Robust algorithms for sequential learning with bandit feedback

Julia Olkhovskaya

Joint work with Gergely Neu, Universitat Pompeu Fabra

Outline

- Multi-armed bandit problem
  - Stochastic setting
  - Adversarial setting
- Contextual bandits
- Learning in episodic MDP
  - MDP basics
  - Online Q-REPS

General sequential learning

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

- The learner chooses **action** A_t,
- The learner gains **reward** $r_t(A_t)$

and observes some **feedback**

General sequential learning

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

- The learner chooses **action** A_t,
- The learner gains **reward** $r_t(A_t)$

and observes some **feedback**

Goal: maximize $\sum_t r_t(A_t)$

The multi-armed bandit

One-Armed Bandit
= Slot Machine

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

- The learner chooses **action** $A_t \in \{1, \ldots, K\}$,
- The learner gains and observes **reward** $r_t(A_t)$

The multi-armed bandit

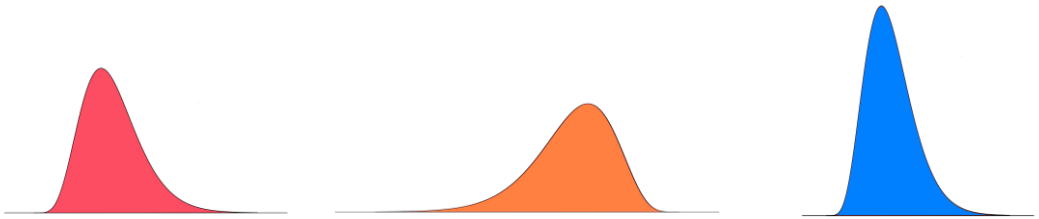
One-Armed Bandit
= Slot Machine

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

- The learner chooses **action** $A_t \in \{1, \ldots, K\}$,
- The learner gains and observes **reward** $r_t(A_t)$

What is r_t?

