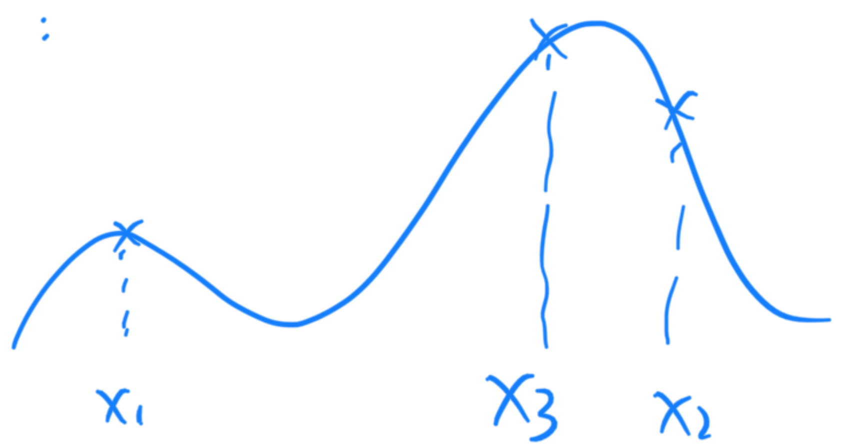
Model Selection in Bandits

Motivation

- MAB (K arms)
- CAB (infinite action/arm set)
  - linear bandits
  - kernelized ~
  - GP ~
  - Lipschitz ~

Goal:



reward function

$f(x)$

$$R(T) = \sum_{t=1}^{T} [f(x^*) - f(x_t)]$$

pseudo regret

$$\mathbb{E}[R(T)] = \mathbb{E}\left[\sum_{t=1}^{T} f(x^*) - f(x_t) \right]$$

expected regret

Common Assumption: $f \in \mathcal{F}$

the environment (function space) is exactly known

E.g. ① linear $f(x) = \langle \theta, x \rangle$

we know. d $\theta \in \mathbb{R}^d$

② GP-bandit

$f(x) \sim GP(0, \Sigma(k))$

GP-UCB: at each step t

$$x_t \in \arg\max_x \mu_{t-1}(x) + \sqrt{\beta}\,\sigma_t(x)$$

$$\sigma_t(x) = k(x,x) - k_t(x)^T (K_t + \sigma^2 I)^2 k_t(x)$$

Question:

what happens if these parameter(s) is unknown?

① can we do ~ if params are known

② if not, what is the best we can do (what algorithms)

Why is model selection "hard" for bandits

* exploration - exploitation
* lack of knowledge of key parameters of F creats a "hard" problem

---

Model Selection in bandits:

Goal: select the best "suitable" algorithm among M candidates: $B^* \in \{B_1 \dots B_M\}$

"competitive" with B^*

Methods

A. "CORRAL"

- CORRAL
- "smooth" CORRAL
- RBBE "regret bound balancing & elimination"
  * stochastic only
  * better than ModCB

B : "test-based"

- relies on statistical test to check whether a base algorithm is "misspecified"

ModCB
* linear models
* not optimal

"model selection for linear contextual bandit" Foster et al. 2019

[$d_1 < d_2 < d_3 < \dots d_M$]

d^*

CORRAL

Meta Alg

B_1 B_2 ... B_M

M candidate / base algorithms

Meta alg: adversarial bandit algorithm

Preliminary: adversarial bandits

stochastic	adversarial
$[P_1 \dots P_K]$	no such assumptions on static reward distributions
	Compare to: "the best arm in hindsight"

$$\mathrm{Reg}(T) = \mathbb{E}\left[\max_{i\in[K]} \sum_{t=1}^{T} r_{t,i} - \sum r_{t,x_t}\right]$$

randomness in the Alg

Exp3: "Exponential weight Alg for exploration - exploitation

① sample $X_t \sim P_t$

② receive reward r_{t,x_t}

③ estimates reward for all arms based on r_{t,x_t}

④ Updates $P_{t,i} \to P_{t+1,i}$ $\forall i \in [K]$

$$\hat{r}_{t,i} = \frac{\mathbb{I}\{X_t = i\}}{P_{t,i}} r_{t,x_t}$$

$$\hat{S}_{t,i} = \sum_{s=1}^{t} \hat{r}_{s,i} \quad \forall i \in [K]$$

* exponential weights

Then: $$P_{t+1,i} = \frac{\exp(\eta \hat{S}_{t,i})}{\sum_j \exp(\cdot)}$$

Regret: $\mathbb{E}[Reg(T)] = O(\sqrt{KT\log(K)})$

$\left(\eta \propto \sqrt{\log(K)/(TK)} \right)$

Problem with Exp 3: (as Meta Alg)

"exponential" weight can shrink too small and "starve" a base algorithm.

