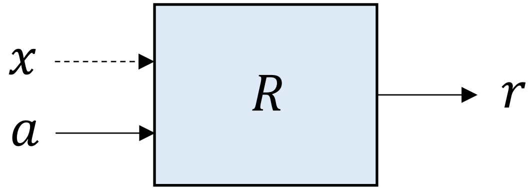
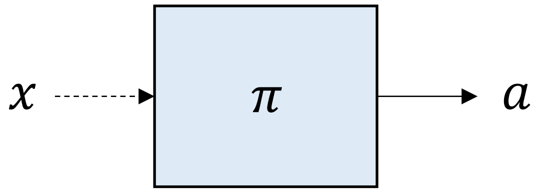
Review: Bandit Techniques

x: context, a: action, r: reward

		MAB	CB
Value-based	 (context, action) to reward	Mean estimation + EG, BE, IGW	Regression + EG, BE, IGW
Policy-based	 context to action distribution	KL-regularized update with reward estimators (EXP3) + baseline, bias, or uniform exploration	PPO/NPG PG + baseline, bias, uniform exploration, clipping

Are we done with bandits?

- Almost, but we have a last important topic: how to deal with continuous action sets? (#actions could be infinite)
- We will go over the 4 regimes once again to deal with continuous actions

	MAB	CB
VB		
PB		

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

Full-Information Policy learning Procedure

Gradient Ascent

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

Choose action μ_t

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

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

When $\pi_\theta = \mathcal{N}(\mu_\theta, \sigma^2 I)$, the KL-regularized policy update

$$\theta_{t+1} = \operatorname*{argmax}_\theta \left\{ \int \left(\pi_\theta(a) - \pi_{\theta_t}(a)\right) r_t(a)\, \mathrm{d}a - \frac{1}{\eta} \mathrm{KL}\left(\pi_\theta, \pi_{\theta_t}\right) \right\}$$

is close to $\mu_{\theta_{t+1}} \leftarrow \mu_{\theta_t} + \eta\sigma \nabla r_t(\mu_{\theta_t})$

Regret Bound of Gradient Ascent

Theorem. If Ω is convex and all reward functions r_t are concave, then Gradient Ascent ensures

$$\text{Regret} = \max_{\mu^\star \in \Omega} \sum_{t=1}^{T} r_t(\mu^\star) - r_t(\mu_t) \leq \frac{\max_{\mu \in \Omega} \|\mu\|_2^2}{\eta} + \eta \sum_{t=1}^{T} \|\nabla r_t\|_2^2$$

This can also be applied to the finite-action setting, but only ensures a $\sqrt{AT}$ regret bound (using exponential weights we get $\sqrt{(\log A)T}$)

Combining with Mean Estimator

$\pi_t(a) = N(\mu_t, \sigma^2 I)$

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

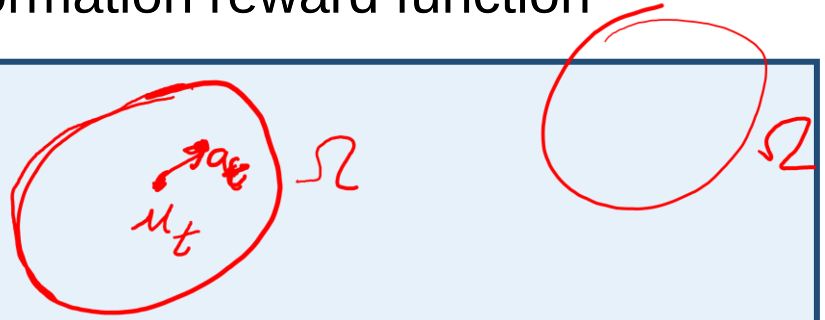
For $t = 1,2,\ldots,T$:

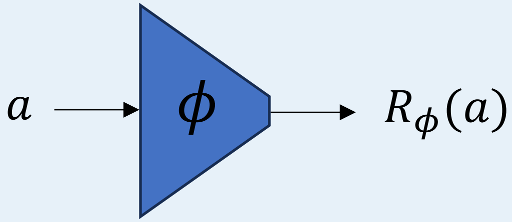
Take action $a_t = \mathcal{P}_\Omega\left(\mu_t + \mathcal{N}(0,\sigma^2 I)\right)$