Stochastic setting



μ_1 μ_2 μ_3

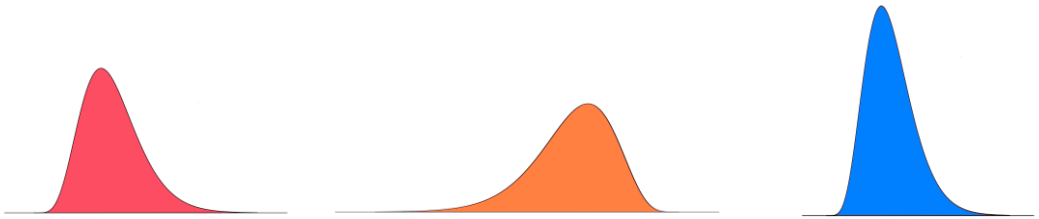
$$r_t(A) \sim \mathcal{D}_A$$

$$\mathbb{E}[r_t(A)] = \mu_A$$

Goal: minimize regret

$$R_T = \max_A \sum_t^T (\mu_A - \mathbb{E}[\mu_{A_t}])$$

Stochastic setting



μ_1 μ_2 μ_3

$$r_t(A) \sim \mathcal{D}_A$$

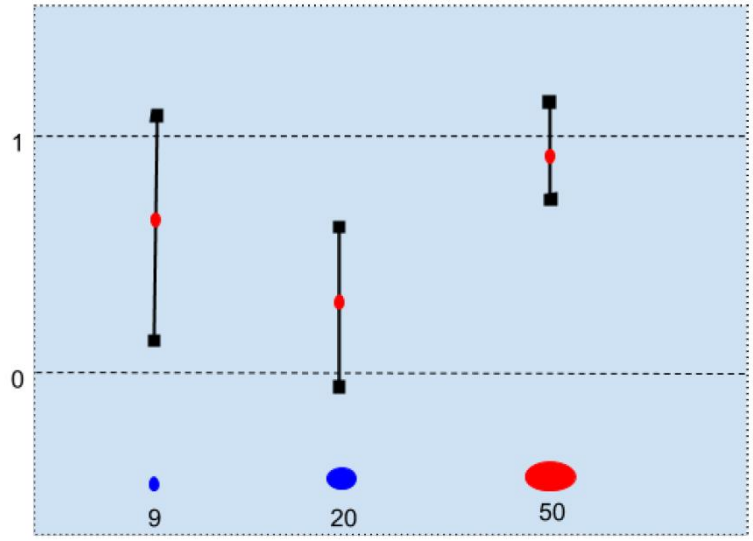
$$\mathbb{E}[r_t(A)] = \mu_A$$

Goal: minimize regret

$$R_T = \max_A \sum_t^T (\mu_A - \mathbb{E}[\mu_{A_t}])$$

UCB Algorithm:

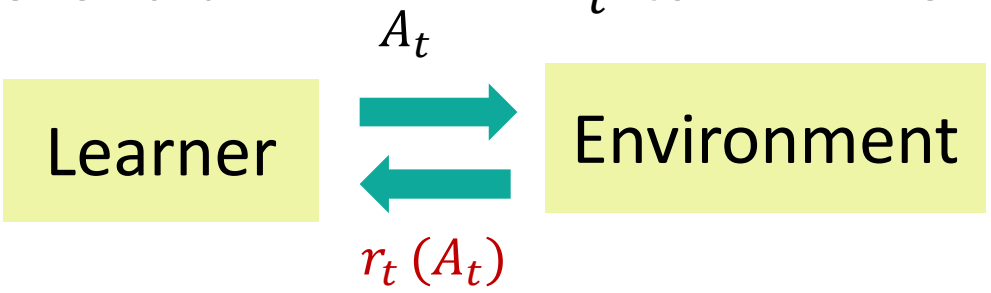
- play $A_t = argmax_A \ \hat{\mu}_{A,t} + \sqrt{\log(t)\,/N_A(t)}$



Adversarial setting

The learner aims to choose A_t to maximize reward

The adversary aims to choose r_t to minimize reward



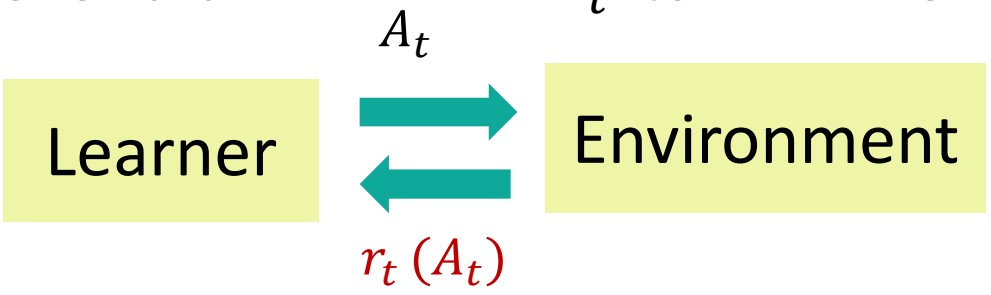
Regret:

$$R_T = \max_A \sum_t^T r_{t,A} - \mathbb{E}\left[\sum_t^T r_{t,A_t}\right]$$

Adversarial setting

The learner aims to choose A_t to maximize reward

The adversary aims to choose r_t to minimize reward



Regret:

$$R_T = \max_A \sum_t^T r_{t,A} - \mathbb{E}\left[\sum_t^T r_{t,A_t}\right]$$

EXP3 Algorithm:

- play $A_t \sim \exp(\sum_{s=1}^{t-1} r_{t,A} I\{A_s = A\})$

- Multi-armed bandit problem
  - Stochastic setting
  - Adversarial setting
- Contextual bandits
- Learning in episodic MDP
  - MDP basics
  - Online Q-REPS

Example: Ad placement

Repeat:

- Website is visited by a **user** (with cookies, profile, etc),
- website chooses **ad** to present to user,
- user **responds** (clicks, leaves, ..)

