# Dealing with Continuous Action Set

# Continuous Action Set

Full-information feedback

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

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

Learner chooses a point $a_t \in \Omega$

Environment reveals a reward function $r_t\colon\ \Omega \to \mathbb{R}$

Bandit feedback

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

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

Learner chooses a point $a_t \in \Omega$

Environment reveals a reward value $r_t(a_t)$

# Continuous Multi-Armed Bandits

With a mean estimator

| | MAB | CB |
|---|---|---|
| VB | ● | |
| PB | | |

# Value-Based Approach (mean estimation)

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

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

# Value-Based Approach (mean estimation)

The mean estimator $R_\phi$ essentially gives us a full-information reward function

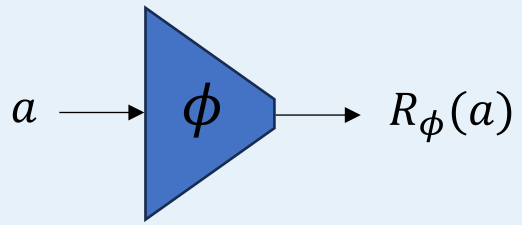
For $t = 1,2,\ldots,T$:

Take action $a_t = \mathcal{P}_\Omega\big(\mu_t + \mathcal{N}(0,\sigma^2 I)\big)$

Receive $r_t(a_t)$

Update the mean estimator:

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




Update policy:

$$\mu_{t+1} = \mathcal{P}_\Omega\big(\mu_t + \eta \nabla_\mu R_\phi(\mu_t)\big)$$

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

# Continuous Contextual Bandits

With a regression oracle

| | MAB | CB |
|---|---|---|
| VB | | ● |
| PB | | |

# Combining with Regression Oracle (a bandit version of DDPG)

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

Receive context $x_t$

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

Receive $r_t(x_t, a_t)$

Update the regression oracle:

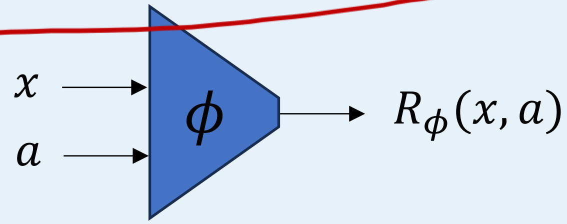
$$\phi \leftarrow \phi - \lambda \nabla_\phi \left[ \big( R_\phi(x_t, a_t) - r_t(x_t, a_t) \big)^2 \right]$$

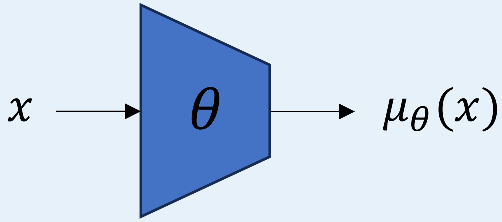
Update policy:

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

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

$\mu_\theta(x) \approx \arg\max_a R_\phi(x,a)$



# Continuous Multi-Armed Bandits

Pure policy-based algorithms

| | MAB | CB |
|---|---|---|
| VB | | |
| PB | • | |

# Pure Policy-Based Approach

**Gradient Ascent (full-information)**

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

Choose action $\mu_t$

Receive reward function $r_t\colon\ \Omega \to \mathbb{R}$

Update action $\mu_{t+1} \leftarrow \mathcal{P}_\Omega(\mu_t + \eta \boxed{\nabla r_t(\mu_t)})$

We face a similar problem as in EXP3: if we only observe $r_t(a_t)$, how can we estimate the **gradient**?

