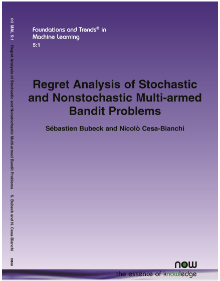
Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems

Sébastien Bubeck
Theory Group

Microsoft Research



Part 1: i.i.d., adversarial, and Bayesian bandit models

i.i.d. multi-armed bandit, Robbins [1952]

i.i.d. multi-armed bandit, Robbins [1952]

Known parameters: number of arms n and (possibly) number of rounds $T \geq n$.

i.i.d. multi-armed bandit, Robbins [1952]

Known parameters: number of arms n and (possibly) number of rounds $T \geq n$.
Unknown parameters: n probability distributions $\nu_1, \ldots, \nu_n$ on $[0, 1]$ with mean $\mu_1, \ldots, \mu_n$ (notation: $\mu^* = \max_{i \in [n]} \mu_i$).

i.i.d. multi-armed bandit, Robbins [1952]

Known parameters: number of arms n and (possibly) number of rounds $T \geq n$.
Unknown parameters: n probability distributions $\nu_1, \ldots, \nu_n$ on $[0,1]$ with mean $\mu_1, \ldots, \mu_n$ (notation: $\mu^* = \max_{i \in [n]} \mu_i$).

Protocol: For each round $t = 1, 2, \ldots, T$, the player chooses $I_t \in [n]$ based on past observations and receives a reward/observation $Y_t \sim \nu_{I_t}$ (independently from the past).

i.i.d. multi-armed bandit, Robbins [1952]

Known parameters: number of arms n and (possibly) number of rounds $T \geq n$.
Unknown parameters: n probability distributions $\nu_1, \ldots, \nu_n$ on $[0, 1]$ with mean $\mu_1, \ldots, \mu_n$ (notation: $\mu^* = \max_{i \in [n]} \mu_i$).

Protocol: For each round $t = 1, 2, \ldots, T$, the player chooses $I_t \in [n]$ based on past observations and receives a reward/observation $Y_t \sim \nu_{I_t}$ (independently from the past).

Performance measure: The cumulative regret is the difference between the player's accumulated reward and the maximum the player could have obtained had she known all the parameters,

$$\overline{R}_T = T\mu^* - \mathbb{E} \sum_{t \in [T]} Y_t.$$

Fundamental tension between **exploration** and **exploitation**. Many applications!

i.i.d. multi-armed bandit: fundamental limitations

How small can we expect $\overline{R}_T$ to be? Consider the 2-armed case where $\nu_1 = \mathrm{Ber}(1/2)$ and $\nu_2 = Ber(1/2 + \xi\Delta)$ where $\xi \in \{-1, 1\}$ is unknown.

i.i.d. multi-armed bandit: fundamental limitations

How small can we expect $\overline{R}_T$ to be? Consider the 2-armed case where $\nu_1 = \mathrm{Ber}(1/2)$ and $\nu_2 = Ber(1/2 + \xi\Delta)$ where $\xi \in \{-1, 1\}$ is unknown.

With τ expected observations from the second arm there is a probability at least $\exp(-\tau\Delta^2)$ to make the wrong guess on the value of ξ.

i.i.d. multi-armed bandit: fundamental limitations

How small can we expect $\overline{R}_T$ to be? Consider the 2-armed case where $\nu_1 = \mathrm{Ber}(1/2)$ and $\nu_2 = \mathit{Ber}(1/2 + \xi\Delta)$ where $\xi \in \{-1, 1\}$ is unknown.

With τ expected observations from the second arm there is a probability at least $\exp(-\tau\Delta^2)$ to make the wrong guess on the value of ξ. Let $\tau(t)$ be the expected number of pulls of arm 2 up to time t when $\xi = -1$.

i.i.d. multi-armed bandit: fundamental limitations

How small can we expect $\overline{R}_T$ to be? Consider the 2-armed case where $\nu_1 = \mathrm{Ber}(1/2)$ and $\nu_2 = Ber(1/2 + \xi\Delta)$ where $\xi \in \{-1, 1\}$ is unknown.

With τ expected observations from the second arm there is a probability at least $\exp(-\tau\Delta^2)$ to make the wrong guess on the value of ξ. Let $\tau(t)$ be the expected number of pulls of arm 2 up to time t when $\xi = -1$.

$$\begin{aligned}
\overline{R}_T(\xi = +1) + \overline{R}_T(\xi = -1) &\geq \Delta\tau(T) + \Delta \sum_{t=1}^{T} \exp(-\tau(t)\Delta^2) \\
&\geq \Delta \min_{t \in [T]} \left(t + T\exp(-t\Delta^2)\right) \\
&\approx \frac{\log(T\Delta^2)}{\Delta}.
\end{aligned}$$

See Bubeck, Perchet and Rigollet [2012] for the details.

i.i.d. multi-armed bandit: fundamental limitations

How small can we expect $\overline{R}_T$ to be? Consider the 2-armed case where $\nu_1 = \mathrm{Ber}(1/2)$ and $\nu_2 = Ber(1/2 + \xi\Delta)$ where $\xi \in \{-1, 1\}$ is unknown.

With τ expected observations from the second arm there is a probability at least $\exp(-\tau\Delta^2)$ to make the wrong guess on the value of ξ. Let $\tau(t)$ be the expected number of pulls of arm 2 up to time t when $\xi = -1$.

$$\begin{aligned}
\overline{R}_T(\xi = +1) + \overline{R}_T(\xi = -1) &\geq \Delta\tau(T) + \Delta \sum_{t=1}^{T} \exp(-\tau(t)\Delta^2) \\
&\geq \Delta \min_{t \in [T]} \left(t + T\exp(-t\Delta^2)\right) \\
&\approx \frac{\log(T\Delta^2)}{\Delta}.
\end{aligned}$$

See Bubeck, Perchet and Rigollet [2012] for the details.

For Δ fixed the lower bound is $\frac{\log(T)}{\Delta}$, and for the worse Δ $(\approx 1/\sqrt{T})$ it is $\sqrt{T}$ (Auer, Cesa-Bianchi, Freund and Schapire [1995]: $\sqrt{Tn}$ for the n-armed case).

i.i.d. multi-armed bandit: fundamental limitations

Notation: $\Delta_i = \mu^* - \mu_i$ and $N_i(t)$ is the number of pulls of arm i up to time t. Then one has $\overline{R}_T = \sum_{i=1}^n \Delta_i \mathbb{E} N_i(T)$.

i.i.d. multi-armed bandit: fundamental limitations

Notation: $\Delta_i = \mu^* - \mu_i$ and $N_i(t)$ is the number of pulls of arm i up to time t. Then one has $\overline{R}_T = \sum_{i=1}^n \Delta_i \mathbb{E} N_i(T)$.

$$\text{For } \; p, q \in [0,1], \;\; \mathrm{kl}(p,q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

i.i.d. multi-armed bandit: fundamental limitations

Notation: $\Delta_i = \mu^* - \mu_i$ and $N_i(t)$ is the number of pulls of arm i up to time t. Then one has $\overline{R}_T = \sum_{i=1}^{n} \Delta_i \mathbb{E} N_i(T)$.

$$\text{For } \; p, q \in [0,1], \;\; \mathrm{kl}(p,q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

Theorem (Lai and Robbins [1985])

Consider a strategy s.t. $\forall a > 0$, we have $\mathbb{E} N_i(T) = o(T^a)$ if $\Delta_i > 0$. Then for any Bernoulli distributions,

$$\liminf_{T \to +\infty} \frac{\overline{R}_T}{\log(T)} \geq \sum_{i : \Delta_i > 0} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu^*)}.$$

i.i.d. multi-armed bandit: fundamental limitations

Notation: $\Delta_i = \mu^* - \mu_i$ and $N_i(t)$ is the number of pulls of arm i up to time t. Then one has $\overline{R}_T = \sum_{i=1}^{n} \Delta_i \mathbb{E} N_i(T)$.

$$\text{For } \; p, q \in [0,1], \;\; \mathrm{kl}(p,q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

Theorem (Lai and Robbins [1985])

Consider a strategy s.t. $\forall a > 0$, we have $\mathbb{E} N_i(T) = o(T^a)$ if $\Delta_i > 0$. Then for any Bernoulli distributions,

$$\liminf_{T \to +\infty} \frac{\overline{R}_T}{\log(T)} \geq \sum_{i : \Delta_i > 0} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu^*)}.$$

Note that $\frac{1}{2\Delta_i} \geq \frac{\Delta_i}{\mathrm{kl}(\mu_i,\mu^*)} \geq \frac{\mu^*(1-\mu^*)}{2\Delta_i}$ so up to a variance-like term the Lai and Robbins lower bound is $\sum_{i : \Delta_i > 0} \frac{\log(T)}{2\Delta_i}$.

i.i.d. multi-armed bandit: fundamental strategy

Hoeffding's inequality: w.p. $\geq 1 - 1/T$, $\forall t \in [T], i \in [n]$,

$$\mu_i \leq \frac{1}{N_i(t)} \sum_{s<t:I_s=i} Y_s + \sqrt{\frac{2\log(T)}{N_i(t)}} =: \mathrm{UCB}_i(t).$$

i.i.d. multi-armed bandit: fundamental strategy

Hoeffding's inequality: w.p. $\geq 1 - 1/T$, $\forall t \in [T], i \in [n]$,

$$\mu_i \leq \frac{1}{N_i(t)} \sum_{s<t : I_s = i} Y_s + \sqrt{\frac{2\log(T)}{N_i(t)}} =: \mathrm{UCB}_i(t).$$

UCB (Upper Confidence Bound) strategy (Lai and Robbins [1985], Agarwal [1995], Auer, Cesa-Bianchi and Fischer [2002]):

$$I_t \in \operatorname*{argmax}_{i \in [n]} \mathrm{UCB}_i(t).$$

i.i.d. multi-armed bandit: fundamental strategy

Hoeffding's inequality: w.p. $\geq 1 - 1/T$, $\forall t \in [T], i \in [n]$,

$$\mu_i \leq \frac{1}{N_i(t)} \sum_{s<t: I_s = i} Y_s + \sqrt{\frac{2\log(T)}{N_i(t)}} =: \mathrm{UCB}_i(t).$$

UCB (Upper Confidence Bound) strategy (Lai and Robbins [1985], Agarwal [1995], Auer, Cesa-Bianchi and Fischer [2002]):

$$I_t \in \operatorname*{argmax}_{i \in [n]} \mathrm{UCB}_i(t).$$

Simple analysis: on a $1 - 2/T$ probability event one has

$$N_i(t) \geq 8\log(T)/\Delta_i^2 \Rightarrow \mathrm{UCB}_i(t) < \mu^* \leq \mathrm{UCB}_{i^*}(t),$$

i.i.d. multi-armed bandit: fundamental strategy

Hoeffding's inequality: w.p. $\geq 1 - 1/T$, $\forall t \in [T], i \in [n]$,

$$\mu_i \leq \frac{1}{N_i(t)} \sum_{s<t: I_s = i} Y_s + \sqrt{\frac{2\log(T)}{N_i(t)}} =: \mathrm{UCB}_i(t).$$

UCB (Upper Confidence Bound) strategy (Lai and Robbins [1985], Agarwal [1995], Auer, Cesa-Bianchi and Fischer [2002]):

$$I_t \in \operatorname*{argmax}_{i \in [n]} \mathrm{UCB}_i(t).$$

Simple analysis: on a $1 - 2/T$ probability event one has

$$N_i(t) \geq 8\log(T)/\Delta_i^2 \Rightarrow \mathrm{UCB}_i(t) < \mu^* \leq \mathrm{UCB}_{i^*}(t),$$

so that $\mathbb{E}N_i(T) \leq 2 + 8\log(T)/\Delta_i^2$ and in fact

$$\overline{R}_T \leq 2 + \sum_{i: \Delta_i > 0} \frac{8\log(T)}{\Delta_i}.$$

i.i.d. multi-armed bandit: going further

1. Optimal constant (replacing 8 by 1/2 in the UCB regret bound) and Lai and Robbins variance-like term (replacing Δ_i by $\mathrm{kl}(\mu_i, \mu^*)$): see Cappé, Garivier, Maillard, Munos and Stoltz [2013].