Goal: achieve the desired user behavior (e.g. maximize clicks)

Linear Contextual Bandits

In each round t $= 1, 2, \ldots, T$

- Nature reveals the context $X_t \in \mathbb{R}^d, X_t \sim \mathcal{D}$
- the learner chooses action $A_t \in [K]$
- adversary picks loss function ℓ_t
- the learner suffers loss $\ell_t(X_t, A_t) = \langle X_t, \theta_{t,A_t} \rangle$

Goal: minimize regret

$$R_T(\pi) = \mathbb{E} \sum_t \left(\ell_t(X_t, A_t) - \ell_t\big(X_t, \pi(X_t)\big) \right)$$

against best policy $\pi: \mathcal{X} \to [K]$

Linear Contextual Bandits

In each round t $= 1, 2, \ldots, T$

- Nature reveals the context $X_t \in \mathbb{R}^d, X_t \sim \mathcal{D}$
- the learner chooses action $A_t \in [K]$
- adversary picks loss function ℓ_t
- the learner suffers loss $\ell_t(X_t, A_t) = \langle X_t, \theta_{t,A_t} \rangle$

Goal: minimize regret

$$R_T(\pi) = \mathbb{E} \sum_t \left(\ell_t(X_t, A_t) - \ell_t\big(X_t, \pi(X_t)\big) \right)$$

against best policy $\pi: \mathcal{X} \to [K]$

Existing results

Best known regret bounds

	i.i.d. context	**adversarial context**
i.i.d. loss	$\sqrt{dKT}$	$\sqrt{dKT}$
adversarial loss	$\sqrt{dKT}$ this work	$\Theta(T)$

Existing results

Best known regret bounds

	i.i.d. context	adversarial context
i.i.d. loss	$\sqrt{dKT}$	$\sqrt{dKT}$
adversarial loss	$\sqrt{dKT}$ this work	$\Theta(T)$

Theorem 1

There is an efficient algorithm with regret of order

$$\sqrt{dKT\log(KT)}$$

for adversarially chosen linear losses and i.i.d. contexts

Algorithm: LinExp3

$$\pi_t(a|x) \propto e^{-\eta\langle x, \widehat{\Theta}_{t-1,a}\rangle}$$

Algorithm: LinExp3

$$\pi_t(a|x) \propto e^{-\eta\langle x, \widehat{\Theta}_{t-1,a}\rangle}$$

$$\widehat{\Theta}_{t,a} = \sum_{k=1}^{t-1} \hat{\theta}_{k,a}$$

Algorithm: LinExp3

$$\pi_t(a|x) \propto e^{-\eta\langle x, \widehat{\Theta}_{t-1,a}\rangle}$$

$$\widehat{\Theta}_{t,a} = \sum_{k=1}^{t-1} \hat{\theta}_{k,a}$$

How to construct $\hat{\theta}_{t,a}$?

Algorithm: LinExp3

$$\pi_t(a|x) \propto e^{-\eta\langle x, \widehat{\Theta}_{t-1,a}\rangle}$$

$$\widehat{\Theta}_{t,a} = \sum_{k=1}^{t-1} \hat{\theta}_{k,a}$$

How to construct $\hat{\theta}_{t,a}$?

Ideal estimator:

$$\Sigma_{t,a} = \mathbb{E}_t[XX^T \mathbb{I}\{\pi_t(X) = a\}]$$

$$\hat{\theta}_{t,a} = \Sigma_{t,a}^{-1} X_t \ell(X_t, a) \mathbb{I}\{\pi_t(X_t) = a\}$$

Algorithm: LinExp3

$$\pi_t(a|x) \propto e^{-\eta\langle x, \widehat{\Theta}_{t-1,a}\rangle}$$

$$\widehat{\Theta}_{t,a} = \sum_{k=1}^{t-1} \hat{\theta}_{k,a}$$

How to construct $\hat{\theta}_{t,a}$?

Ideal estimator:

$$\Sigma_{t,a} = \mathbb{E}_t[XX^T \mathbb{I}\{\pi_t(X) = a\}]$$

$$\hat{\theta}_{t,a} = \Sigma_{t,a}^{-1} X_t \ell(X_t, a) \mathbb{I}\{\pi_t(X_t) = a\}$$

Question: how to estimate $\Sigma_{t,a}^{-1}$?

