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)

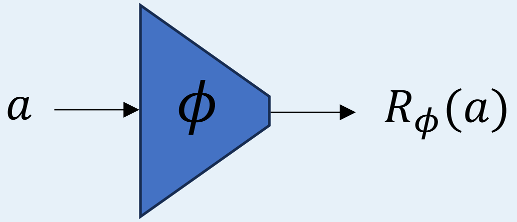
- Use supervised learning to learn a reward function $R_\phi(a)$

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

Value-Based Approach (mean estimation)

The mean estimator R_ϕ essentially gives us a full-information reward function

```
For $t = 1,2,\dots,T$:
    Take action $a_t = \mathcal{P}_\Omega\left(\mu_t + \mathcal{N}(0,\sigma^2 I)\right)$
    Receive $r_t(a_t)$
    Update the mean estimator:
        $\phi \leftarrow \phi - \lambda \nabla_\phi \left[\left(R_\phi(a_t) - r_t(a_t)\right)^2\right]$
    Update policy:
        $\underline{\mu_{t+1}} = \mathcal{P}_\Omega\left(\mu_t + \eta \nabla_\mu R_\phi(\mu_t)\right)$
```



mimic $\underset{\mu}{\text{argmax}}\ R_\phi(\mu)$

Think of this as a continuous-action counterpart of ϵ-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 $a_t = \mathcal{P}_\Omega\left(\mu_\theta(x_t) + \mathcal{N}(0,\sigma^2 I)\right)$

Receive $r_t(x_t, a_t)$

Update the regression oracle:

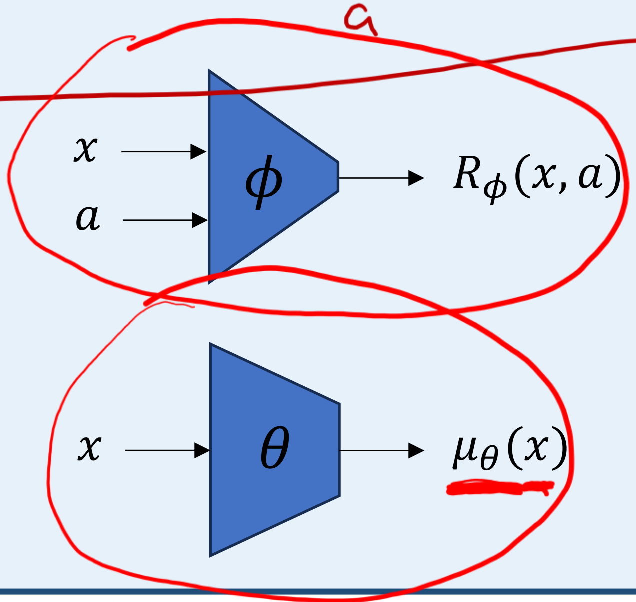
$$\phi \leftarrow \phi - \lambda \nabla_\phi \left[\left(R_\phi(x_t,a_t) - r_t(x_t,a_t)\right)^2\right]$$

Update policy:

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

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

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



mimic $\arg\max_\theta R_\phi(x_t, \mu_\theta(x_t))$

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

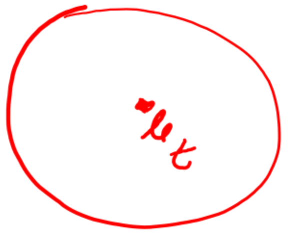
gradient

$$\left[\nabla_\mu r_t(\mu)\right]\Big|_{\mu=\mu_t}$$

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

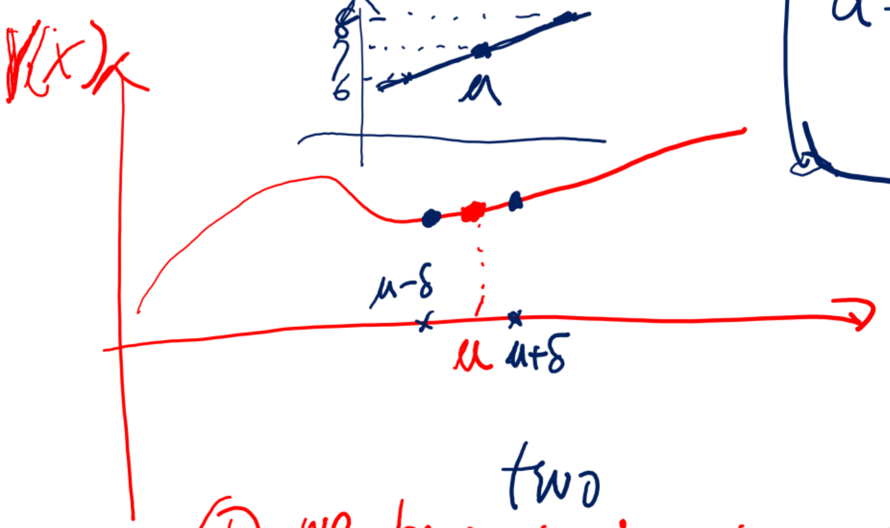
(Nearly) Unbiased Gradient Estimator

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



(Nearly) Unbiased Gradient Estimator (1/3)

Consider $d = 1$



$$\boxed{a = \begin{cases} \mu+\delta & w.p.\ \frac{1}{2} \\ \mu-\delta & w.p.\ \frac{1}{2} \end{cases}}$$

$$\boxed{g = \begin{cases} \frac{r(\mu+\delta)}{\delta} & \text{if } a=\mu+\delta \\ -\frac{r(\mu-\delta)}{\delta} & \text{if } a=\mu-\delta \end{cases}}$$

$$g = \begin{cases} \frac{\delta}{\delta} & w.p.\ \frac{1}{2} \\ \frac{-\delta}{\delta} & w.p.\ \frac{1}{2} \end{cases}$$

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

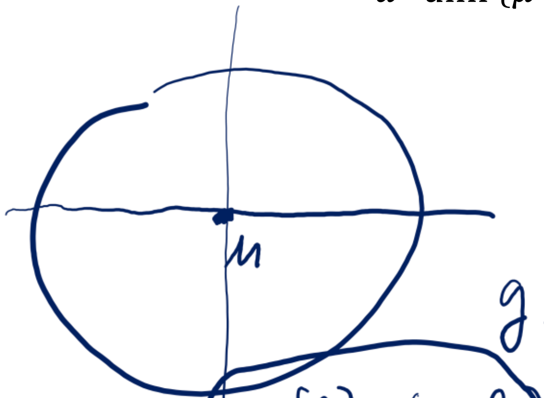
$$= \frac{1}{2}\left(\frac{r(\mu+\delta)}{\delta}\right) + \frac{1}{2}\left(-\frac{r(\mu-\delta)}{\delta}\right) = \mathbb{E}[g]$$

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

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

① we have ~~one~~ two chance to sample a point a and receive $r(a)$

② We want generate g such that $\mathbb{E}[g] = \frac{dr(x)}{dx}\Big|_{x=\mu}$



$$g = (g_1, g_2)$$

$$g = \begin{cases} (2g_1, 0), & w.p.\ \frac{1}{2} \\ (0, 2g_2) & w.p.\ \frac{1}{2} \end{cases}$$

$$\boxed{\mathbb{E}[g] = (g_1, g_2)}$$

(Nearly) Unbiased Gradient Estimator (2/3)

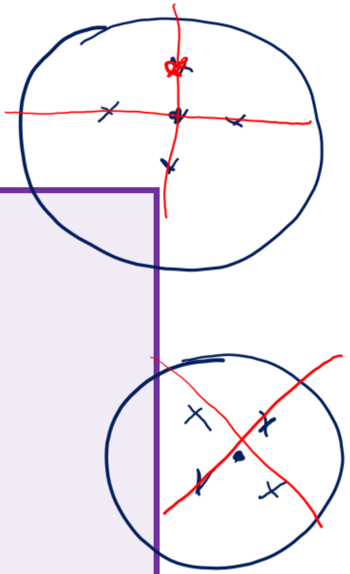
Uniformly randomly choose a direction $i_t \in \{1, 2, \ldots, d\}$

