \boldsymbol{\theta}))^2}{2\sigma(s, \boldsymbol{\theta})^2}\right)$$

  - Mu can be any parameterized function

$$\mu(s, \boldsymbol{\theta}) \doteq \boldsymbol{\theta}_\mu^T \mathbf{x}(s)$$

  - The parameter's function Sigma must be positive.

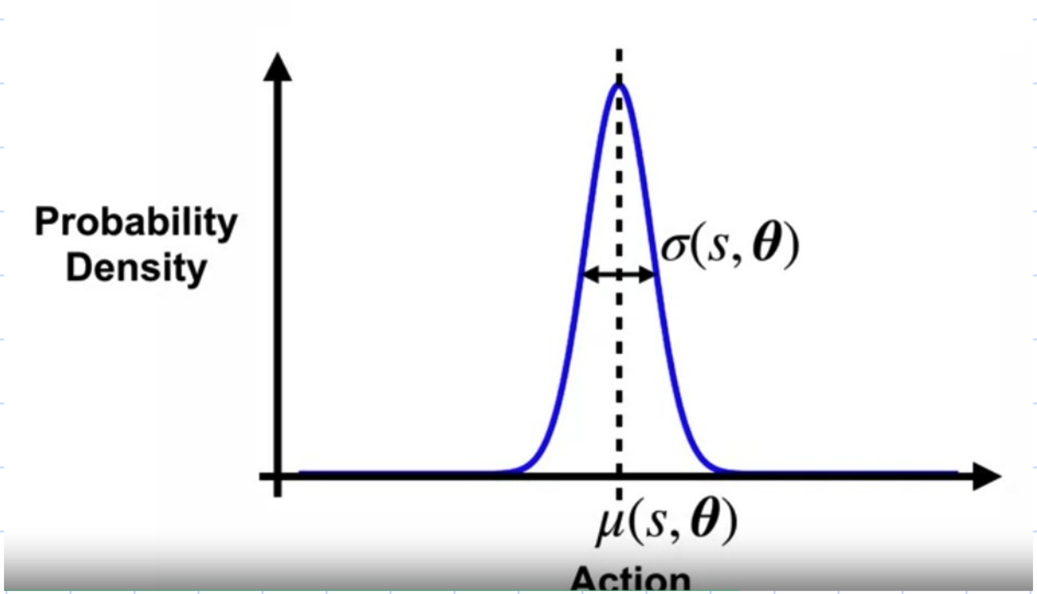
$$\sigma(s, \boldsymbol{\theta}) \doteq \exp\left(\boldsymbol{\theta}_\sigma^T \mathbf{x}(s)\right)$$

  - The policy parameters

$$\boldsymbol{\theta} \doteq \begin{bmatrix} \boldsymbol{\theta}_\mu \\ \boldsymbol{\theta}_\sigma \end{bmatrix}$$

Gaussian policies for continuous actions

- Gaussian Policies in Action
  - Sigma essentially controls the degree of exploration.
  - We typically initialize the variance to be large so that a wide range of actions are tried.



Gaussian policies for continuous actions

- Gradient of the Log of the Gaussian policy

$$\nabla \ln \pi(a \mid s, \boldsymbol{\theta}_\mu) = \frac{1}{\sigma(s, \boldsymbol{\theta})^2}(a - \mu(s, \boldsymbol{\theta}))\mathbf{x}(s)$$

$$\nabla \ln \pi(a \mid s, \boldsymbol{\theta}_\sigma) = \left(\frac{(a - \mu(s, \boldsymbol{\theta}))^2}{\sigma(s, \boldsymbol{\theta})^2} - 1 \right)\mathbf{x}(s)$$

Gaussian policies for continuous actions

- Advantages of Continuous actions
  - It might not be straightforward to choose a proper discrete set of action
  - Continuous actions allow us to generalize over actions
  - Expressiveness: allow for a finer-grained control over actions, enabling agents to perform a wide range of subtle and precise movements.
  - Smoothness: often lead to smoother and more natural policies, as agents can smoothly transition between different action values.

Gaussian policies for continuous actions

- Advantages of Continuous actions
  - Efficiency: can lead to more efficient exploration and learning
  - Generalization: facilitate better generalization across similar actions
  - Optimization: are amenable to optimization techniques that rely on gradient-based methods, such as policy gradients or actor-critic algorithms

Summary

- Understand Actor-Critic with Softmax policies
- Understand Gaussian policies for continuous actions

FPT Education
FPT UNIVERSITY

Q & A