Matrix Geometric Resampling

$$\Sigma_{t,a}^{-1} = \beta \sum_{k=0}^{\infty} \left(1 - \beta\, \Sigma_{t,a}\right)^{k}$$

Matrix Geometric Resampling

$$\Sigma_{t,a}^{-1} = \beta \sum_{k=0}^{\infty} \left(1 - \beta\, \Sigma_{t,a}\right)^{k}$$

$$B_{k,a} = X(k)X(k)^{T}\, \mathbb{I}\{\pi(X(k)) = a\}$$

$$\hat{\Sigma}_{t,a}^{-1} = \beta \sum_{k=0}^{\infty} \prod_{j=0}^{k} \left(1 - \beta\, \mathrm{B}_{j,a}\right)$$

Matrix Geometric Resampling

$$\Sigma_{t,a}^{-1} = \beta \sum_{k=0}^{\infty} \left(1 - \beta\, \Sigma_{t,a}\right)^{k}$$

$$B_{k,a} = X(k)X(k)^{T}\, \mathbb{I}\{\pi(X(k)) = a\}$$

$$\hat{\Sigma}_{t,a}^{-1} = \beta \sum_{k=0}^{M} \prod_{j=0}^{k} \left(1 - \beta\, \mathrm{B}_{j,a}\right)$$

$$\hat{\theta}_{t,a} = \hat{\Sigma}_{t,a}^{-1}\, X_t X_t^{T} \theta_{t,a}\, \mathbb{I}\{\pi_t(X_t) = a\}$$

Main result 2

What happens when the losses are only nearly-linear?

$$\ell_t(X_t, a) = \langle X_t, \theta_{t,a} \rangle + \varepsilon_{t,a}(X_t)$$

Main result 2

What happens when the losses are only nearly-linear?

$$\ell_t(X_t, a) = \langle X_t, \theta_{t,a} \rangle + \varepsilon_{t,a}(X_t)$$

Theorem 2

Assuming that $\varepsilon_{t,a}(X_t) \leq \varepsilon$, there is an efficient algorithm with regret of order

$$T^{2/3}(dK \log(KT))^{1/3} + \varepsilon\sqrt{d}T$$

for adversarially chosen losses and i.i.d. contexts

Algorithm: LinExp3

$$\pi_t(a|x) \propto e^{-\eta\langle x, \widehat{\Theta}_{t-1,a}\rangle}$$

$$\widehat{\Theta}_{t,a} = \sum_{k=1}^{t-1} \hat{\theta}_{k,a}$$

$\mathbb{E}_t[\hat{\theta}_{t,a}] - \theta_{t,a}$ **can be huge!**

Algorithm: LinExp3

$$\pi_t(a|x) \propto e^{-\eta \langle x, \widehat{\Theta}_{t-1,a} \rangle}$$

$$\widehat{\Theta}_{t,a} = \sum_{k=1}^{t-1} \hat{\theta}_{k,a}$$

$\mathbb{E}_t[\hat{\theta}_{t,a}] - \theta_{t,a}$ **can be huge!**

New estimator:

$\Sigma = \mathbb{E}_t[XX^\top]$ (assumed to be known)

$$\hat{\theta}_{t,a} = \frac{1}{\pi_t(a|X_t)} \cdot \Sigma^{-1} X_t \ell_t(X_t, a) \mathbb{I}\{\pi_t(X_t) = a\}$$

Open problems

Is the term $\varepsilon\sqrt{dT}$ optimal?