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



$$z_t = \begin{cases} \delta e_1 & \text{w.p. } \frac{1}{2d} \\ -\delta e_1 & \text{w.p. } \frac{1}{2d} \\ \delta e_2 \\ -\delta e_2 \\ \vdots \\ \delta e_d \\ -\delta e_d \end{cases}$$

$$H_t = \frac{\delta^2}{d} I$$

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

$$H_t^{-1} \underbrace{\mathbb{E}_{z_t}\left[z_t z_t^T\right]}_{H_t} v = v$$

Assume $r_t(a) = v^T a + b$ (v is the gradient)

$$\boxed{\text{Prove } \mathbb{E}_{z_t}\left[g_t\right] = v}$$

$$\mathbb{E}_{z_t}\left[r_t(a_t) H_t^{-1} z_t\right] = \mathbb{E}_{z_t}\left[(v^T a_t + b) H_t^{-1} z_t\right]$$

$$= \mathbb{E}_{z_t}\left[\left(\underbrace{v^T \mu_t}_{\text{const}} + v^T z_t \underbrace{+b}_{\text{const}}\right) H_t^{-1} z_t\right] = \mathbb{E}_{z_t}\left[(v^T z_t) H_t^{-1} z_t\right]$$

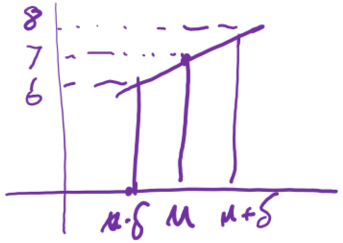
$$= \mathbb{E}_{z_t}\left[H_t^{-1} z_t z_t^T v\right]$$

Baseline

$$\frac{1}{2}\frac{8}{\delta} + \frac{1}{2}\frac{-6}{\delta} = \frac{1}{\delta}$$

$$g = \begin{cases} \frac{r(\mu+\delta)}{\delta} = \frac{8}{\delta} \text{ if choose } \mu+\delta \\ \frac{-r(\mu-\delta)}{\delta} = \frac{-6}{\delta} \text{ if choose } \mu-\delta \end{cases}$$

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



$$g = \begin{cases} \frac{r(\mu+\delta)-7}{\delta} = \frac{1}{\delta} \\ \frac{-(r(\mu-\delta)-7)}{\delta} = \frac{1}{\delta} \end{cases}$$

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 $t = 1,2,\ldots,T$:
    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
        $g_t = (r_t(a_t) - b_t)H_t^{-1} z_t$        where $H_t := \mathbb{E}_{z\sim\mathcal{D}}[zz^\top]$
    Update policy:
        $\mu_{t+1} = \Pi_\Omega(\mu_t + \eta g_t)$
```

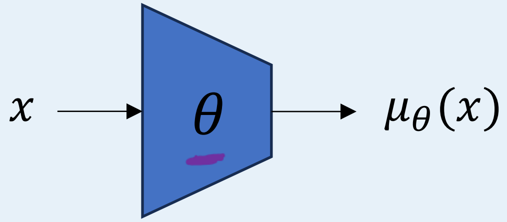
Gaussian $(0, \sigma^2 I)$

Continuous Contextual Bandits

Pure policy-based algorithms

	MAB	CB
VB		
PB		●

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
        $g_t = (r_t(x_t, a_t) - b_t(x_t))H_t^{-1}z_t$    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:
        $\theta_{t+1} \leftarrow \theta_t + \eta$ [an estimator of $\nabla_\theta r_t(x_t, \mu_\theta(x_t))$ at $\theta = \theta_t$]
```

$$= \boxed{\nabla_\mu r_t(x_t, \mu)\Big|_{\mu=\mu_\theta(x_t)}} \quad \boxed{\nabla_\theta \mu_\theta(x_t)}$$

$$g_t$$

Gradient Ascent with Gradient Estimator (PG)

Gradient Ascent with Gradient Estimator (PG)

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

Receive context x_t

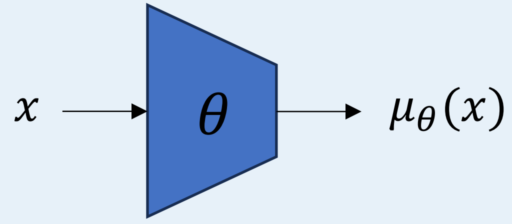
Let $a_t = \mu_{\theta_t}(x_t) + z_t$ where $z_t \sim \mathcal{D}$

Receive $r_t(x_t, a_t)$

Define

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

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

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

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

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

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

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

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

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 $= \boxed{\frac{\nabla_\theta \pi_\theta(a_t|x_t)}{\pi_\theta(a_t|x_t)} r_t(x_t,a_t)} = \nabla_\theta \log \pi_\theta(a_t|x_t) r_t(x_t,a_t)$