# (Nearly) Unbiased Gradient Estimator

**Goal:** construct $g_t \in \mathbb{R}^d$ such that $\mathbb{E}[g_t] \approx \nabla r_t(\mu_t)$ with only $r_t(a_t)$ feedback

# (Nearly) Unbiased Gradient Estimator (1/3)

Consider $d = 1$

$$\nabla r(\mu) \approx \frac{r(\mu+\delta) - r(\mu-\delta)}{2\delta}$$

$$= \frac{1}{2}\left(\frac{r(\mu+\delta)}{\delta}\right) + \frac{1}{2}\left(-\frac{r(\mu-\delta)}{\delta}\right)$$

$$= \mathbb{E}_{\beta \sim \mathrm{unif}\{-1,1\}}\left[\frac{\beta \cdot r(\mu+\beta\delta)}{\delta}\right]$$

$$= \mathbb{E}_{a \sim \mathrm{unif}\{\mu-\delta, \mu+\delta\}}\left[\frac{(a-\mu)r(a)}{\delta^2}\right]$$

# (Nearly) Unbiased Gradient Estimator (2/3)

Uniformly randomly choose a direction $i_t \in \{1, 2, \dots, d\}$

Uniformly randomly choose $\beta_t \in \{1, -1\}$

Sample $a_t = \mu_t + \delta \beta_t \mathrm{e}_{i_t}$

Observe $r_t(a_t)$

Define $g_t = \frac{d r_t(a_t)}{\delta} \beta_t \mathrm{e}_{i_t}$

# (Nearly) Unbiased Gradient Estimator (3/3)

Choose $z_t \sim \mathcal{D}$ with $\mathbb{E}_{z\sim\mathcal{D}}[z] = 0$

Sample $a_t = \mu_t + z_t$

Observe $r_t(a_t)$

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

# Baseline

$$g_t = (r_t(a_t) - b_t)H_t^{-1}z_t$$

Besides controlling the extent of exploration, it also affects the **variance** of the gradient

# 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 $\|z_t\| \le \delta$ 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

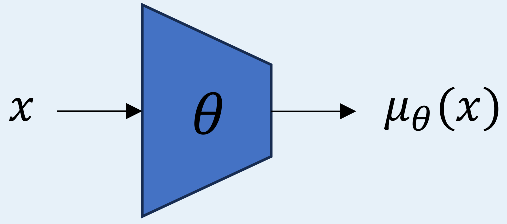
| | MAB | CB |
|---|---|---|
| VB | | |
| PB | | • |

# Gradient Ascent with Gradient Estimator (PG)

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

Receive context $x_t$

Let $a_t = \mu_{\theta_t}(x_t) + z_t$ where $z_t \sim \mathcal{D}$




Receive $r_t(x_t, a_t)$

Define

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

Recall: $g_t$ is an estimator for $\nabla_\mu r_t(x_t, \mu)\big|_{\mu=\mu_{\theta_t}(x_t)}$

Update policy:

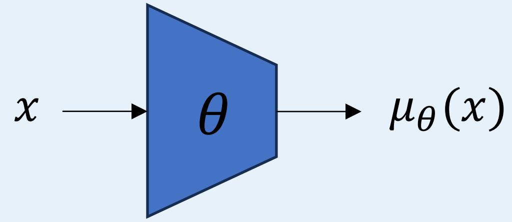
$$\theta_{t+1} \leftarrow \theta_t + \eta \left[\text{an estimator of } \nabla_\theta r_t\big(x_t, \mu_\theta(x_t)\big) \text{ at } \theta = \theta_t\right]$$

# Gradient Ascent with Gradient Estimator (PG)

# Gradient Ascent with Gradient Estimator (PG)

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

Receive context $x_t$




Let $a_t = \mu_{\theta_t}(x_t) + z_t$ where $z_t \sim \mathcal{D}$

Receive $r_t(x_t, a_t)$

Define

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

Recall: $g_t$ is an estimator for $\nabla_\mu r_t(x_t, \mu)\big|_{\mu=\mu_{\theta_t}(x_t)}$

Update policy:

$$\theta_{t+1} \leftarrow \theta_t + \eta\, \nabla_\theta \langle \mu_\theta(x_t), g_t \rangle|_{\theta=\theta_t}$$

*c.f.* finite action case

$$\nabla_\theta \langle \pi_\theta(\cdot\,|x_t), \hat{r}_t \rangle|_{\theta=\theta_t}$$