- Lattimore and Szepesvári (2019): for large K, no linear-bandit algorithm can remove $\varepsilon\sqrt{dT}$
- Foster and Rakhlin (2020) improve dependence to $\varepsilon\sqrt{KT}$ for i.i.d. losses and small K

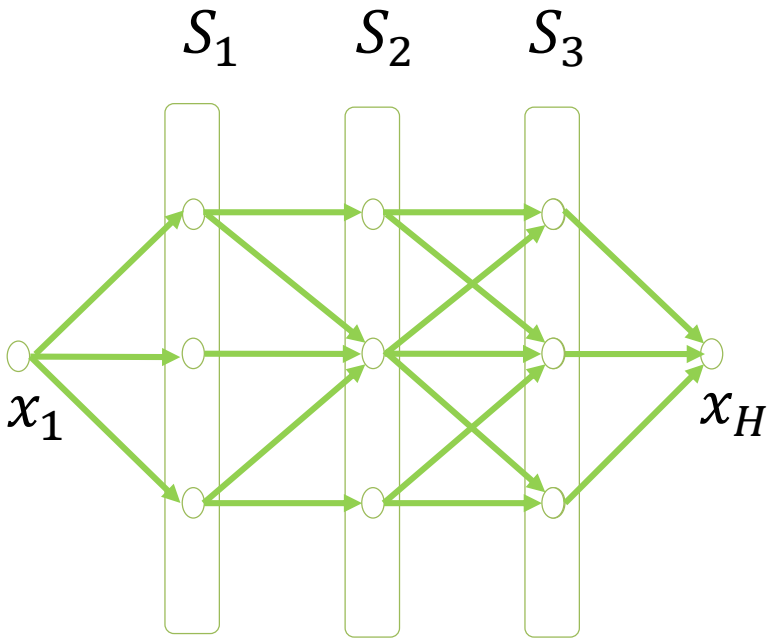
Is it possible to achieve $\sqrt{TdK} + \varepsilon\sqrt{\min\{K, d\}\,T}$ when the losses are adversarial?

- Open problem!

- Multi-armed bandit problem
  - Stochastic setting
  - Adversarial setting
- Contextual bandits
- Learning in episodic MDP
  - MDP basics
  - Online Q-REPS

Episodic MDP with adversarial rewards

- State set $\mathbf{S}$ (not necessary finite)
- Finite action set $\mathbf{A}$
- Episode length $\mathbf{H}$
- Transition function $\mathbf{P}$: $\mathrm{S} \times \mathrm{A} \times \mathrm{S} \rightarrow [0,1]$
- Reward function $\mathbf{r_t}$: $\mathrm{S} \times \mathrm{A} \rightarrow [0,1]$



Assumption [Jin et al. 2020]:

$\exists \varphi(x,a) \in \mathbf{R}^d$:

$\boldsymbol{P}(\cdot\,|x,a) = \langle m(\cdot), \varphi(x,a) \rangle$

$\boldsymbol{r_t}(x,a) = \langle \beta_t, \varphi(x,a) \rangle$

Online learning in episodic MDPs

For each episode: t = 1, …, T:

- Learner chooses policy π_t,
- Adversary selects reward function r_t,
- Learner traverses path $\mathrm{u_t} = (x_{1,t}, a_{1,t}), (x_{2,t}, a_{2,t}), \dots, (x_{H-1,t}, a_{H-1,t})$,
- Learner gains $\sum_{h=1}^{H-1} r_t(x_{h,t}, a_{h,t})$,
- The learner observes $\mathrm{r_t}(x, a)$ for all $(x, a) \in u_t$.

Online learning in episodic MDPs

For each episode: t = 1, …, T:

- Learner chooses policy π_t,
- Adversary selects reward function r_t,
- Learner traverses path $\mathrm{u_t} = (x_{1,t}, a_{1,t}), (x_{2,t}, a_{2,t}), \dots, (x_{H-1,t}, a_{H-1,t})$,
- Learner gains $\sum_{h=1}^{H-1} r_t(x_{h,t}, a_{h,t})$,
- The learner observes $\mathrm{r_t}(x, a)$ for all $(x, a) \in u_t$.

$$\rho_t(\pi) = \mathbf{E}\left[\sum_{h=1}^{H-1} r_t(x_{h,t}, a_{h,t}) \middle| \pi_t = \pi\right]$$

Performance is measured in terms of **regret**:

$$R_T = \max_{\pi} \sum_{t=1}^{T} \mathbf{E}[\rho_t(\pi) - \rho_t(\pi_t)]$$

Previous works

Tabular MDP:

Regret decomposition

[Even-Dar et al. (2009), Neu et al. (2010)]