CORRAL (different from Exp 3)

linear weights instead of exponential

"less extreme"

$$P_{t+1,i} \propto \left(-\eta \sum_{s=1}^{t} \hat{r}_{s,i} + Z\right)^{-1}$$

normalization

- - - - - - - - - - - - - - -

Input: M base Algo $\{B_1 \ldots B_M\}$

initial lr η (adaptive η)

- initialize all base algorithms
- $\gamma = 1/T$ (lower bound on sampling prob)

$P_{1,j} = 1/M \quad \forall j = 1 \ldots M$

$\bar{p}_{1,j} = p_{1,j}$

For $t = 1, .., T$ do:

- sample $j_t \sim \bar{p}_t$
- B_{j_t} fr one step
- receive reward

  context

  noise

  $\underline{r_t} = f(x_t \sim B_{j_t}) + \delta_t$
- send feedback

  $\frac{r_t}{\bar{p}_{t,j_t}} \mathbb{1}\{j = j_t\}$ to all B_j $j \in M$

- Update $p_t \to p_{t+1}$ via OMD

  $(y_t, r_t, \bar{p}_t, j_t)$

- smooth $p_t \to \bar{p}_t$

con:
* requires "stability" property

pro:
* very general

"Smooth" CORRAL Pacchiano et al. 2020

* base algos only upates when selected

* original reward r_{t,i_t} back to base
* only works in stochastic settings
* memory : a nightmare

smooth wrapper

Base B_j

if selected at t:

$x_t \sim B_j$

$r_t = f(x_t) + \xi_t$

$r_t^{(2)}(i^*)$ is bounded w.h.p

Smoothed B_j

(internal state s)

if selected at t:

step 1:

$x_t^{(1)} \sim B_j$

$r_t^{(1)} = f(x_t^{(1)}) + \xi_t$

step 2:

$q \sim \text{uniform}(1, 2 \dots s)$

$x_t^{(2)} \sim B_{j,q}$

$r_t^{(2)} = f \cdots -$

update $s = s+1$

(expected)

Regret decomposition

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} f(x^*) - f(x_t)\right]$$

$$= \mathbb{E}\left[\sum^{T} f(x^*) - f(x_{i_t^*, t})\right]$$

$$\underbrace{\quad}_{t=1}\quad \#\mathrm{I}$$

$$+\ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} f(x_{i^*,t}) - f(x_t)\right]}_{\#\mathrm{II}}$$

$\#\mathrm{II}$: "model selection" cost

$\propto \mathrm{poly}(M)$

$\#\mathrm{I}$. regret of base i^*

w.r.t optimal action/policy

* has high-probability bounds!

- Assumes all B_i, $i \in [M]$ has high-prob bound $U_i(t, \delta)$ w.p. $1-\delta$ if not misspecified

- $R(T) = \tilde{O}(\mathrm{poly}(M)\, \underline{U_{i^*}(t, \delta)})$ w.p. $1-\delta$

Linear bandit results (model-selection)

$- \,(\ldots\,\ldots)$ if d^* is known

$O(d_* \sqrt{T})$ if ...

& $\mathcal{X}$ is infinite (action space)

smooth CORRAL

$$\tilde{O}(d_*^2 \sqrt{T}) \qquad d_* \le D$$

$$d_{1,2\ldots M} = [1, 2, 2^2, \ldots 2^{\ln(D)}]$$

Mod CB: $\tilde{O}\left((K^{\frac{1}{4}} T^{\frac{3}{4}}) + \sqrt{KTd_*}\right)$

* only works under finite-arm setting

K arms

RBBE : $\tilde{O}(d_*^2 \sqrt{T})$

lower bound :

$\Omega(\sqrt{dT})$ for $d \le \sqrt{T}$

finite (K) armed setting

More complex settings for model selection

{ Lipschitz constant L

{ smoothness,

kernel parameters { lengthscale, etc..
for GP / kernelized bandits

...