i.i.d. multi-armed bandit: going further

1. Optimal constant (replacing 8 by $1/2$ in the UCB regret bound) and Lai and Robbins variance-like term (replacing Δ_i by $\mathrm{kl}(\mu_i, \mu^*)$): see Cappé, Garivier, Maillard, Munos and Stoltz [2013].
2. In many applications one is merely interested in *finding* the best arm (instead of maximizing cumulative reward): this is the best arm identification problem. For the fundamental strategies see Even-Dar, Mannor and Mansour [2006] for the fixed-confidence setting (see also Jamieson and Nowak [2014] for a recent short survey) and Audibert, Bubeck and Munos [2010] for the fixed budget setting. Key takeaway: one needs of order $\mathbf{H} := \sum_i \Delta_i^{-2}$ rounds to find the best arm.

i.i.d. multi-armed bandit: going further

1. Optimal constant (replacing 8 by $1/2$ in the UCB regret bound) and Lai and Robbins variance-like term (replacing Δ_i by $\mathrm{kl}(\mu_i, \mu^*)$): see Cappé, Garivier, Maillard, Munos and Stoltz [2013].
2. In many applications one is merely interested in *finding* the best arm (instead of maximizing cumulative reward): this is the best arm identification problem. For the fundamental strategies see Even-Dar, Mannor and Mansour [2006] for the fixed-confidence setting (see also Jamieson and Nowak [2014] for a recent short survey) and Audibert, Bubeck and Munos [2010] for the fixed budget setting. Key takeaway: one needs of order $\mathbf{H} := \sum_i \Delta_i^{-2}$ rounds to find the best arm.
3. The UCB analysis extends to sub-Gaussian reward distributions. For heavy-tailed distributions, say with $1+\varepsilon$ moment for some $\varepsilon \in (0,1]$, one can get a regret that scales with $\Delta_i^{-1/\varepsilon}$ (instead of Δ_i^{-1}) by using a robust mean estimator, see Bubeck, Cesa-Bianchi and Lugosi [2012].

Adversarial multi-armed bandit, Auer, Cesa-Bianchi, Freund and Schapire [1995, 2001]

For $t = 1, \ldots, T$, the player chooses $I_t \in [n]$ based on previous observations, and simultaneously an adversary chooses a loss vector $\ell_t \in [0,1]^n$. The player's loss/observation is $\ell_t(I_t)$.

Adversarial multi-armed bandit, Auer, Cesa-Bianchi, Freund and Schapire [1995, 2001]

For $t = 1, \ldots, T$, the player chooses $I_t \in [n]$ based on previous observations, and simultaneously an adversary chooses a loss vector $\ell_t \in [0,1]^n$. The player's loss/observation is $\ell_t(I_t)$.
The regret and pseudo-regret are defined as:

$$R_T = \max_{i \in [n]} \sum_{t \in [T]} (\ell_t(I_t) - \ell_t(i)), \quad \overline{R}_T = \max_{i \in [n]} \mathbb{E} \sum_{t \in [T]} (\ell_t(I_t) - \ell_t(i)).$$

Adversarial multi-armed bandit, Auer, Cesa-Bianchi, Freund and Schapire [1995, 2001]

For $t = 1, \ldots, T$, the player chooses $I_t \in [n]$ based on previous observations, and simultaneously an adversary chooses a loss vector $\ell_t \in [0,1]^n$. The player's loss/observation is $\ell_t(I_t)$.
The regret and pseudo-regret are defined as:

$$R_T = \max_{i \in [n]} \sum_{t \in [T]} (\ell_t(I_t) - \ell_t(i)), \quad \overline{R}_T = \max_{i \in [n]} \mathbb{E} \sum_{t \in [T]} (\ell_t(I_t) - \ell_t(i)).$$

Obviously $\mathbb{E} R_T \geq \overline{R}_T$ and there is equality in the oblivious case ($\equiv$ adversary's choice are independent of the player's choice). The case where $\ell_1, \ldots, \ell_T$ is an i.i.d. sequence corresponds to the i.i.d. case we just studied. In particular we have a $\sqrt{Tn}$ lower bound.

Adversarial multi-armed bandit, fundamental strategy

Exponential weights strategy for *full information* (ℓ_t is observed at the end of round t): play I_t at random from p_t where

$$p_{t+1}(i) = \frac{1}{Z_{t+1}} p_t(i) \exp(-\eta \ell_t(i)).$$

Adversarial multi-armed bandit, fundamental strategy

Exponential weights strategy for *full information* (ℓ_t is observed at the end of round t): play I_t at random from p_t where

$$p_{t+1}(i) = \frac{1}{Z_{t+1}} p_t(i) \exp(-\eta \ell_t(i)).$$

In five lines one can show $\overline{R}_T \leq \sqrt{2T \log(n)}$ with $p_1(i) = 1/n$:

Adversarial multi-armed bandit, fundamental strategy

Exponential weights strategy for *full information* (ℓ_t is observed at the end of round t): play I_t at random from p_t where

$$p_{t+1}(i) = \frac{1}{Z_{t+1}} p_t(i) \exp(-\eta \ell_t(i)).$$

In five lines one can show $\overline{R}_T \leq \sqrt{2T \log(n)}$ with $p_1(i) = 1/n$:

$$\mathrm{Ent}(\delta_j \| p_t) - \mathrm{Ent}(\delta_j \| p_{t+1}) = \log \frac{p_{t+1}(j)}{p_t(j)} = \log \frac{1}{Z_{t+1}} - \eta \ell_t(j)$$

Adversarial multi-armed bandit, fundamental strategy

Exponential weights strategy for *full information* (ℓ_t is observed at the end of round t): play I_t at random from p_t where

$$p_{t+1}(i) = \frac{1}{Z_{t+1}} p_t(i) \exp(-\eta \ell_t(i)).$$

In five lines one can show $\overline{R}_T \leq \sqrt{2T \log(n)}$ with $p_1(i) = 1/n$:

$$\mathrm{Ent}(\delta_j \| p_t) - \mathrm{Ent}(\delta_j \| p_{t+1}) = \log \frac{p_{t+1}(j)}{p_t(j)} = \log \frac{1}{Z_{t+1}} - \eta \ell_t(j)$$

$$\psi_t := \log \mathbb{E}_{I \sim p_t} \exp(-\eta(\ell_t(I) - \mathbb{E}_{I' \sim p_t} \ell_t(I'))) = \eta \mathbb{E} \ell_t(I') + \log(Z_{t+1})$$

Adversarial multi-armed bandit, fundamental strategy

Exponential weights strategy for *full information* (ℓ_t is observed at the end of round t): play I_t at random from p_t where

$$p_{t+1}(i) = \frac{1}{Z_{t+1}} p_t(i) \exp(-\eta \ell_t(i)).$$

In five lines one can show $\overline{R}_T \leq \sqrt{2T \log(n)}$ with $p_1(i) = 1/n$:

$$\mathrm{Ent}(\delta_j \| p_t) - \mathrm{Ent}(\delta_j \| p_{t+1}) = \log \frac{p_{t+1}(j)}{p_t(j)} = \log \frac{1}{Z_{t+1}} - \eta \ell_t(j)$$

$$\psi_t := \log \mathbb{E}_{I \sim p_t} \exp(-\eta(\ell_t(I) - \mathbb{E}_{I' \sim p_t} \ell_t(I'))) = \eta \mathbb{E} \ell_t(I') + \log(Z_{t+1})$$

$$\eta \sum_t \left(\sum_i p_t(i) \ell_t(i) - \ell_t(j) \right) = \mathrm{Ent}(\delta_j \| p_1) - \mathrm{Ent}(\delta_j \| p_{T+1}) + \sum_t \psi_t$$

Adversarial multi-armed bandit, fundamental strategy

Exponential weights strategy for *full information* (ℓ_t is observed at the end of round t): play I_t at random from p_t where

$$p_{t+1}(i) = \frac{1}{Z_{t+1}} p_t(i) \exp(-\eta \ell_t(i)).$$

In five lines one can show $\overline{R}_T \leq \sqrt{2T \log(n)}$ with $p_1(i) = 1/n$:

$$\mathrm{Ent}(\delta_j \| p_t) - \mathrm{Ent}(\delta_j \| p_{t+1}) = \log \frac{p_{t+1}(j)}{p_t(j)} = \log \frac{1}{Z_{t+1}} - \eta \ell_t(j)$$

$$\psi_t := \log \mathbb{E}_{I \sim p_t} \exp(-\eta(\ell_t(I) - \mathbb{E}_{I' \sim p_t} \ell_t(I'))) = \eta \mathbb{E} \ell_t(I') + \log(Z_{t+1})$$

$$\eta \sum_t \left(\sum_i p_t(i) \ell_t(i) - \ell_t(j) \right) = \mathrm{Ent}(\delta_j \| p_1) - \mathrm{Ent}(\delta_j \| p_{T+1}) + \sum_t \psi_t$$

Using that $\ell_t \geq 0$ one has $\psi_t \leq \frac{\eta^2}{2} \mathbb{E} \ell_t(i)^2$ thus $\overline{R}_T \leq \frac{\log(n)}{\eta} + \frac{\eta T}{2}$

Adversarial multi-armed bandit, fundamental strategy

Exp3: replace ℓ_t by $\widetilde{\ell}_t$ in the exponential weights strategy, where

$$\widetilde{\ell}_t(i) = \frac{\ell_t(I_t)}{p_t(i)} \mathbb{1}\{i = I_t\}.$$

Key property: $\mathbb{E}_{I_t \sim p_t} \widetilde{\ell}_t(i) = \ell_t(i)$.

Adversarial multi-armed bandit, fundamental strategy

Exp3: replace ℓ_t by $\widetilde{\ell}_t$ in the exponential weights strategy, where

$$\widetilde{\ell}_t(i) = \frac{\ell_t(I_t)}{p_t(i)} \mathbb{1}\{i = I_t\}.$$

Key property: $\mathbb{E}_{I_t \sim p_t} \widetilde{\ell}_t(i) = \ell_t(i)$. Thus with the analysis from the previous slide:

$$\overline{R}_T \leq \frac{\log(n)}{\eta} + \frac{\eta}{2} \mathbb{E} \sum_t \mathbb{E}_{I \sim p_t} \widetilde{\ell}_t(I)^2.$$

Adversarial multi-armed bandit, fundamental strategy

Exp3: replace ℓ_t by $\widetilde{\ell}_t$ in the exponential weights strategy, where

$$\widetilde{\ell}_t(i) = \frac{\ell_t(I_t)}{p_t(i)} \mathbb{1}\{i = I_t\}.$$

Key property: $\mathbb{E}_{I_t \sim p_t} \widetilde{\ell}_t(i) = \ell_t(i)$. Thus with the analysis from the previous slide:

$$\overline{R}_T \leq \frac{\log(n)}{\eta} + \frac{\eta}{2} \mathbb{E} \sum_t \mathbb{E}_{I \sim p_t} \widetilde{\ell}_t(I)^2.$$

Amazingly the variance term is automatically controlled:

$$\mathbb{E}_{I_t, I \sim p_t} \widetilde{\ell}_t(I)^2 \leq \mathbb{E}_{I_t, I \sim p_t} \frac{\mathbb{1}\{I = I_t\}}{p_t(I_t)^2} = \mathbb{E}_{I \sim p_t} \frac{1}{p_t(I)} = n.$$

Adversarial multi-armed bandit, fundamental strategy

Exp3: replace ℓ_t by $\widetilde{\ell}_t$ in the exponential weights strategy, where

$$\widetilde{\ell}_t(i) = \frac{\ell_t(I_t)}{p_t(i)} \mathbb{1}\{i = I_t\}.$$

Key property: $\mathbb{E}_{I_t \sim p_t} \widetilde{\ell}_t(i) = \ell_t(i)$. Thus with the analysis from the previous slide:

$$\overline{R}_T \leq \frac{\log(n)}{\eta} + \frac{\eta}{2} \mathbb{E} \sum_t \mathbb{E}_{I \sim p_t} \widetilde{\ell}_t(I)^2.$$

Amazingly the variance term is automatically controlled:

$$\mathbb{E}_{I_t, I \sim p_t} \widetilde{\ell}_t(I)^2 \leq \mathbb{E}_{I_t, I \sim p_t} \frac{\mathbb{1}\{I = I_t\}}{p_t(I_t)^2} = \mathbb{E}_{I \sim p_t} \frac{1}{p_t(I)} = n.$$

Thus with $\eta = \sqrt{2n \log(n)/T}$ one gets $\overline{R}_T \leq \sqrt{2Tn \log(n)}$.