Idea: decomposition into local bandit problems.

$$\boldsymbol{R}_T = \tilde{O}(H|S|\sqrt{T|A|}/\alpha)$$

α – smallest visitation probability

Linear optimization

[Zimin and Neu (2013), Jin et al. (2020)]

Idea: study occupancy measure

$$\mu^{\pi}(x,a) = P((x,a) \in u_t | \pi_t = \pi)$$

Expected reward: $\langle \mu^{\pi}, \sum_{s=1}^{t-1} r_t \rangle$

$$\boldsymbol{R}_T = \tilde{O}\left(\sqrt{HT|A||S|}\right)$$ no α!

No results on the case of infinite S!

Ideas that may work for infinite state space

MDP with infinite state space:

Regret decomposition

Decomposition into local bandit problems.

\+ Q-functions as rewards for bandit problem, and $Q_t(x,a) = \langle \zeta_t, \varphi(x,a) \rangle$

\- α is still there

Linear optimization

Occupancy measure $\mu^\pi(x,a) = P((x,a) \in u_t | \pi_t = \pi)$.

Expected reward: $\langle \mu^\pi, \sum_{s=1}^{t-1} r_t \rangle$

\- $\mu^\pi(x,a) \in \Delta^{|A|\times|S|}$

\+ ?

Ideas that may work for infinite state space

MDP with infinite state space:

Regret decomposition

Decomposition into local bandit problems.

+ Q-functions as rewards for bandit problem, and $Q_t(x,a) = \langle \zeta_t, \varphi(x,a) \rangle$

- α is still there

Linear optimization

Occupancy measure

$\mu^\pi(x,a) = P((x,a) \in u_t | \pi_t = \pi)$

Expected reward: $\langle \mu^\pi, \sum_{s=1}^{t-1} r_t \rangle$

- $\mu^\pi(x,a) \in \Delta^{|A|\times|S|}$

+ Q-REPS [Bas-Serrano et al. (2020)]:

$\max_{\mu} \langle \mu, \sum_{s=1}^{t-1} r_t \rangle$ ➡ $\min_{\theta \in R^d} G(\theta)$

Q-REPS by Bas-Serrano et al. (2020):

$$\max_{\mu,d} \left\{ \langle \mu, r \rangle \right\}$$

$$s.t.\ \forall h\colon \sum_{a} d_h(x,a) = \sum_{x',a'} P(x|x',a')\mu_h(x',a')\ ,$$

$$\sum_{x,a} d_h(x,a)\,\varphi(x,a) = \sum_{x,a} \mu_h(x,a)\,\varphi(x,a)$$

Q-REPS by Bas-Serrano et al. (2020):

$$\max_{\mu,d} \left\{ \langle \mu, r \rangle - \frac{1}{\eta} D(\mu || \mu_0) - \frac{1}{\alpha} D_c(d || \mu_0) \right\}$$

$$s.t.\ \forall h\colon \sum_a d_h(x,a) = \sum_{x',a'} P(x|x',a') \mu_h(x',a')\ ,$$

$$\sum_{x,a} d_h(x,a)\, \varphi(x,a) = \sum_{x,a} \mu_h(x,a)\, \varphi(x,a)$$

Online Q-REPS:

$$\max_{\mu,d}\left\{\langle \mu, r\rangle - \frac{1}{\eta} D(\mu||\mu_0) - \frac{1}{\alpha} D_c(d||\mu_0)\right\}$$

Lagrange multipliers:

$$s.t.\ \forall h\colon \sum_a d_h(x,a) = \sum_{x',a'} P(x|x',a')\mu_h(x',a')\,, \qquad \times V$$

$$\sum_{x,a} d_h(x,a)\,\varphi(x,a) = \sum_{x,a} \mu_h(x,a)\,\varphi(x,a) \qquad \times \theta$$

$$Q_\theta(x,a) = \langle \theta, \varphi(x,a)\rangle$$

$$V_\theta(x) = \frac{1}{\alpha}\log \sum_a \pi_0(a|x)\, e^{\alpha Q_\theta(x,a)}$$

$$\Delta_\theta(x,a) = r(x,a) + \sum_{x' \in X_{h+1}} P(x'|x,a) V_\theta(x') - Q_\theta(x,a)$$