Gradient Ascent with Gradient Estimator (PG)

```
For $t = 1,2,\ldots,T$:
    Receive context $x_t$
    Let $a_t \sim \pi_{\theta_t}(\cdot\,|x_t)$
    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

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

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

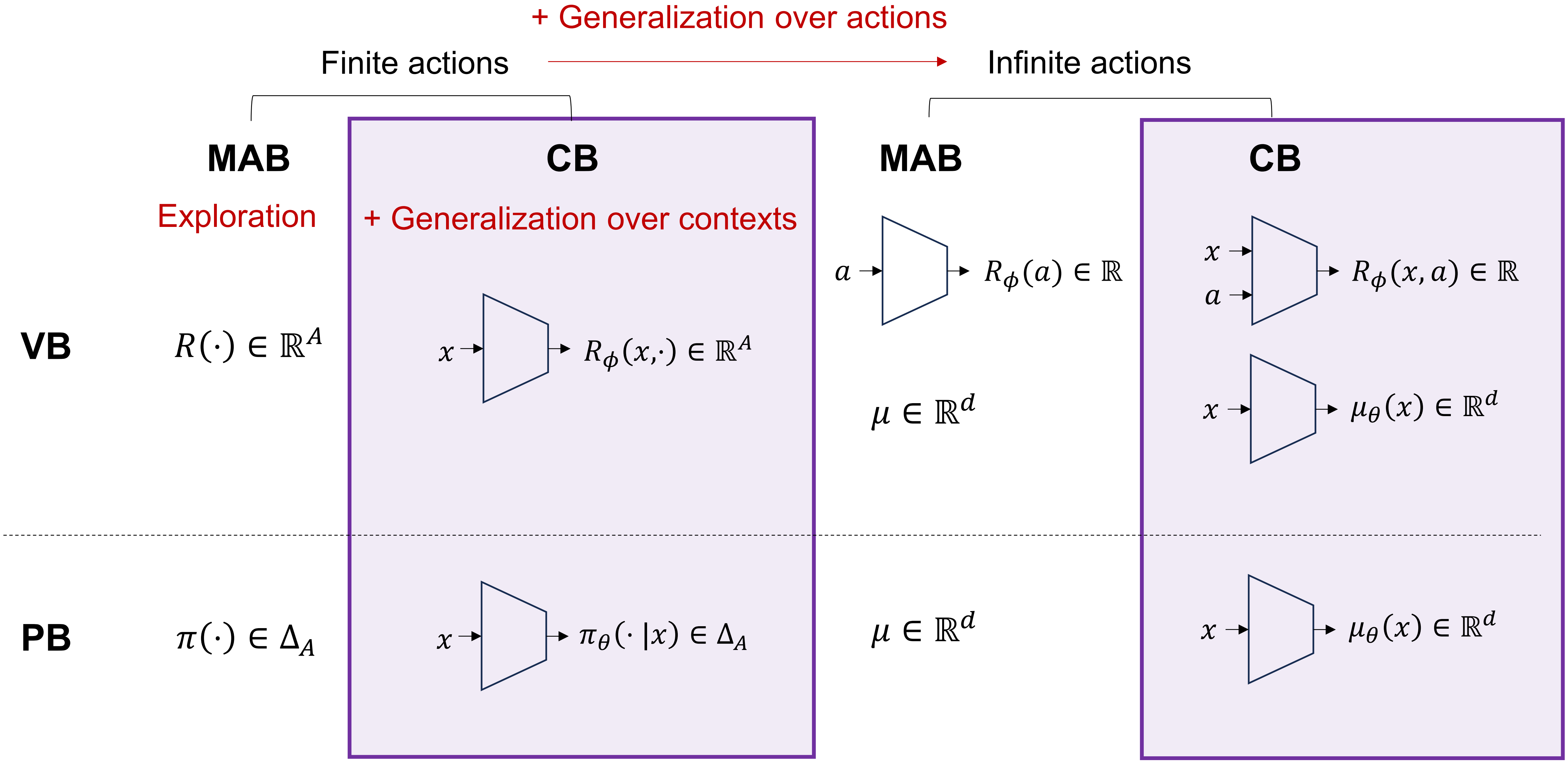
c.f. PG update

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

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

Summary for Bandits

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



+ Generalization over actions

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