Adversarial multi-armed bandit, going further

1. With the modified loss estimate $\frac{\ell_t(I_t)\mathbb{1}\{i=I_t\}+\beta}{p_t(I_t)}$ one can prove high probability bounds on R_T, and by integrating the deviations one can show $\mathbb{E}R_T = O(\sqrt{Tn\log(n)})$.

Adversarial multi-armed bandit, going further

1. With the modified loss estimate $\frac{\ell_t(I_t)\mathbb{1}\{i=I_t\}+\beta}{p_t(I_t)}$ one can prove high probability bounds on R_T, and by integrating the deviations one can show $\mathbb{E}R_T = O(\sqrt{Tn\log(n)})$.
2. The extraneous logarithmic factor in the pseudo-regret upper can be removed, see Audibert and Bubeck [2009]. Conjecture: one cannot remove the log factor for the expected regret, that is for any strategy there exists an adaptive adversary such that $\mathbb{E}R_T = \Omega(\sqrt{Tn\log(n)})$.

Adversarial multi-armed bandit, going further

1. With the modified loss estimate $\frac{\ell_t(I_t)\mathbb{1}\{i=I_t\}+\beta}{p_t(I_t)}$ one can prove high probability bounds on R_T, and by integrating the deviations one can show $\mathbb{E}R_T = O(\sqrt{Tn\log(n)})$.
2. The extraneous logarithmic factor in the pseudo-regret upper can be removed, see Audibert and Bubeck [2009]. Conjecture: one cannot remove the log factor for the expected regret, that is for any strategy there exists an adaptive adversary such that $\mathbb{E}R_T = \Omega(\sqrt{Tn\log(n)})$.
3. T can be replaced by various measure of "variance" in the loss sequence, see e.g., Hazan and Kale [2009].

Adversarial multi-armed bandit, going further

1. With the modified loss estimate $\frac{\ell_t(I_t)\mathbb{1}\{i=I_t\}+\beta}{p_t(I_t)}$ one can prove high probability bounds on R_T, and by integrating the deviations one can show $\mathbb{E}R_T = O(\sqrt{Tn\log(n)})$.
2. The extraneous logarithmic factor in the pseudo-regret upper can be removed, see Audibert and Bubeck [2009]. Conjecture: one cannot remove the log factor for the expected regret, that is for any strategy there exists an adaptive adversary such that $\mathbb{E}R_T = \Omega(\sqrt{Tn\log(n)})$.
3. T can be replaced by various measure of "variance" in the loss sequence, see e.g., Hazan and Kale [2009].
4. There exists strategies which guarantee simultaneously $\overline{R}_T = \widetilde{O}(\sqrt{Tn})$ in the adversarial model and $\overline{R}_T = \widetilde{O}(\sum_i \Delta_i^{-1})$ in the i.i.d. model, see Bubeck and Slivkins [2012].

Adversarial multi-armed bandit, going further

1. With the modified loss estimate $\frac{\ell_t(I_t)\mathbb{1}\{i=I_t\}+\beta}{p_t(I_t)}$ one can prove high probability bounds on R_T, and by integrating the deviations one can show $\mathbb{E}R_T = O(\sqrt{Tn\log(n)})$.
2. The extraneous logarithmic factor in the pseudo-regret upper can be removed, see Audibert and Bubeck [2009]. Conjecture: one cannot remove the log factor for the expected regret, that is for any strategy there exists an adaptive adversary such that $\mathbb{E}R_T = \Omega(\sqrt{Tn\log(n)})$.
3. T can be replaced by various measure of "variance" in the loss sequence, see e.g., Hazan and Kale [2009].
4. There exists strategies which guarantee simultaneously $\overline{R}_T = \widetilde{O}(\sqrt{Tn})$ in the adversarial model and $\overline{R}_T = \widetilde{O}(\sum_i \Delta_i^{-1})$ in the i.i.d. model, see Bubeck and Slivkins [2012].
5. Graph feedback structure, regret with respect to S switches, label efficient, switching cost...

Bayesian multi-armed bandit, Thompson [1933]

Set of models $\{(\nu_1(\theta), \ldots, \nu_n(\theta)), \theta \in \Theta\}$ and prior distribution π_0 over Θ. The Bayesian regret is defined as

$$BR_T(\pi_0) = \mathbb{E}_{\theta \sim \pi_0} \overline{R}_T(\nu_1(\theta), \ldots, \nu_n(\theta)).$$

Bayesian multi-armed bandit, Thompson [1933]

Set of models $\{(\nu_1(\theta), \dots, \nu_n(\theta)), \theta \in \Theta\}$ and prior distribution π_0 over Θ. The Bayesian regret is defined as

$$BR_T(\pi_0) = \mathbb{E}_{\theta \sim \pi_0} \overline{R}_T(\nu_1(\theta), \dots, \nu_n(\theta)).$$

In principle the strategy minimizing the Bayesian regret can be computed by dynamic programming on the potentially huge state space $\mathcal{P}(\Theta)$.

Bayesian multi-armed bandit, Thompson [1933]

Set of models $\{(\nu_1(\theta), \dots, \nu_n(\theta)), \theta \in \Theta\}$ and prior distribution π_0 over Θ. The Bayesian regret is defined as

$$BR_T(\pi_0) = \mathbb{E}_{\theta \sim \pi_0} \overline{R}_T(\nu_1(\theta), \dots, \nu_n(\theta)).$$

In principle the strategy minimizing the Bayesian regret can be computed by dynamic programming on the potentially huge state space $\mathcal{P}(\Theta)$. The celebrated Gittins index theorem gives sufficient condition to dramatically reduce the computational complexity of implementing the optimal Bayesian strategy under a strong product assumption on π_0.

Bayesian multi-armed bandit, Thompson [1933]

Set of models $\{(\nu_1(\theta), \dots, \nu_n(\theta)), \theta \in \Theta\}$ and prior distribution π_0 over Θ. The Bayesian regret is defined as

$$BR_T(\pi_0) = \mathbb{E}_{\theta \sim \pi_0} \overline{R}_T(\nu_1(\theta), \dots, \nu_n(\theta)).$$

In principle the strategy minimizing the Bayesian regret can be computed by dynamic programming on the potentially huge state space $\mathcal{P}(\Theta)$. The celebrated Gittins index theorem gives sufficient condition to dramatically reduce the computational complexity of implementing the optimal Bayesian strategy under a strong product assumption on π_0.

Notation: π_t denotes the posterior distribution on θ at time t.

Bayesian multi-armed bandit, Gittins index

Theorem (Gittins [1979])

Consider the product and γ-discounted case: $\Theta = \times_i \Theta_i$, $\nu_i(\theta) := \nu(\theta_i)$, $\pi_0 = \otimes_i \pi_0(i)$, and furthermore one is interested in maximizing $\mathbb{E} \sum_{t \geq 0} \gamma^t Y_t$.

Bayesian multi-armed bandit, Gittins index

Theorem (Gittins [1979])

Consider the product and γ-discounted case: $\Theta = \times_i \Theta_i$, $\nu_i(\theta) := \nu(\theta_i)$, $\pi_0 = \otimes_i \pi_0(i)$, and furthermore one is interested in maximizing $\mathbb{E} \sum_{t \geq 0} \gamma^t Y_t$. The optimal Bayesian strategy is to pick at time s the arm maximizing:

$$\sup \left\{ \lambda \in \mathbb{R} : \sup_{\tau} \mathbb{E} \left(\sum_{t < \tau} \gamma^t X_t + \frac{\gamma^\tau}{1-\gamma} \lambda \right) \geq \frac{1}{1-\gamma} \lambda \right\},$$

where the expectation is over (X_t) drawn from $\nu(\theta)$ with $\theta \sim \pi_s(i)$, and the supremum is taken over all stopping times τ.

Bayesian multi-armed bandit, Gittins index

Theorem (Gittins [1979])

Consider the product and γ-discounted case: $\Theta = \times_i \Theta_i$, $\nu_i(\theta) := \nu(\theta_i)$, $\pi_0 = \otimes_i \pi_0(i)$, and furthermore one is interested in maximizing $\mathbb{E} \sum_{t \geq 0} \gamma^t Y_t$. The optimal Bayesian strategy is to pick at time s the arm maximizing:

$$\sup \left\{ \lambda \in \mathbb{R} : \sup_{\tau} \mathbb{E} \left(\sum_{t < \tau} \gamma^t X_t + \frac{\gamma^\tau}{1-\gamma} \lambda \right) \geq \frac{1}{1-\gamma} \lambda \right\},$$

where the expectation is over (X_t) drawn from $\nu(\theta)$ with $\theta \sim \pi_s(i)$, and the supremum is taken over all stopping times τ.

For much more (implementation for exponential families, interpretation as a multitoken Markov game, ...) see Dumitriu, Tetali and Winkler [2003], Gittins, Glazebrook, Weber [2011], Kaufmann [2014].

Bayesian multi-armed bandit, Gittins index

Weber [1992] gives an exquisite proof of Gittins theorem. Let

$$\lambda_t(i) := \sup \left\{ \lambda \in \mathbb{R} : \sup_{\tau} \mathbb{E} \sum_{t<\tau} \gamma^t (X_t - \lambda) \geq 0 \right\}$$

the Gittins index of arm i at time t, which we interpret as the *maximum charge* one is willing to pay to play arm i given the current information.

Bayesian multi-armed bandit, Gittins index

Weber [1992] gives an exquisite proof of Gittins theorem. Let

$$\lambda_t(i) := \sup \left\{ \lambda \in \mathbb{R} : \sup_{\tau} \mathbb{E} \sum_{t<\tau} \gamma^t (X_t - \lambda) \geq 0 \right\}$$

the Gittins index of arm i at time t, which we interpret as the *maximum charge* one is willing to pay to play arm i given the current information. The *prevailing charge* is defined as $\min_{s \leq t} \lambda_s(i)$ (i.e. whenever the prevailing charge is too high we just drop it to the fair level).

Bayesian multi-armed bandit, Gittins index

Weber [1992] gives an exquisite proof of Gittins theorem. Let

$$\lambda_t(i) := \sup \left\{ \lambda \in \mathbb{R} : \sup_{\tau} \mathbb{E} \sum_{t<\tau} \gamma^t (X_t - \lambda) \geq 0 \right\}$$

the Gittins index of arm i at time t, which we interpret as the *maximum charge* one is willing to pay to play arm i given the current information. The *prevailing charge* is defined as $\min_{s \leq t} \lambda_s(i)$ (i.e. whenever the prevailing charge is too high we just drop it to the fair level).

1. Discounted sum of prevailing charge for played arms is an upper bound (in expectation) on the discounted sum of rewards.

Bayesian multi-armed bandit, Gittins index

Weber [1992] gives an exquisite proof of Gittins theorem. Let

$$\lambda_t(i) := \sup \left\{ \lambda \in \mathbb{R} : \sup_{\tau} \mathbb{E} \sum_{t<\tau} \gamma^t (X_t - \lambda) \geq 0 \right\}$$

the Gittins index of arm i at time t, which we interpret as the *maximum charge* one is willing to pay to play arm i given the current information. The *prevailing charge* is defined as $\min_{s \leq t} \lambda_s(i)$ (i.e. whenever the prevailing charge is too high we just drop it to the fair level).

1. Discounted sum of prevailing charge for played arms is an upper bound (in expectation) on the discounted sum of rewards.
2. Since the prevailing charge is nonincreasing, the discounted sum of prevailing charge is maximized if we always pick the arm with maximum prevailing charge.

Bayesian multi-armed bandit, Gittins index

Weber [1992] gives an exquisite proof of Gittins theorem. Let

$$\lambda_t(i) := \sup \left\{ \lambda \in \mathbb{R} : \sup_{\tau} \mathbb{E} \sum_{t<\tau} \gamma^t (X_t - \lambda) \geq 0 \right\}$$

the Gittins index of arm i at time t, which we interpret as the *maximum charge* one is willing to pay to play arm i given the current information. The *prevailing charge* is defined as $\min_{s \leq t} \lambda_s(i)$ (i.e. whenever the prevailing charge is too high we just drop it to the fair level).