Solution:

$$\theta^* = argmin_\theta \left\{\frac{1}{\eta}\log \sum_{x,a} \mu_0(x,a) e^{\eta \Delta_\theta(x,a)}\right\}$$

$$\pi^*(a|x) \propto e^{\alpha Q_{\theta^*}(x,a)}$$

d – dimension optimization problem

Online learning

Estimator of $\hat{r}_t$: Matrix Geometric Resampling

$$\Sigma_{t,h} = E_{\pi_t}\left[\varphi\left(X_{t,h}, A_{t,h}\right)\varphi\left(X_{t,h}, A_{t,h}\right)^T\right]$$
$$\hat{\beta}_{t,h} = \Sigma_{t,h}^{-1}\,\varphi\left(X_{t,h}, A_{t,h}\right) r_t(X_{t,h}, A_{t,h})$$

For t=1, …, T:

- $\mu_t, \mathrm{d_t} = \mathrm{argmax}_{\mu,d}\left\{\sum_{s=1}^{t-1}\langle\mu, \hat{r}_s\rangle - \frac{1}{\eta}D(\mu||\mu_0) - \frac{1}{\alpha}D_c(d||\mu_0)\right\}$
  + constraints
- generate path $u_t \sim \pi(d_t)$
- compute $\hat{r}_t$

Online Q-REPS:

$$\mu_t, \mathrm{d_t} = \mathrm{argmax}_{\mu,d} \left\{ \sum_{s=1}^{t-1} \langle \mu, \hat{r}_s \rangle - \frac{1}{\eta} D(\mu || \mu_0) - \frac{1}{\alpha} D_c(d || \mu_0) \right\}$$

Lagrange multipliers:

$$s.t.\ \forall h: \sum_a d_h(x,a) = \sum_{x',a'} P(x|x',a') \mu_h(x',a')\ , \qquad \times V$$

$$\sum_{x,a} d_h(x,a)\, \varphi(x,a)\varphi(x,a)^T = \sum_{x,a} \mu_h(x,a)\, \varphi(x,a)\varphi(x,a)^T \qquad \times Z$$

$$Q_Z(x,a) = \varphi(x,a) Z \varphi(x,a)^T$$

$$V_Z(x) = \frac{1}{\alpha} \log \sum_a \pi_0(a|x)\, e^{\alpha Q_Z(x,a)}$$

$$\Delta_Z(x,a) = \sum_{s=1}^{t-1} \hat{r}_s + \sum_{x' \in X_{h+1}} P(x'|x,a) V_Z(x') - Q_Z(x,a)$$

Solution:

$$Z_t^* = argmin_Z \left\{ \frac{1}{\eta} \log \sum_{x,a} \mu_0(x,a) e^{\eta \Delta_Z(x,a)} \right\}$$

$$\pi_t^*(a|x) \propto \mathrm{e}^{\alpha Q_{Z_t^*}(x,a)}$$

d^2 – dimension optimization problem

Online Q-REPS:

$$Z_t^* = argmin_Z \left\{ \frac{1}{\eta} \log \sum_{x,a} \mu_0(x,a) e^{\eta \Delta_Z(x,a)} \right\} = argmin_Z \{G_t(Z)\}$$

$$\pi_t^*(a|x) \propto e^{\alpha Q_{Z_t^*}(x,a)}$$

- convex
- $\alpha + \eta$-smooth
- need oracle to compute $\sum_{x'} P(x'|x,a) V_Z(x')$

Theorem:

Let $\hat{Z}$: $G_t(Z^*) - G_t(\hat{Z}) \leq \varepsilon$. Then,

$$\boldsymbol{R}_T = \tilde{O}\left(\sqrt{dHT}(D(\mu^*||\mu_0) + D_C(d^*||\mu_0)) + \sqrt{\varepsilon} T^{5/4} (dH)^{1/4}\right).$$

Thanks!