# Gradient Ascent with Gradient Estimator (PG)

**An alternative expression:**

When $\mathcal{D} = \mathcal{N}(0, H_t)$, we have

$$\nabla_\theta \langle \mu_\theta(x_t), g_t \rangle \qquad = \qquad \nabla_\theta \log \pi_\theta(a_t | x_t)(r_t(x_t, a_t) - b_t(x_t))$$

$g_t = (r_t(x_t, a_t) - b_t(x_t))H_t^{-1} z_t$

$H_t = \mathbb{E}_{z \sim \mathcal{D}}[zz^\top]$

$a_t = \mu_\theta(x_t) + z_t$

$\pi_\theta(\cdot | x_t) = \mathcal{N}(\mu_\theta(x_t), H_t)$

$$\pi_\theta(a | x_t) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(H_t)^{\frac{1}{2}}} e^{-\frac{1}{2}(a - \mu_\theta(x_t))^\top H_t^{-1}(a - \mu_\theta(x_t))}$$

# Gradient Ascent with Gradient Estimator (PG)

$\nabla_\theta \log \pi_\theta(a_t|x_t)(r_t(x_t,a_t) - b_t(x_t))$ is a general and direct way to construct gradient estimator in the parameter space:

$$V(\theta) = \int \pi_\theta(a|x_t)\, r_t(x_t,a)\, \mathrm{d}a$$

$$\nabla_\theta V(\theta) = \int \nabla_\theta \pi_\theta(a|x_t)\, r_t(x_t,a)\, \mathrm{d}a = \int \pi_\theta(a|x_t) \frac{\nabla_\theta \pi_\theta(a|x_t)}{\pi_\theta(a|x_t)} r_t(x_t,a)\, \mathrm{d}a$$

Unbiased estimator for $\nabla_\theta V(\theta)$:

Sample $a_t \sim \pi_\theta(\cdot\,|x_t)$ and define estimator $= \frac{\nabla_\theta \pi_\theta(a_t|x_t)}{\pi_\theta(a_t|x_t)} r_t(x_t,a_t) = \nabla_\theta \log \pi_\theta(a_t|x_t) r_t(x_t,a_t)$

# Gradient Ascent with Gradient Estimator (PG)

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

Receive context $x_t$

Let $a_t \sim \pi_{\theta_t}(\cdot \mid x_t)$

Receive $r_t(x_t, a_t)$

Update policy:

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

# PPO

PPO update

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

$$\approx \operatorname*{argmax}_{\theta} \left\{ \langle \mu_\theta(x_t), g_t \rangle - \frac{1}{2\eta\sigma^2} \left\| \mu_\theta(x_t) - \mu_{\theta_t}(x_t) \right\|^2 \right\}$$

*c.f.* PG update

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

$$\approx \operatorname*{argmax}_{\theta} \left\{ \langle \mu_\theta(x_t), g_t \rangle - \frac{1}{2\eta} \|\theta - \theta_t\|^2 \right\}$$

# Summary for Bandits

3 main challenges in online RL: Exploration, Generalization, (Temporal) Credit Assignment

+ Generalization over actions

Finite actions → Infinite actions

| | Finite actions: MAB | Finite actions: CB | Infinite actions: MAB | Infinite actions: CB |
|---|---|---|---|---|
| | Exploration | + Generalization over contexts | | |
| **VB** | $R(\cdot) \in \mathbb{R}^A$ | $x \rightarrow R_\phi(x,\cdot) \in \mathbb{R}^A$ | $a \rightarrow R_\phi(a) \in \mathbb{R}$<br>$\mu \in \mathbb{R}^d$ | $x, a \rightarrow R_\phi(x,a) \in \mathbb{R}$<br>$x \rightarrow \mu_\theta(x) \in \mathbb{R}^d$ |
| **PB** | $\pi(\cdot) \in \Delta_A$ | $x \rightarrow \pi_\theta(\cdot\,\vert x) \in \Delta_A$ | $\mu \in \mathbb{R}^d$ | $x \rightarrow \mu_\theta(x) \in \mathbb{R}^d$ |