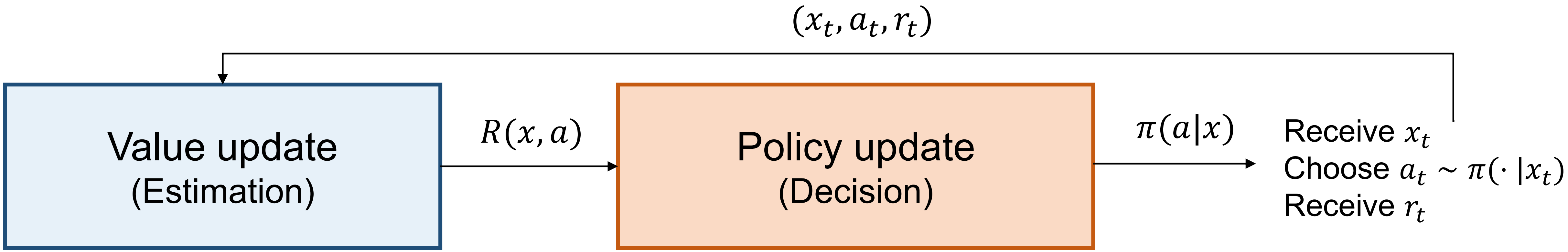
# Dealing with Continuous Action Set

# OpenAI SpinningUp

- https://spinningup.openai.com/en/latest/

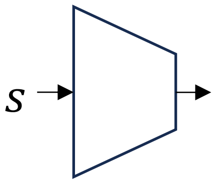
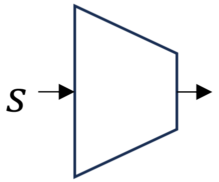
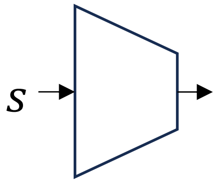
# A Unified View / Summary for Algorithms Discussed So Far

# A Unified View




| Contextual Bandit | Value<br>$R(x, a)$ | Policy<br>$\pi(a\|x)$ |
|---|---|---|
| Value-based approach (HW1) | $x \to$ [network] $\to R_\phi(x,\cdot) \in \mathbb{R}^A$ | Induced by $R_\phi$ |
| Policy-based approach (HW2) | $\hat{r}(x, a)$ constructed from real samples (importance weighting needed) | $x \to$ [network] $\to \pi_\theta(\cdot\,\|x) \in \Delta_A$ |

# A Unified View

| **MDP** | Value<br>$Q(s,a)$ | Policy<br>$\pi(a\|s)$ |
|---|---|---|
| Value-based approach<br>(HW3: DQN) |  $Q_\phi(s,\cdot) \in \mathbb{R}^A$ | Induced by $Q_\phi$ |
| Policy-based approach<br>(HW4: PPO) | $\hat{Q}(s,a)$ constructed from<br>1. Pure real samples (MC estimator), or<br>2. Real samples + $V_\phi(s)$ (TD estimator)<br> $V_\phi(s) \in \mathbb{R}$<br>(importance weighting needed) |  $\pi_\theta(\cdot\|s) \in \Delta_A$ |

# Contextual Bandits
# with Continuous Actions

# Contextual Bandits with Continuous Actions

**Given:** Action set $\Omega \subseteq \mathbb{R}^d$

For time $t = 1, 2, \dots, T$:

Environment reveals a context $x_t$

Learner chooses an action $a_t \in \Omega$

Environment reveals a reward value $r_t(x_t, a_t) = R(x_t, a_t) + \text{noise}$

# Value-Based Approach

- Use supervised learning to learn a reward function $R_\phi(x, a)$

- How to perform the exploration strategies (like $\epsilon$-Greedy)?
  - How to find $\text{argmax}_a R_\phi(x, a)$?
  - Usually, there needs to be another **policy learning procedure** that helps to find $\text{argmax}_a R_\phi(x, a)$
  - Then we can explore as $a_t = \text{argmax}_a R_\phi(x, a) + \mathcal{N}(0, \sigma^2 I)$

# Value-Based Approach

But more precisely, this is a combination of value and policy approaches

For $t = 1,2,\ldots,T$:

Receive context $x_t$

Take action $a_t = \mathcal{P}_\Omega\big(\mu_\theta(x_t) + \mathcal{N}(0, \sigma^2 I)\big) \in \mathbb{R}^d$

Receive $r_t(x_t, a_t)$

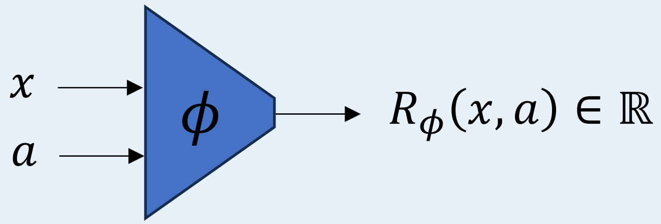
Update the reward model:

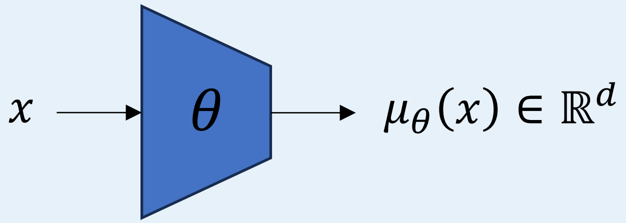
$$\phi \leftarrow \phi - \lambda \nabla_\phi \left[\left(R_\phi(x_t, a_t) - r_t(x_t, a_t)\right)^2\right]$$

Update policy:

$$\theta \leftarrow \theta + \eta \nabla_\theta R_\phi(x_t, \mu_\theta(x_t))$$

Equivalent policy parametrization
$\pi_\theta(\cdot\,|x) = \mathcal{N}(\mu_\theta(x), \sigma^2 I)$




Think of this as a continuous-action counterpart of $\epsilon$-Greedy

# Gradient Ascent with Gradient Estimator

Arbitrarily initialize $\mu_1 \in \Omega$

For $t = 1,2, \dots, T$:

Let $a_t = \Pi_\Omega(\mu_t + z_t)$ where $z_t \sim \mathcal{D}$ (assume that $\|z_t\| \leq \delta$ always holds)

Receive $r_t(a_t)$

Define

$$g_t = (r_t(a_t) - b_t)H_t^{-1} z_t \qquad \text{where } H_t \coloneqq \mathbb{E}_{z\sim\mathcal{D}}[zz^\top]$$

Update policy:

$$\mu_{t+1} = \Pi_\Omega\left(\mu_t + \eta g_t\right)$$

# Continuous Contextual Bandits

Pure policy-based algorithms

# Gradient Ascent / PPO

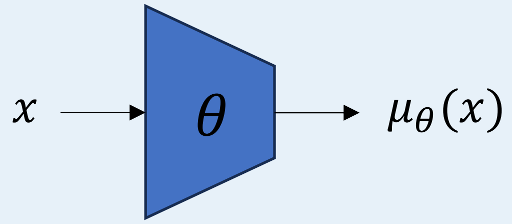
For $t = 1,2,\dots,T$:

Receive context $x_t$

Let $a_t = \mu_{\theta_t}(x_t) + \mathcal{N}(0, \sigma^2 I)$

Receive $r_t(x_t, a_t)$

Equivalent policy parametrization
$\pi_\theta(\cdot\,|x) = \mathcal{N}(\mu_\theta(x), \sigma^2 I)$




Update policy: (Ideally)

$$\theta_{t+1} \leftarrow \theta_t + \alpha \nabla_\theta \int_a \pi_\theta(a|x_t)\, R(x_t, a)\, \mathrm{d}a \Big|_{\theta=\theta_t}$$

or $$\theta_{t+1} \leftarrow \operatorname*{argmax}_\theta \int_a \pi_\theta(a|x_t)\, R(x_t, a)\, \mathrm{d}a - \frac{1}{\eta} \mathrm{KL}(\pi_\theta(\cdot\,|x_t), \pi_{\theta_t}(\cdot\,|x_t))$$

# Gradient Ascent / PPO

Review: how did we do this in finite-action case?

# Gradient Ascent / PPO

$$\pi_\theta(a|x) = \frac{1}{(2\pi\sigma^2)^{d/2}}\left(-\frac{1}{2\sigma^2}\|a-\mu_\theta(x)\|^2\right)$$

For $t = 1,2,\dots,T$:

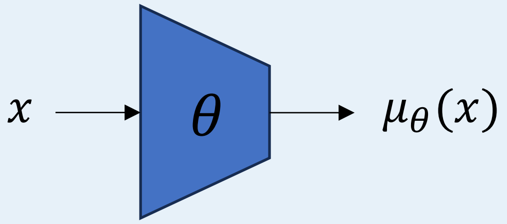
Receive context $x_t$

Let $a_t = \mu_{\theta_t}(x_t) + \mathcal{N}(0,\sigma^2 I)$

Receive $r_t(x_t, a_t)$

Equivalent policy parametrization
$\pi_\theta(\cdot\,|x) = \mathcal{N}(\mu_\theta(x), \sigma^2 I)$

Plug this in




Update policy:

$$\theta_{t+1} \leftarrow \theta_t + \alpha \frac{\nabla_\theta \pi_\theta(a_t|x_t)}{\pi_{\theta_t}(a_t|x_t)}\left(r_t(x_t,a_t) - b_t(x_t)\right) = \theta_t + \alpha \nabla_\theta \log \pi_\theta(a_t|x_t)\left(r_t(x_t,a_t) - b_t(x_t)\right)$$

or

$$\theta_{t+1} \leftarrow \underset{\theta}{\operatorname{argmax}} \frac{\pi_\theta(a_t|x_t)}{\pi_{\theta_t}(a_t|x_t)}\left(r_t(x_t,a_t) - b_t(x_t)\right) - \frac{1}{\eta}\underbrace{\mathrm{KL}(\pi_\theta(\cdot\,|x_t), \pi_{\theta_t}(\cdot\,|x_t))}_{\frac{1}{2\sigma^2}\left\|\mu_\theta(x_t) - \mu_{\theta_t}(x_t)\right\|^2}$$