1. Discounted sum of prevailing charge for played arms is an upper bound (in expectation) on the discounted sum of rewards.
2. Since the prevailing charge is nonincreasing, the discounted sum of prevailing charge is maximized if we always pick the arm with maximum prevailing charge.
3. Gittins index does exactly 2. and that in this case 1. is an equality. Q.E.D.

Bayesian multi-armed bandit, Thompson Sampling (TS)

In machine learning we want (i) strategies that can deal with complicated priors, and (ii) guarantees for misspecified priors. This is why we have to go beyond the Gittins index theory.

Bayesian multi-armed bandit, Thompson Sampling (TS)

In machine learning we want (i) strategies that can deal with complicated priors, and (ii) guarantees for misspecified priors. This is why we have to go beyond the Gittins index theory.

In his 1933 paper Thompson proposed the following strategy: sample $\theta' \sim \pi_t$ and play $I_t \in \mathrm{argmax}\, \mu_i(\theta')$.

Bayesian multi-armed bandit, Thompson Sampling (TS)

In machine learning we want (i) strategies that can deal with complicated priors, and (ii) guarantees for misspecified priors. This is why we have to go beyond the Gittins index theory.

In his 1933 paper Thompson proposed the following strategy: sample $\theta' \sim \pi_t$ and play $I_t \in \mathrm{argmax}\, \mu_i(\theta')$.

Theoretical guarantees for this highly practical strategy have long remained elusive. Recently Agrawal and Goyal [2012] and Kaufmann, Korda and Munos [2012] proved that TS with Bernoulli reward distributions and uniform prior on the parameters achieves $\overline{R}_T = O\left(\sum_i \frac{\log(T)}{\Delta_i}\right)$ (note that this is the frequentist regret!).

Bayesian multi-armed bandit, Thompson Sampling (TS)

In machine learning we want (i) strategies that can deal with complicated priors, and (ii) guarantees for misspecified priors. This is why we have to go beyond the Gittins index theory.

In his 1933 paper Thompson proposed the following strategy: sample $\theta' \sim \pi_t$ and play $I_t \in \mathrm{argmax}\, \mu_i(\theta')$.

Theoretical guarantees for this highly practical strategy have long remained elusive. Recently Agrawal and Goyal [2012] and Kaufmann, Korda and Munos [2012] proved that TS with Bernoulli reward distributions and uniform prior on the parameters achieves $\overline{R}_T = O\left(\sum_i \frac{\log(T)}{\Delta_i}\right)$ (note that this is the frequentist regret!).

Guha and Munagala [2014] conjecture that, for product priors, TS is a 2-approximation to the optimal Bayesian strategy for the objective of minimizing the number of pulls on suboptimal arms.

Bayesian multi-armed bandit, Russo and Van Roy [2014] information ratio analysis

Assume a prior in the adversarial model, that is a prior over $(\ell_1, \dots, \ell_T) \in [0,1]^{n \times T}$, and let $\mathbb{E}_t$ denote the posterior distribution (given $\ell_1(I_1), \dots, \ell_{t-1}(I_{t-1})$).

Bayesian multi-armed bandit, Russo and Van Roy [2014] information ratio analysis

Assume a prior in the adversarial model, that is a prior over $(\ell_1, \dots, \ell_T) \in [0,1]^{n \times T}$, and let $\mathbb{E}_t$ denote the posterior distribution (given $\ell_1(I_1), \dots, \ell_{t-1}(I_{t-1})$). We introduce

$$r_t(i) = \mathbb{E}_t(\ell_t(i) - \ell_t(i^*)), \text{ and } v_t(i) = \mathrm{Var}_t(\mathbb{E}_t(\ell_t(i)|i^*)).$$

Bayesian multi-armed bandit, Russo and Van Roy [2014] information ratio analysis

Assume a prior in the adversarial model, that is a prior over $(\ell_1, \dots, \ell_T) \in [0,1]^{n \times T}$, and let $\mathbb{E}_t$ denote the posterior distribution (given $\ell_1(I_1), \dots, \ell_{t-1}(I_{t-1})$). We introduce

$$r_t(i) = \mathbb{E}_t(\ell_t(i) - \ell_t(i^*)), \text{ and } v_t(i) = \mathrm{Var}_t(\mathbb{E}_t(\ell_t(i)|i^*)).$$

Key observation (next slide):

$$\mathbb{E} \sum_{t \le T} v_t(I_t) \le \frac{1}{2} H(i^*)$$

Bayesian multi-armed bandit, Russo and Van Roy [2014] information ratio analysis

Assume a prior in the adversarial model, that is a prior over $(\ell_1, \dots, \ell_T) \in [0,1]^{n \times T}$, and let $\mathbb{E}_t$ denote the posterior distribution (given $\ell_1(I_1), \dots, \ell_{t-1}(I_{t-1})$). We introduce

$$r_t(i) = \mathbb{E}_t(\ell_t(i) - \ell_t(i^*)), \text{ and } v_t(i) = \mathrm{Var}_t(\mathbb{E}_t(\ell_t(i)|i^*)).$$

Key observation (next slide):

$$\mathbb{E} \sum_{t \le T} v_t(I_t) \le \frac{1}{2} H(i^*)$$

which implies:

$$\begin{aligned} & \forall t, \mathbb{E}_t r_t(I_t) \le \sqrt{C \ \mathbb{E}_t v_t(I_t)} \\ \Rightarrow \quad & \mathbb{E} \sum_{t=1}^{T} r_t(I_t) \le \sum_{t=1}^{T} \sqrt{C \ \mathbb{E} v_t(I_t)} \\ \Rightarrow \quad & BR_T \le \sqrt{C \ T \ H(i^*)/2}. \end{aligned}$$

Bayesian multi-armed bandit, accumulation of information

$$v_t(i) = \mathrm{Var}_t(\mathbb{E}_t(\ell_t(i)|i^*)),\ \pi_t(j) = \mathbb{P}_t(i^* = j),\ \mathbb{E}\sum_{t \le T} v_t(I_t) \le \frac{1}{2}H(x^*)$$

Bayesian multi-armed bandit, accumulation of information

$$v_t(i) = \mathrm{Var}_t(\mathbb{E}_t(\ell_t(i)|i^*)),\ \pi_t(j) = \mathbb{P}_t(i^* = j),\ \mathbb{E} \sum_{t \leq T} v_t(I_t) \leq \frac{1}{2} H(x^*)$$

Equipped with Pinsker's inequality and basic information theory concepts (such as the mutual information $\mathbb{I}$) one has:

$$\begin{aligned} v_t(i) &= \sum_j \pi_t(j)(\mathbb{E}_t(\ell_t(i)|i^* = j) - \mathbb{E}_t(\ell_t(i)))^2 \\ &\leq \frac{1}{2} \sum_j \pi_t(j) \mathrm{Ent}(\mathcal{L}_t(\ell_t(i)|i^* = j) \| \mathcal{L}_t(\ell_t(i))) \\ &= \frac{1}{2} \mathbb{I}_t(\ell_t(i), i^*) = H_t(i^*) - H_t(i^*|\ell_t(i)). \end{aligned}$$

Bayesian multi-armed bandit, accumulation of information

$$v_t(i) = \mathrm{Var}_t(\mathbb{E}_t(\ell_t(i)|i^*)),\ \pi_t(j) = \mathbb{P}_t(i^* = j),\ \mathbb{E} \sum_{t \leq T} v_t(I_t) \leq \frac{1}{2} H(x^*)$$

Equipped with Pinsker's inequality and basic information theory concepts (such as the mutual information $\mathbb{I}$) one has:

$$\begin{aligned} v_t(i) &= \sum_j \pi_t(j)(\mathbb{E}_t(\ell_t(i)|i^* = j) - \mathbb{E}_t(\ell_t(i)))^2 \\ &\leq \frac{1}{2} \sum_j \pi_t(j)\mathrm{Ent}(\mathcal{L}_t(\ell_t(i)|i^* = j) \| \mathcal{L}_t(\ell_t(i))) \\ &= \frac{1}{2} \mathbb{I}_t(\ell_t(i), i^*) = H_t(i^*) - H_t(i^*|\ell_t(i)). \end{aligned}$$

Thus $\mathbb{E} v_t(I_t) \leq \frac{1}{2}\mathbb{E}(H_t(i^*) - H_{t+1}(i^*))$.

Bayesian multi-armed bandit, TS' information ratio

Let $\bar{\ell}_t(i) = \mathbb{E}_t \ell_t(i)$ and $\bar{\ell}_t(i,j) = \mathbb{E}_t(\ell_t(i) | i^* = j)$. Then

$$\mathbb{E}_t r_t(I_t) \leq \sqrt{C \ \mathbb{E}_t v_t(I_t)}$$

$$\Leftrightarrow \mathbb{E}_t \bar{\ell}_t(I_t) - \sum_i \pi_t(i) \bar{\ell}_t(i,i) \leq \sqrt{C \ \mathbb{E}_t \sum_j \pi_t(j) (\bar{\ell}_t(I_t, j) - \bar{\ell}_t(I_t))^2}$$

Bayesian multi-armed bandit, TS' information ratio

Let $\bar{\ell}_t(i) = \mathbb{E}_t \ell_t(i)$ and $\bar{\ell}_t(i,j) = \mathbb{E}_t(\ell_t(i) | i^* = j)$. Then

$$\mathbb{E}_t r_t(I_t) \leq \sqrt{C \ \mathbb{E}_t v_t(I_t)}$$

$$\Leftrightarrow \mathbb{E}_t \bar{\ell}_t(I_t) - \sum_i \pi_t(i) \bar{\ell}_t(i,i) \leq \sqrt{C \ \mathbb{E}_t \sum_j \pi_t(j) (\bar{\ell}_t(I_t, j) - \bar{\ell}_t(I_t))^2}$$

For TS the following shows that one can take $C = n$:

$$\begin{aligned} \mathbb{E}_t \bar{\ell}_t(I_t) - \sum_i \pi_t(i) \bar{\ell}_t(i,i) &= \sum_i \pi_t(i) (\bar{\ell}_t(i) - \bar{\ell}_t(i,i)) \\ &\leq \sqrt{n \sum_i \pi_t(i)^2 (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2} \\ &\leq \sqrt{n \sum_{i,j} \pi_t(i) \pi_t(j) (\bar{\ell}_t(i) - \bar{\ell}_t(i,j))^2}. \end{aligned}$$

Thus TS always satisfies $BR_T \leq \sqrt{TnH(i^*)} \leq \sqrt{Tn \log(n)}$.

Bayesian multi-armed bandit, TS' information ratio

Let $\bar{\ell}_t(i) = \mathbb{E}_t \ell_t(i)$ and $\bar{\ell}_t(i,j) = \mathbb{E}_t(\ell_t(i) | i^* = j)$. Then

$$\mathbb{E}_t r_t(I_t) \leq \sqrt{C \ \mathbb{E}_t v_t(I_t)}$$

$$\Leftrightarrow \mathbb{E}_t \bar{\ell}_t(I_t) - \sum_i \pi_t(i) \bar{\ell}_t(i,i) \leq \sqrt{C \ \mathbb{E}_t \sum_j \pi_t(j) (\bar{\ell}_t(I_t, j) - \bar{\ell}_t(I_t))^2}$$

For TS the following shows that one can take $C = n$:

$$\begin{aligned}
\mathbb{E}_t \bar{\ell}_t(I_t) - \sum_i \pi_t(i) \bar{\ell}_t(i,i) &= \sum_i \pi_t(i) (\bar{\ell}_t(i) - \bar{\ell}_t(i,i)) \\
&\leq \sqrt{n \sum_i \pi_t(i)^2 (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2} \\
&\leq \sqrt{n \sum_{i,j} \pi_t(i) \pi_t(j) (\bar{\ell}_t(i) - \bar{\ell}_t(i,j))^2}.
\end{aligned}$$

Thus TS always satisfies $BR_T \leq \sqrt{TnH(i^*)} \leq \sqrt{Tn \log(n)}$. Side note: by the minimax theorem this implies there exists a strategy for the oblivious adversarial model with regret $\sqrt{Tn \log(n)}$.