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:

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

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

Receive $r_t(x_t, a_t)$

Update the regression oracle:

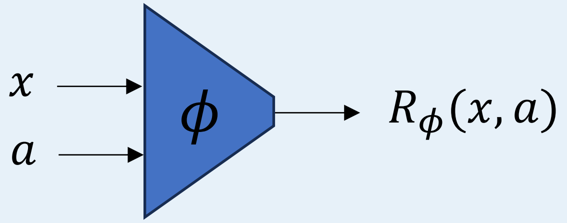
$$\phi \leftarrow \phi - \lambda \nabla_\phi \left[\left(R_\phi(x_t, a_t) - r_t(x_t, a_t) \right)^2 \right]$$

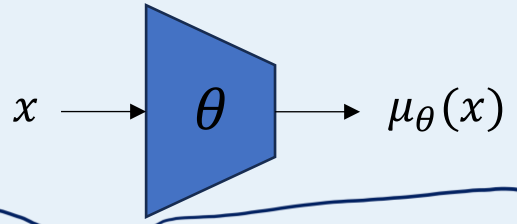
Update policy: $\operatorname*{argmax}_\theta R_\phi(x_t, \mu_\theta(x_t))$

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

$[\mu \leftarrow \mu + \eta \nabla_\mu R_\phi(\mu)]$ (w/o context)

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





$\operatorname*{argmax}_a R_\phi(x,a)$

Continuous Multi-Armed Bandits

Pure policy-based algorithms

	MAB	CB
VB		
PB	•	

Pure Policy-Based Approach

Gradient Ascent

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

Choose action μ_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**?

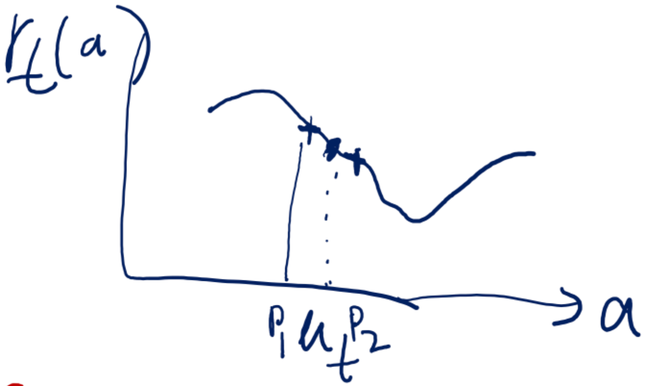
g_t

(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

$r_t : \mathbb{R} \to \mathbb{R}$

actn



$r_t(p_2) > r_t(p_1)$

$r_t(p_1)$

$\propto r_t(p_2)$

Sample $p_1 = \mu_t - \sigma$ get $r_t(p_1)$

$p_2 = \mu_t + \sigma$ $r_t(p_2)$

$$\boxed{g_t^{(2\text{-point})} = \frac{r_t(p_2) - r_t(p_1)}{p_2 - p_1}}$$

Create randomized g_t such that

$$\mathbb{E}(g_t) = g_t^{(2\text{-point})}$$

Sample $a_t \sim \text{unif}(p_1, p_2)$

create
$$\boxed{g_t = \begin{cases} \dfrac{2 r_t(p_2)}{p_2 - p_1} & \text{if } a_t = p_2 \\[2ex] \dfrac{-2 r_t(p_1)}{p_2 - p_1} & \text{if } a_t = p_1 \end{cases}}$$

$= \dfrac{2 r_t(p_2)}{2\sigma}$

$= \dfrac{-2 r_t(p_1)}{2\sigma}$

$\mathbb{E}[g_t]$

$$= \frac{1}{2}\left(\frac{2 r_t(p_2)}{p_2 - p_1}\right) + \frac{1}{2}\left(\frac{-2 r_t(p_1)}{p_2 - p_1}\right)$$