Summary of basic results

1. In the i.i.d. model UCB attains a regret of $O\left(\sum_i \frac{\log(T)}{\Delta_i}\right)$ and by Lai and Robbins' lower bound this is optimal (up to a multiplicative variance term).
2. In the adversarial model Exp3 attains a regret of $O(\sqrt{Tn\log(n)})$ and this is optimal up to the logarithmic term.
3. In the Bayesian model, Gittins index gives an *optimal* strategy for the case of product priors. For general priors Thompson Sampling is a more flexible strategy. Its Bayesian regret is controlled by the entropy of the optimal decision. Moreover TS with an uninformative prior has frequentist guarantees comparable to UCB.

Part 2: Linear, non-linear, and contextual bandit

The linear bandit problem, Auer [2002]

Known parameters: compact action set $\mathcal{A} \subset \mathbb{R}^n$, adversary's action set $\mathcal{L} \subset \mathbb{R}^n$, number of rounds T.

The linear bandit problem, Auer [2002]

Known parameters: compact action set $\mathcal{A} \subset \mathbb{R}^n$, adversary's action set $\mathcal{L} \subset \mathbb{R}^n$, number of rounds T.

Protocol: For each round $t = 1, 2, \ldots, T$, the adversary chooses a loss vector $\ell_t \in \mathcal{L}$ and simultaneously the player chooses $a_t \in \mathcal{A}$ based on past observations and receives a loss/observation $Y_t = \ell_t^\top a_t$.

$$R_T = \mathbb{E} \sum_{t=1}^{T} \ell_t^\top a_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{T} \ell_t^\top a.$$

The linear bandit problem, Auer [2002]

Known parameters: compact action set $\mathcal{A} \subset \mathbb{R}^n$, adversary's action set $\mathcal{L} \subset \mathbb{R}^n$, number of rounds T.

Protocol: For each round $t = 1, 2, \dots, T$, the adversary chooses a loss vector $\ell_t \in \mathcal{L}$ and simultaneously the player chooses $a_t \in \mathcal{A}$ based on past observations and receives a loss/observation $Y_t = \ell_t^\top a_t$.

$$R_T = \mathbb{E} \sum_{t=1}^{T} \ell_t^\top a_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{T} \ell_t^\top a.$$

Other models: In the i.i.d. model we assume that there is some underlying $\theta \in \mathcal{L}$ such that $\mathbb{E}(Y_t | a_t) = \theta^\top a_t$. In the Bayesian model we assume that we have a prior distribution ν over the sequence $(\ell_1, \dots, \ell_T)$ (in this case the expectation in R_T is also over $(\ell_1, \dots, \ell_T) \sim \nu$). Alternatively we could assume a prior over θ.

The linear bandit problem, Auer [2002]

Known parameters: compact action set $\mathcal{A} \subset \mathbb{R}^n$, adversary's action set $\mathcal{L} \subset \mathbb{R}^n$, number of rounds T.

Protocol: For each round $t = 1, 2, \dots, T$, the adversary chooses a loss vector $\ell_t \in \mathcal{L}$ and simultaneously the player chooses $a_t \in \mathcal{A}$ based on past observations and receives a loss/observation $Y_t = \ell_t^\top a_t$.

$$R_T = \mathbb{E} \sum_{t=1}^{T} \ell_t^\top a_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{T} \ell_t^\top a.$$

Other models: In the i.i.d. model we assume that there is some underlying $\theta \in \mathcal{L}$ such that $\mathbb{E}(Y_t | a_t) = \theta^\top a_t$. In the Bayesian model we assume that we have a prior distribution ν over the sequence $(\ell_1, \dots, \ell_T)$ (in this case the expectation in R_T is also over $(\ell_1, \dots, \ell_T) \sim \nu$). Alternatively we could assume a prior over θ.

Example: Part 1 was about $\mathcal{A} = \{e_1, \dots, e_n\}$ and $\mathcal{L} = [0, 1]^n$.

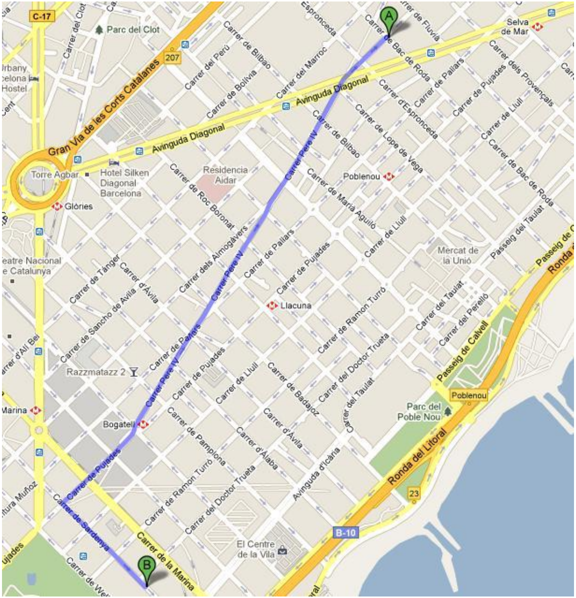
The linear bandit problem, Auer [2002]

Known parameters: compact action set $\mathcal{A} \subset \mathbb{R}^n$, adversary's action set $\mathcal{L} \subset \mathbb{R}^n$, number of rounds T.
Protocol: For each round $t = 1, 2, \dots, T$, the adversary chooses a loss vector $\ell_t \in \mathcal{L}$ and simultaneously the player chooses $a_t \in \mathcal{A}$ based on past observations and receives a loss/observation $Y_t = \ell_t^\top a_t$.

$$R_T = \mathbb{E} \sum_{t=1}^{T} \ell_t^\top a_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{T} \ell_t^\top a.$$

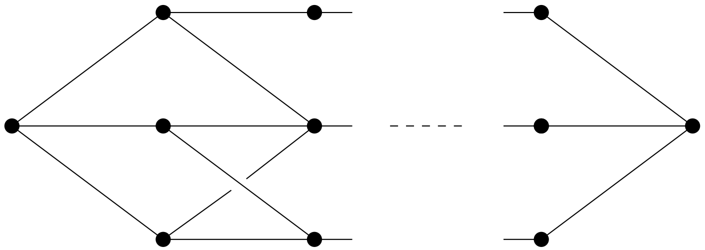
Other models: In the i.i.d. model we assume that there is some underlying $\theta \in \mathcal{L}$ such that $\mathbb{E}(Y_t | a_t) = \theta^\top a_t$. In the Bayesian model we assume that we have a prior distribution ν over the sequence $(\ell_1, \dots, \ell_T)$ (in this case the expectation in R_T is also over $(\ell_1, \dots, \ell_T) \sim \nu$). Alternatively we could assume a prior over θ.
Example: Part 1 was about $\mathcal{A} = \{e_1, \dots, e_n\}$ and $\mathcal{L} = [0,1]^n$.
Assumption: unless specified otherwise we assume $\mathcal{L} = \mathcal{A}^\circ := \{\ell : \sup_{a \in \mathcal{A}} |\ell^\top a| \le 1\}$.

Example: path planning



Example: path planning

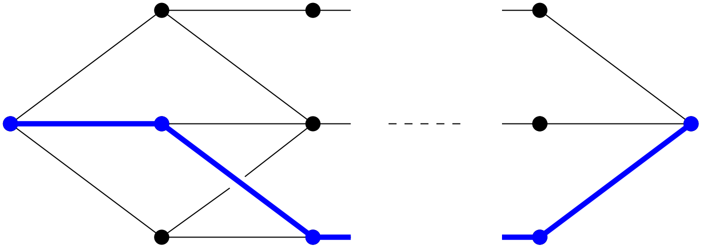
Adversary



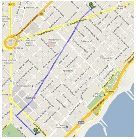
Player

Example: path planning

Adversary

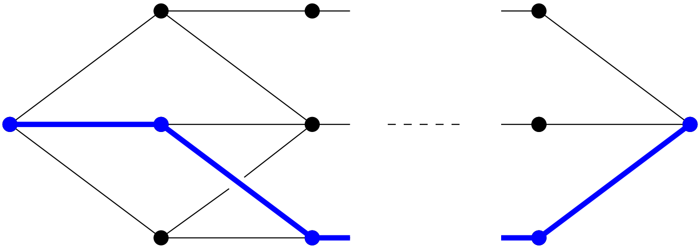


Player

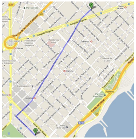


Example: path planning

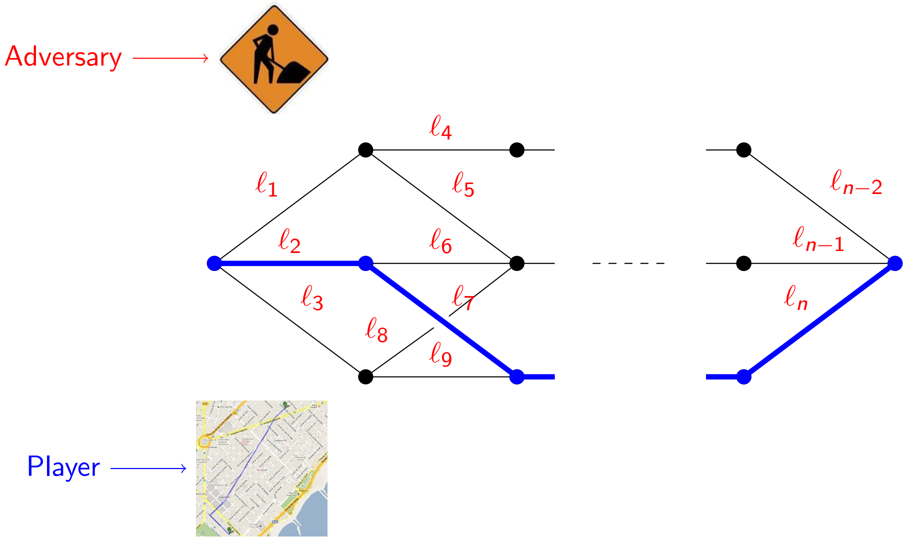
Adversary $\longrightarrow$



Player $\longrightarrow$

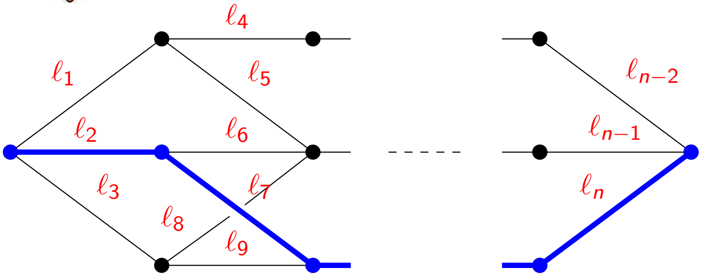


Example: path planning

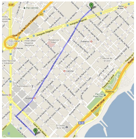


Example: path planning

Adversary $\longrightarrow$



Player $\longrightarrow$

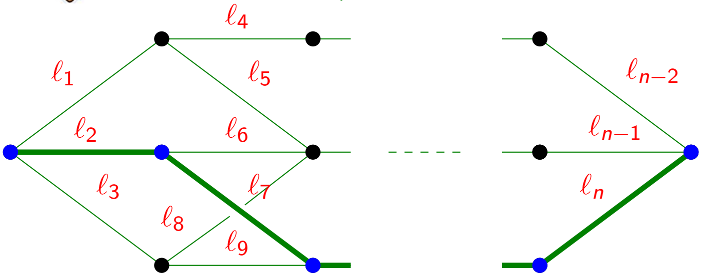


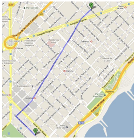
loss suffered: $\ell_2 + \ell_7 + \ldots + \ell_n$

Example: path planning

Adversary ⟶ 

Feedback: Full Info: $\ell_1, \ell_2, \ldots, \ell_n$

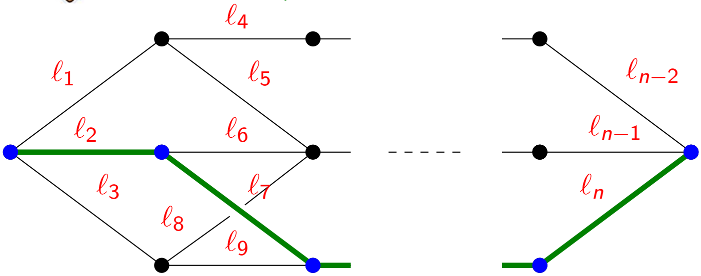


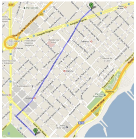
Player ⟶ 

loss suffered: $\ell_2 + \ell_7 + \ldots + \ell_n$

Example: path planning

Adversary ⟶  Feedback:
$$\begin{cases} \text{Full Info:} & \ell_1, \ell_2, \ldots, \ell_n \\ \text{Semi-Bandit:} & \ell_2, \ell_7, \ldots, \ell_n \end{cases}$$

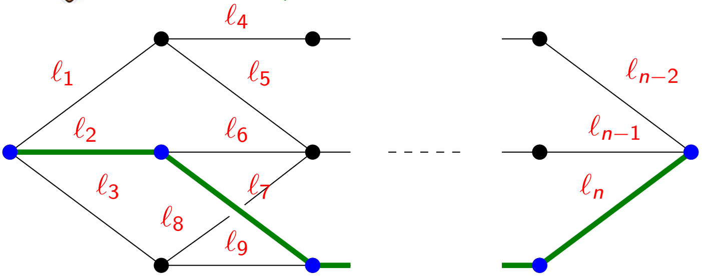


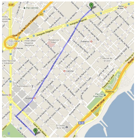
Player ⟶ 

loss suffered: $\ell_2 + \ell_7 + \ldots + \ell_n$

Example: path planning

Adversary $\longrightarrow$  Feedback: $\begin{cases} \text{Full Info:} & \ell_1, \ell_2, \ldots, \ell_n \\ \text{Semi-Bandit:} & \ell_2, \ell_7, \ldots, \ell_n \\ \text{Bandit:} & \ell_2 + \ell_7 + \ldots + \ell_n \end{cases}$



Player $\longrightarrow$  loss suffered: $\ell_2 + \ell_7 + \ldots + \ell_n$

Thompson Sampling for linear bandit after RVR14

Assume $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$. Recall from Part 1 that TS satisfies

$$\sum_i \pi_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)) \leq \sqrt{C \sum_{i,j} \pi_t(i)\pi_t(j)(\bar{\ell}_t(i,j) - \bar{\ell}_t(i))^2}$$

$$\Rightarrow R_T \leq \sqrt{C \;\; T \;\; \log(|\mathcal{A}|)/2},$$

where $\bar{\ell}_t(i) = \mathbb{E}_t \ell_t(i)$ and $\bar{\ell}_t(i,j) = \mathbb{E}_t(\ell_t(i) | i^* = j)$.

Thompson Sampling for linear bandit after RVR14

Assume $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$. Recall from Part 1 that TS satisfies

$$\sum_i \pi_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)) \leq \sqrt{C \sum_{i,j} \pi_t(i)\pi_t(j)(\bar{\ell}_t(i,j) - \bar{\ell}_t(i))^2}$$

$$\Rightarrow R_T \leq \sqrt{C \;\; T \;\; \log(|\mathcal{A}|)/2},$$

where $\bar{\ell}_t(i) = \mathbb{E}_t \ell_t(i)$ and $\bar{\ell}_t(i,j) = \mathbb{E}_t(\ell_t(i) | i^* = j)$.

Writing $\bar{\ell}_t(i) = a_i^\top \bar{\ell}_t$, $\bar{\ell}_t(i,j) = a_i^\top \bar{\ell}_t^j$, and $(M_{i,j}) = \left(\sqrt{\pi_t(i)\pi_t(j)} a_i^\top (\bar{\ell}_t - \bar{\ell}_t^j)\right)$ we want to show that

$$\mathrm{Tr}(M) \leq \sqrt{C} \|M\|_F.$$

Thompson Sampling for linear bandit after RVR14

Assume $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$. Recall from Part 1 that TS satisfies

$$\sum_i \pi_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)) \leq \sqrt{C \sum_{i,j} \pi_t(i)\pi_t(j)(\bar{\ell}_t(i,j) - \bar{\ell}_t(i))^2}$$
$$\Rightarrow R_T \leq \sqrt{C \ T \ \log(|\mathcal{A}|)/2},$$

where $\bar{\ell}_t(i) = \mathbb{E}_t \ell_t(i)$ and $\bar{\ell}_t(i,j) = \mathbb{E}_t(\ell_t(i) | i^* = j)$.

Writing $\bar{\ell}_t(i) = a_i^\top \bar{\ell}_t$, $\bar{\ell}_t(i,j) = a_i^\top \bar{\ell}_t^j$, and $(M_{i,j}) = \left(\sqrt{\pi_t(i)\pi_t(j)} a_i^\top (\bar{\ell}_t - \bar{\ell}_t^j)\right)$ we want to show that

$$\mathrm{Tr}(M) \leq \sqrt{C} \|M\|_F.$$

Using the eigenvalue formula for the trace and the Frobenius norm one can see that $\mathrm{Tr}(M)^2 \leq \mathrm{rank}(M) \|M\|_F^2$.

Thompson Sampling for linear bandit after RVR14

Assume $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$. Recall from Part 1 that TS satisfies

$$\sum_i \pi_t(i)(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)) \le \sqrt{C \sum_{i,j} \pi_t(i)\pi_t(j)(\bar{\ell}_t(i,j) - \bar{\ell}_t(i))^2}$$

$$\Rightarrow R_T \le \sqrt{C \;\, T \;\, \log(|\mathcal{A}|)/2},$$

where $\bar{\ell}_t(i) = \mathbb{E}_t \ell_t(i)$ and $\bar{\ell}_t(i,j) = \mathbb{E}_t(\ell_t(i) | i^* = j)$.

Writing $\bar{\ell}_t(i) = a_i^\top \bar{\ell}_t$, $\bar{\ell}_t(i,j) = a_i^\top \bar{\ell}^j_t$, and $(M_{i,j}) = \left(\sqrt{\pi_t(i)\pi_t(j)} a_i^\top (\bar{\ell}_t - \bar{\ell}^j_t)\right)$ we want to show that

$$\mathrm{Tr}(M) \le \sqrt{C} \|M\|_F.$$

Using the eigenvalue formula for the trace and the Frobenius norm one can see that $\mathrm{Tr}(M)^2 \le \mathrm{rank}(M)\|M\|_F^2$. Moreover the rank of M is at most n since $M = UV^\top$ where $U, V \in \mathbb{R}^{|\mathcal{A}| \times n}$ (the i^{th} row of U is $\sqrt{\pi_t(i)} a_i$ and for V it is $\sqrt{\pi_t(i)}(\bar{\ell}_t - \bar{\ell}^i_t)$).

Thompson Sampling for linear bandit after RVR14

1. TS satisfies $R_T \leq \sqrt{nT \log(|\mathcal{A}|)}$. To appreciate the improvement recall that without the linear structure one would get a regret of order $\sqrt{|\mathcal{A}| T}$ and that $\mathcal{A}$ can be exponential in the dimension n (think of the path planning example).

Thompson Sampling for linear bandit after RVR14

1. TS satisfies $R_T \leq \sqrt{nT\log(|\mathcal{A}|)}$. To appreciate the improvement recall that without the linear structure one would get a regret of order $\sqrt{|\mathcal{A}|T}$ and that $\mathcal{A}$ can be exponential in the dimension n (think of the path planning example).
2. Provided that one can efficiently sample from the posterior on ℓ_t (or on θ), TS just requires at each step one linear optimization over $\mathcal{A}$.

Thompson Sampling for linear bandit after RVR14

1. TS satisfies $R_T \leq \sqrt{nT \log(|\mathcal{A}|)}$. To appreciate the improvement recall that without the linear structure one would get a regret of order $\sqrt{|\mathcal{A}|T}$ and that $\mathcal{A}$ can be exponential in the dimension n (think of the path planning example).
2. Provided that one can efficiently sample from the posterior on ℓ_t (or on θ), TS just requires at each step one linear optimization over $\mathcal{A}$.
3. TS regret bound is optimal in the following sense. W.l.og. one can assume $|\mathcal{A}| \leq (10T)^n$ and thus TS satisfies $R_T = O(n\sqrt{T \log(T)})$ for any action set. Furthermore one can show that there exists an action set and a prior such that for any strategy one has $R_T = \Omega(n\sqrt{T})$, see Dani, Hayes and Kakade [2008], Rusmevichientong and Tsitsiklis [2010], and Audibert, Bubeck and Lugosi [2011, 2014].

Adversarial linear bandit after Dani, Hayes, Kakade [2008]

Recall from Part 1 that exponential weights satisfies for any $\widetilde{\ell}_t$ such that $\mathbb{E}\widetilde{\ell}_t(i) = \ell_t(i)$ and $\widetilde{\ell}_t(i) \geq 0$,

$$R_T \leq \frac{\max_i \mathrm{Ent}(\delta_i \| p_1)}{\eta} + \frac{\eta}{2}\mathbb{E}\sum_t \mathbb{E}_{I\sim p_t}\widetilde{\ell}_t(I)^2.$$

Adversarial linear bandit after Dani, Hayes, Kakade [2008]

Recall from Part 1 that exponential weights satisfies for any $\widetilde{\ell}_t$ such that $\mathbb{E}\widetilde{\ell}_t(i) = \ell_t(i)$ and $\widetilde{\ell}_t(i) \geq 0$,

$$R_T \leq \frac{\max_i \mathrm{Ent}(\delta_i \| p_1)}{\eta} + \frac{\eta}{2}\mathbb{E}\sum_t \mathbb{E}_{I \sim p_t}\widetilde{\ell}_t(I)^2.$$

DHK08 proposed the following (beautiful) unbiased estimator for the linear case:

$$\widetilde{\ell}_t = \Sigma_t^{-1} a_t a_t^\top \ell_t \text{ where } \Sigma_t = \mathbb{E}_{a \sim p_t}(a a^\top).$$

Adversarial linear bandit after Dani, Hayes, Kakade [2008]

Recall from Part 1 that exponential weights satisfies for any $\widetilde{\ell}_t$ such that $\mathbb{E}\widetilde{\ell}_t(i) = \ell_t(i)$ and $\widetilde{\ell}_t(i) \geq 0$,

$$R_T \leq \frac{\max_i \mathrm{Ent}(\delta_i \| p_1)}{\eta} + \frac{\eta}{2}\mathbb{E}\sum_t \mathbb{E}_{I\sim p_t}\widetilde{\ell}_t(I)^2.$$

DHK08 proposed the following (beautiful) unbiased estimator for the linear case:

$$\widetilde{\ell}_t = \Sigma_t^{-1} a_t a_t^\top \ell_t \text{ where } \Sigma_t = \mathbb{E}_{a\sim p_t}(aa^\top).$$

Again, amazingly, the variance is automatically controlled:

$$\mathbb{E}(\mathbb{E}_{a\sim p_t}(\widetilde{\ell}_t^\top a)^2) = \mathbb{E}\widetilde{\ell}_t^\top \Sigma_t \widetilde{\ell}_t \leq \mathbb{E}a_t^\top \Sigma_t^{-1} a_t = \mathbb{E}\mathrm{Tr}(\Sigma_t^{-1} a_t a_t) = n.$$

Adversarial linear bandit after Dani, Hayes, Kakade [2008]

Recall from Part 1 that exponential weights satisfies for any $\widetilde{\ell}_t$ such that $\mathbb{E}\widetilde{\ell}_t(i) = \ell_t(i)$ and $\widetilde{\ell}_t(i) \geq 0$,

$$R_T \leq \frac{\max_i \mathrm{Ent}(\delta_i \| p_1)}{\eta} + \frac{\eta}{2}\mathbb{E}\sum_t \mathbb{E}_{I \sim p_t}\widetilde{\ell}_t(I)^2.$$

DHK08 proposed the following (beautiful) unbiased estimator for the linear case:

$$\widetilde{\ell}_t = \Sigma_t^{-1} a_t a_t^\top \ell_t \text{ where } \Sigma_t = \mathbb{E}_{a \sim p_t}(a a^\top).$$

Again, amazingly, the variance is automatically controlled:

$$\mathbb{E}(\mathbb{E}_{a \sim p_t}(\widetilde{\ell}_t^\top a)^2) = \mathbb{E}\widetilde{\ell}_t^\top \Sigma_t \widetilde{\ell}_t \leq \mathbb{E} a_t^\top \Sigma_t^{-1} a_t = \mathbb{E}\mathrm{Tr}(\Sigma_t^{-1} a_t a_t) = n.$$

Up to the issue that $\widetilde{\ell}_t$ can take negative values this suggests the "optimal" $\sqrt{nT \log(|\mathcal{A}|)}$ regret bound.