(Nearly) Unbiased Gradient Estimator (1/3)

$e_i = \begin{array}{c} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \end{array} \leftarrow i\text{-th}$

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

$a_t - \mu_t$

$\mu_t + \delta \beta_t e_i$

μ_t

μ_t

(Nearly) Unbiased Gradient Estimator (2/3)

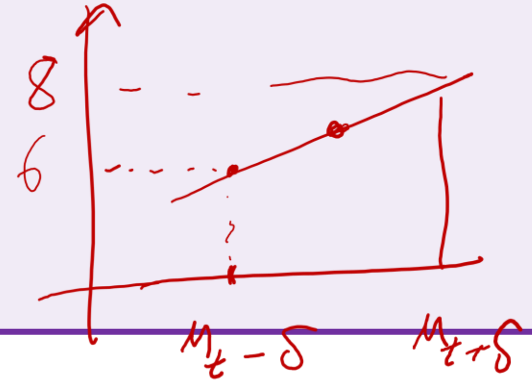
Uniformly randomly choose z_t from the unit sphere $\mathbb{S}_d = \{z \in \mathbb{R}^d : \ \|z\|_2 = 1\}$

Sample $a_t = \mu_t + \delta z_t$

Observe $r_t(a_t)$

Define $g_t = \frac{d(r_t(a_t) - b_t)}{\delta} z_t$

$\boxed{r_t(\mu_t)}$



$$b = 0 \begin{cases} \text{w.p. } \frac{1}{2} \Rightarrow \frac{-6}{\delta} \\ \text{w.p. } \frac{1}{2} \Rightarrow \frac{8}{\delta} \end{cases}$$

$$b = 7 \begin{cases} \text{w.p. } \frac{1}{2} \Rightarrow \frac{-1}{\delta} \\ \text{w.p. } \frac{1}{2} \Rightarrow \frac{+1}{\delta} \end{cases}$$

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

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

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

Update policy:

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

Regret Bound of Gradient Ascent with Gradient Estimator

Theorem. If Ω is convex and all reward functions r_t are concave, then Gradient Ascent with Gradient estimator ensures

$$\text{Regret} = \max_{\mu^\star \in \Omega} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mu^\star) - r_t(\mu_t)\right] \le \frac{\max_{\mu \in \Omega} \|\mu\|_2^2}{\eta} + \eta \sum_{t=1}^{T} \|g_t\|_2^2 + \sum_{t=1}^{T} \text{bias}_t$$

Decrease with δ

Increase with δ

Continuous Contextual Bandits

Pure policy-based algorithms

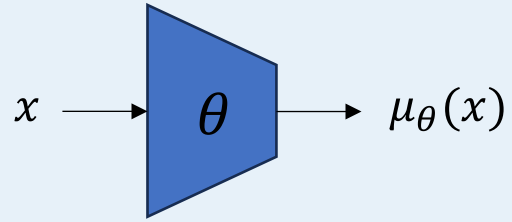
	MAB	CB
VB		
PB		•

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

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

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

Update policy:

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

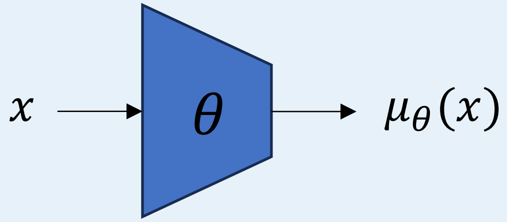
Gradient Ascent with Gradient Estimator (PG)

Gradient Ascent with Gradient Estimator (PG)

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

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:

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

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

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

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

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

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

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

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

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

$\boldsymbol{c.f.}$ **The other approach:**

Create g_t as a gradient estimator in the action space (by sampling around **mean action** μ_θ)

Then construct gradient estimator in the parameter space as $\nabla_\theta \langle \mu_\theta, g_t \rangle$

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:

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

Question

What about PPO objective

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

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

c.f. the PG objective:

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

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

Summary for Bandits

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