Adversarial linear bandit, further development

1. The non-negativity issue of $\widetilde{\ell}_t$ is a manifestation of the need for an added exploration. DHK08 used a suboptimal exploration which led to an additional $\sqrt{n}$ in the regret. This was later improved in Bubeck, Cesa-Bianchi, and Kakade [2012] with an exploration based on the John's ellipsoid (smallest ellipsoid containing $\mathcal{A}$).

Adversarial linear bandit, further development

1. The non-negativity issue of $\widetilde{\ell}_t$ is a manifestation of the need for an added exploration. DHK08 used a suboptimal exploration which led to an additional $\sqrt{n}$ in the regret. This was later improved in Bubeck, Cesa-Bianchi, and Kakade [2012] with an exploration based on the John's ellipsoid (smallest ellipsoid containing $\mathcal{A}$).
2. Sampling the exp. weights is usually computationally difficult, see Cesa-Bianchi and Lugosi [2009] for some exceptions.

Adversarial linear bandit, further development

1. The non-negativity issue of $\widetilde{\ell}_t$ is a manifestation of the need for an added exploration. DHK08 used a suboptimal exploration which led to an additional $\sqrt{n}$ in the regret. This was later improved in Bubeck, Cesa-Bianchi, and Kakade [2012] with an exploration based on the John's ellipsoid (smallest ellipsoid containing $\mathcal{A}$).
2. Sampling the exp. weights is usually computationally difficult, see Cesa-Bianchi and Lugosi [2009] for some exceptions.
3. Abernethy, Hazan and Rakhlin [2008] proposed an alternative (beautiful) strategy based on mirror descent. The key idea is to use a n-self-concordant barrier for $\mathrm{conv}(\mathcal{A})$ as a mirror map and to sample points uniformly in Dikin ellipses. This method's regret is suboptimal by a factor $\sqrt{n}$ and the computational efficiency depends on the barrier being used.

Adversarial linear bandit, further development

1. The non-negativity issue of $\widetilde{\ell}_t$ is a manifestation of the need for an added exploration. DHK08 used a suboptimal exploration which led to an additional $\sqrt{n}$ in the regret. This was later improved in Bubeck, Cesa-Bianchi, and Kakade [2012] with an exploration based on the John's ellipsoid (smallest ellipsoid containing $\mathcal{A}$).
2. Sampling the exp. weights is usually computationally difficult, see Cesa-Bianchi and Lugosi [2009] for some exceptions.
3. Abernethy, Hazan and Rakhlin [2008] proposed an alternative (beautiful) strategy based on mirror descent. The key idea is to use a n-self-concordant barrier for $\mathrm{conv}(\mathcal{A})$ as a mirror map and to sample points uniformly in Dikin ellipses. This method's regret is suboptimal by a factor $\sqrt{n}$ and the computational efficiency depends on the barrier being used.
4. Bubeck and Eldan [2014]'s entropic barrier allows for a much more information-efficient sampling than AHR08. This gives another strategy with optimal regret which is efficient when $\mathcal{A}$ is convex (and one can do linear optimization on $\mathcal{A}$).

Adversarial combinatorial bandit after Audibert, Bubeck and Lugosi [2011, 2014]

Combinatorial setting: $\mathcal{A} \subset \{0,1\}^n$, $\max_a \|a\|_1 = m$, $\mathcal{L} = [0,1]^n$.

Adversarial combinatorial bandit after Audibert, Bubeck and Lugosi [2011, 2014]

Combinatorial setting: $\mathcal{A} \subset \{0,1\}^n$, $\max_a \|a\|_1 = m$, $\mathcal{L} = [0,1]^n$.

1. Full information case goes back to the end of the 90's (Warmuth and co-authors), semi-bandit and bandit were introduced in Audibert, Bubeck and Lugosi [2011] (following several papers that studied specific sets $\mathcal{A}$).

Adversarial combinatorial bandit after Audibert, Bubeck and Lugosi [2011, 2014]

Combinatorial setting: $\mathcal{A} \subset \{0,1\}^n$, $\max_a \|a\|_1 = m$, $\mathcal{L} = [0,1]^n$.

1. Full information case goes back to the end of the 90's (Warmuth and co-authors), semi-bandit and bandit were introduced in Audibert, Bubeck and Lugosi [2011] (following several papers that studied specific sets $\mathcal{A}$).
2. This is a natural setting to study FPL-type (Follow the Perturbed Leader) strategies, see e.g. Kalai and Vempala [2004] and more recently Devroye, Lugosi and Neu [2013].

Adversarial combinatorial bandit after Audibert, Bubeck and Lugosi [2011, 2014]

Combinatorial setting: $\mathcal{A} \subset \{0,1\}^n$, $\max_a \|a\|_1 = m$, $\mathcal{L} = [0,1]^n$.

1. Full information case goes back to the end of the 90's (Warmuth and co-authors), semi-bandit and bandit were introduced in Audibert, Bubeck and Lugosi [2011] (following several papers that studied specific sets $\mathcal{A}$).
2. This is a natural setting to study FPL-type (Follow the Perturbed Leader) strategies, see e.g. Kalai and Vempala [2004] and more recently Devroye, Lugosi and Neu [2013].
3. ABL11: Exponential weights is provably suboptimal in this setting! This is in sharp contrast with the case where $\mathcal{L} = \mathcal{A}^\circ$.

Adversarial combinatorial bandit after Audibert, Bubeck and Lugosi [2011, 2014]

Combinatorial setting: $\mathcal{A} \subset \{0,1\}^n$, $\max_a \|a\|_1 = m$, $\mathcal{L} = [0,1]^n$.

1. Full information case goes back to the end of the 90's (Warmuth and co-authors), semi-bandit and bandit were introduced in Audibert, Bubeck and Lugosi [2011] (following several papers that studied specific sets $\mathcal{A}$).
2. This is a natural setting to study FPL-type (Follow the Perturbed Leader) strategies, see e.g. Kalai and Vempala [2004] and more recently Devroye, Lugosi and Neu [2013].
3. ABL11: Exponential weights is provably suboptimal in this setting! This is in sharp contrast with the case where $\mathcal{L} = \mathcal{A}^{\circ}$.
4. Optimal regret in the semi-bandit case is $\sqrt{mnT}$ and it can be achieved with mirror descent and the natural unbiased estimator for the semi-bandit situation.

Adversarial combinatorial bandit after Audibert, Bubeck and Lugosi [2011, 2014]

Combinatorial setting: $\mathcal{A} \subset \{0,1\}^n$, $\max_a \|a\|_1 = m$, $\mathcal{L} = [0,1]^n$.

1. Full information case goes back to the end of the 90's (Warmuth and co-authors), semi-bandit and bandit were introduced in Audibert, Bubeck and Lugosi [2011] (following several papers that studied specific sets $\mathcal{A}$).
2. This is a natural setting to study FPL-type (Follow the Perturbed Leader) strategies, see e.g. Kalai and Vempala [2004] and more recently Devroye, Lugosi and Neu [2013].
3. ABL11: Exponential weights is provably suboptimal in this setting! This is in sharp contrast with the case where $\mathcal{L} = \mathcal{A}^\circ$.
4. Optimal regret in the semi-bandit case is $\sqrt{mnT}$ and it can be achieved with mirror descent and the natural unbiased estimator for the semi-bandit situation.
5. For the bandit case the bound for exponential weights from the previous slides gives $m\sqrt{mnT}$. However the lower bound from ABL14 is $m\sqrt{nT}$, which is conjectured to be tight.

Preliminaries for the i.i.d. case: a primer on least squares

Assume $Y_t = \theta^\top a_t + \xi_t$ where (ξ_t) is an i.i.d. sequence of centered and sub-Gaussian real-valued random variables. The (regularized) least squares estimator for θ based on $\mathbb{Y}_t = (Y_1, \ldots, Y_{t-1})^\top$ is, with $\mathbb{A}_t = (a_1 \ldots a_{t-1}) \in \mathbb{R}^{n \times t-1}$ and $\Sigma_t = \lambda \mathrm{I}_n + \sum_{s=1}^{t-1} a_s a_s^\top$:

$$\hat{\theta}_t = \Sigma_t^{-1} \mathbb{A}_t \mathbb{Y}_t$$

Preliminaries for the i.i.d. case: a primer on least squares

Assume $Y_t = \theta^\top a_t + \xi_t$ where (ξ_t) is an i.i.d. sequence of centered and sub-Gaussian real-valued random variables. The (regularized) least squares estimator for θ based on $\mathbb{Y}_t = (Y_1, \ldots, Y_{t-1})^\top$ is, with $\mathbb{A}_t = (a_1 \ldots a_{t-1}) \in \mathbb{R}^{n \times t-1}$ and $\Sigma_t = \lambda \mathrm{I}_n + \sum_{s=1}^{t-1} a_s a_s^\top$:

$$\hat{\theta}_t = \Sigma_t^{-1} \mathbb{A}_t \mathbb{Y}_t$$

Observe that we can also write $\theta = \Sigma_t^{-1}(\mathbb{A}_t(\mathbb{Y}_t + \varepsilon_t) + \lambda\theta)$ where $\varepsilon_t = (\mathbb{E}(Y_1|a_1) - Y_1, \ldots, \mathbb{E}(Y_{t-1}|a_{t-1}) - Y_{t-1})^\top$

Preliminaries for the i.i.d. case: a primer on least squares

Assume $Y_t = \theta^\top a_t + \xi_t$ where (ξ_t) is an i.i.d. sequence of centered and sub-Gaussian real-valued random variables. The (regularized) least squares estimator for θ based on $\mathbb{Y}_t = (Y_1, \ldots, Y_{t-1})^\top$ is, with $\mathbb{A}_t = (a_1 \ldots a_{t-1}) \in \mathbb{R}^{n \times t-1}$ and $\Sigma_t = \lambda \mathrm{I}_n + \sum_{s=1}^{t-1} a_s a_s^\top$:

$$\hat{\theta}_t = \Sigma_t^{-1} \mathbb{A}_t \mathbb{Y}_t$$

Observe that we can also write $\theta = \Sigma_t^{-1}(\mathbb{A}_t(\mathbb{Y}_t + \varepsilon_t) + \lambda\theta)$ where $\varepsilon_t = (\mathbb{E}(Y_1|a_1) - Y_1, \ldots, \mathbb{E}(Y_{t-1}|a_{t-1}) - Y_{t-1})^\top$ so that

$$\|\theta - \hat{\theta}_t\|_{\Sigma_t} = \|\mathbb{A}_t \varepsilon_t + \lambda\theta\|_{\Sigma_t^{-1}} \leq \|\mathbb{A}_t \varepsilon_t\|_{\Sigma_t^{-1}} + \sqrt{\lambda}\|\theta\|.$$

Preliminaries for the i.i.d. case: a primer on least squares

Assume $Y_t = \theta^\top a_t + \xi_t$ where (ξ_t) is an i.i.d. sequence of centered and sub-Gaussian real-valued random variables. The (regularized) least squares estimator for θ based on $\mathbb{Y}_t = (Y_1, \ldots, Y_{t-1})^\top$ is, with $\mathbb{A}_t = (a_1 \ldots a_{t-1}) \in \mathbb{R}^{n \times t-1}$ and $\Sigma_t = \lambda \mathrm{I}_n + \sum_{s=1}^{t-1} a_s a_s^\top$:

$$\hat{\theta}_t = \Sigma_t^{-1} \mathbb{A}_t \mathbb{Y}_t$$

Observe that we can also write $\theta = \Sigma_t^{-1}(\mathbb{A}_t(\mathbb{Y}_t + \varepsilon_t) + \lambda\theta)$ where $\varepsilon_t = (\mathbb{E}(Y_1|a_1) - Y_1, \ldots, \mathbb{E}(Y_{t-1}|a_{t-1}) - Y_{t-1})^\top$ so that

$$\|\theta - \hat{\theta}_t\|_{\Sigma_t} = \|\mathbb{A}_t \varepsilon_t + \lambda\theta\|_{\Sigma_t^{-1}} \leq \|\mathbb{A}_t \varepsilon_t\|_{\Sigma_t^{-1}} + \sqrt{\lambda}\|\theta\|.$$

A basic martingale argument (see e.g., Abbasi-Yadkori, Pál and Szepesvári [2011]) shows that w.p. $\geq 1 - \delta$, $\forall t \geq 1$,

$$\|\mathbb{A}_t \varepsilon_t\|_{\Sigma_t^{-1}} \leq \sqrt{\mathrm{logdet}(\Sigma_t) + \log(1/(\delta^2 \lambda^n))}.$$

Preliminaries for the i.i.d. case: a primer on least squares

Assume $Y_t = \theta^\top a_t + \xi_t$ where (ξ_t) is an i.i.d. sequence of centered and sub-Gaussian real-valued random variables. The (regularized) least squares estimator for θ based on $\mathbb{Y}_t = (Y_1, \dots, Y_{t-1})^\top$ is, with $\mathbb{A}_t = (a_1 \dots a_{t-1}) \in \mathbb{R}^{n \times t-1}$ and $\Sigma_t = \lambda \mathrm{I}_n + \sum_{s=1}^{t-1} a_s a_s^\top$:

$$\hat{\theta}_t = \Sigma_t^{-1} \mathbb{A}_t \mathbb{Y}_t$$

Observe that we can also write $\theta = \Sigma_t^{-1}(\mathbb{A}_t(\mathbb{Y}_t + \varepsilon_t) + \lambda\theta)$ where $\varepsilon_t = (\mathbb{E}(Y_1|a_1) - Y_1, \dots, \mathbb{E}(Y_{t-1}|a_{t-1}) - Y_{t-1})^\top$ so that

$$\|\theta - \hat{\theta}_t\|_{\Sigma_t} = \|\mathbb{A}_t \varepsilon_t + \lambda\theta\|_{\Sigma_t^{-1}} \le \|\mathbb{A}_t \varepsilon_t\|_{\Sigma_t^{-1}} + \sqrt{\lambda}\|\theta\|.$$

A basic martingale argument (see e.g., Abbasi-Yadkori, Pál and Szepesvári [2011]) shows that w.p. $\ge 1 - \delta$, $\forall t \ge 1$,

$$\|\mathbb{A}_t \varepsilon_t\|_{\Sigma_t^{-1}} \le \sqrt{\mathrm{logdet}(\Sigma_t) + \log(1/(\delta^2 \lambda^n))}.$$

Note that $\mathrm{logdet}(\Sigma_t) \le n \log(\mathrm{Tr}(\Sigma_t)/n) \le n \log(\lambda + t/n)$ (w.l.o.g. we assumed $\|a_t\| \le 1$).

i.i.d. linear bandit after DHK08, RT10, AYPS11

Let $\beta = 2\sqrt{n \log(T)}$, and $\mathcal{E}_t = \{\theta' : \|\theta' - \hat{\theta}_t\|_{\Sigma_t} \leq \beta\}$. We showed that w.p. $\geq 1 - 1/T^2$ one has $\theta \in \mathcal{E}_t$ for all $t \in [T]$.

i.i.d. linear bandit after DHK08, RT10, AYPS11

Let $\beta = 2\sqrt{n \log(T)}$, and $\mathcal{E}_t = \{\theta' : \|\theta' - \hat{\theta}_t\|_{\Sigma_t} \leq \beta\}$. We showed that w.p. $\geq 1 - 1/T^2$ one has $\theta \in \mathcal{E}_t$ for all $t \in [T]$.

The appropriate generalization of UCB is to select: $(\widetilde{\theta}_t, a_t) = \mathrm{argmin}_{(\theta', a) \in \mathcal{E}_t \times \mathcal{A}} \theta'^{\top} a$ (this optimization is NP-hard in general, more on that next slide).

i.i.d. linear bandit after DHK08, RT10, AYPS11

Let $\beta = 2\sqrt{n \log(T)}$, and $\mathcal{E}_t = \{\theta' : \|\theta' - \hat{\theta}_t\|_{\Sigma_t} \leq \beta\}$. We showed that w.p. $\geq 1 - 1/T^2$ one has $\theta \in \mathcal{E}_t$ for all $t \in [T]$.

The appropriate generalization of UCB is to select: $(\widetilde{\theta}_t, a_t) = \mathrm{argmin}_{(\theta', a) \in \mathcal{E}_t \times \mathcal{A}} \, \theta'^\top a$ (this optimization is NP-hard in general, more on that next slide). Then one has on the high-probability event:

$$\sum_{t=1}^{T} \theta^\top (a_t - a^*) \leq \sum_{t=1}^{T} (\theta - \widetilde{\theta}_t)^\top a_t \leq \beta \sum_{t=1}^{T} \|a_t\|_{\Sigma_t^{-1}} \leq \beta \sqrt{T \sum_t \|a_t\|^2_{\Sigma_t^{-1}}}.$$

i.i.d. linear bandit after DHK08, RT10, AYPS11

Let $\beta = 2\sqrt{n \log(T)}$, and $\mathcal{E}_t = \{\theta' : \|\theta' - \hat{\theta}_t\|_{\Sigma_t} \leq \beta\}$. We showed that w.p. $\geq 1 - 1/T^2$ one has $\theta \in \mathcal{E}_t$ for all $t \in [T]$.

The appropriate generalization of UCB is to select: $(\widetilde{\theta}_t, a_t) = \mathrm{argmin}_{(\theta', a) \in \mathcal{E}_t \times \mathcal{A}} \theta'^{\top} a$ (this optimization is NP-hard in general, more on that next slide). Then one has on the high-probability event:

$$\sum_{t=1}^{T} \theta^{\top}(a_t - a^*) \leq \sum_{t=1}^{T} (\theta - \widetilde{\theta}_t)^{\top} a_t \leq \beta \sum_{t=1}^{T} \|a_t\|_{\Sigma_t^{-1}} \leq \beta \sqrt{T \sum_t \|a_t\|^2_{\Sigma_t^{-1}}}.$$

To control the sum of squares we observe that:

$$\det(\Sigma_{t+1}) = \det(\Sigma_t) \det(\mathrm{I}_n + \Sigma_t^{-1/2} a_t (\Sigma_t^{-1/2} a_t)^{\top}) = \det(\Sigma_t)(1 + \|a_t\|^2_{\Sigma_t^{-1}})$$

so that (assuming $\lambda \geq 1$)

$$\log\det(\Sigma_{T+1}) - \log\det(\Sigma_1) = \sum_t \log(1 + \|a_t\|^2_{\Sigma_t^{-1}}) \geq \frac{1}{2} \sum_t \|a_t\|^2_{\Sigma_t^{-1}}.$$

i.i.d. linear bandit after DHK08, RT10, AYPS11

Let $\beta = 2\sqrt{n \log(T)}$, and $\mathcal{E}_t = \{\theta' : \|\theta' - \hat{\theta}_t\|_{\Sigma_t} \leq \beta\}$. We showed that w.p. $\geq 1 - 1/T^2$ one has $\theta \in \mathcal{E}_t$ for all $t \in [T]$.

The appropriate generalization of UCB is to select: $(\widetilde{\theta}_t, a_t) = \mathrm{argmin}_{(\theta', a) \in \mathcal{E}_t \times \mathcal{A}} \theta'^\top a$ (this optimization is NP-hard in general, more on that next slide). Then one has on the high-probability event:

$$\sum_{t=1}^{T} \theta^\top (a_t - a^*) \leq \sum_{t=1}^{T} (\theta - \widetilde{\theta}_t)^\top a_t \leq \beta \sum_{t=1}^{T} \|a_t\|_{\Sigma_t^{-1}} \leq \beta \sqrt{T \sum_t \|a_t\|^2_{\Sigma_t^{-1}}}.$$

To control the sum of squares we observe that:

$$\det(\Sigma_{t+1}) = \det(\Sigma_t) \det(\mathrm{I}_n + \Sigma_t^{-1/2} a_t (\Sigma_t^{-1/2} a_t)^\top) = \det(\Sigma_t)(1 + \|a_t\|^2_{\Sigma_t^{-1}})$$

so that (assuming $\lambda \geq 1$)

$$\log\det(\Sigma_{T+1}) - \log\det(\Sigma_1) = \sum_t \log(1 + \|a_t\|^2_{\Sigma_t^{-1}}) \geq \frac{1}{2} \sum_t \|a_t\|^2_{\Sigma_t^{-1}}.$$

Putting things together we see that the regret is $O(n \log(T) \sqrt{T})$.

What's the point of i.i.d. linear bandit?

So far we did not get any real benefit from the i.i.d. assumption (the regret guarantee we obtained is the same as for the adversarial model). To me the key benefit is in the simplicity of the i.i.d. algorithm which makes it easy to incorporate further assumptions.

What's the point of i.i.d. linear bandit?

So far we did not get any real benefit from the i.i.d. assumption (the regret guarantee we obtained is the same as for the adversarial model). To me the key benefit is in the simplicity of the i.i.d. algorithm which makes it easy to incorporate further assumptions.

1. Sparsity of θ: instead of regularization with ℓ_2-norm to define $\hat{\theta}$ one could regularize with ℓ_1-norm, see e.g., Johnson, Sivakumar and Banerjee [2016].

What's the point of i.i.d. linear bandit?

So far we did not get any real benefit from the i.i.d. assumption (the regret guarantee we obtained is the same as for the adversarial model). To me the key benefit is in the simplicity of the i.i.d. algorithm which makes it easy to incorporate further assumptions.

1. Sparsity of θ: instead of regularization with ℓ_2-norm to define $\hat{\theta}$ one could regularize with ℓ_1-norm, see e.g., Johnson, Sivakumar and Banerjee [2016].
2. Computational constraint: instead of optimizing over $\mathcal{E}_t$ to define $\widetilde{\theta}_t$ one could optimize over an ℓ_1-ball containing $\mathcal{E}_t$ (this would cost an extra $\sqrt{n}$ in the regret bound).

What's the point of i.i.d. linear bandit?

So far we did not get any real benefit from the i.i.d. assumption (the regret guarantee we obtained is the same as for the adversarial model). To me the key benefit is in the simplicity of the i.i.d. algorithm which makes it easy to incorporate further assumptions.

1. Sparsity of θ: instead of regularization with ℓ_2-norm to define $\hat{\theta}$ one could regularize with ℓ_1-norm, see e.g., Johnson, Sivakumar and Banerjee [2016].
2. Computational constraint: instead of optimizing over $\mathcal{E}_t$ to define $\widetilde{\theta}_t$ one could optimize over an ℓ_1-ball containing $\mathcal{E}_t$ (this would cost an extra $\sqrt{n}$ in the regret bound).
3. Generalized linear model: $\mathbb{E}(Y_t|a_t) = \sigma(\theta^\top a_t)$ for some known increasing $\sigma : \mathbb{R} \to \mathbb{R}$, see Filippi, Cappe, Garivier and Szepesvari [2011].

What's the point of i.i.d. linear bandit?

So far we did not get any real benefit from the i.i.d. assumption (the regret guarantee we obtained is the same as for the adversarial model). To me the key benefit is in the simplicity of the i.i.d. algorithm which makes it easy to incorporate further assumptions.

1. Sparsity of θ: instead of regularization with ℓ_2-norm to define $\hat{\theta}$ one could regularize with ℓ_1-norm, see e.g., Johnson, Sivakumar and Banerjee [2016].
2. Computational constraint: instead of optimizing over $\mathcal{E}_t$ to define $\widetilde{\theta}_t$ one could optimize over an ℓ_1-ball containing $\mathcal{E}_t$ (this would cost an extra $\sqrt{n}$ in the regret bound).
3. Generalized linear model: $\mathbb{E}(Y_t|a_t) = \sigma(\theta^\top a_t)$ for some known increasing $\sigma : \mathbb{R} \to \mathbb{R}$, see Filippi, Cappe, Garivier and Szepesvari [2011].
4. $\log(T)$-regime: if $\mathcal{A}$ is finite (note that a polytope is effectively finite for us) one can get $n^2 \log^2(T)/\Delta$ regret:
$$R_T \leq \mathbb{E} \sum_{t=1}^{T} \frac{(\theta^\top(a_t - a^*))^2}{\Delta} \leq \frac{\beta^2}{\Delta} \mathbb{E} \sum_{t=1}^{T} \|a_t\|^2_{\Sigma_t^{-1}} \lesssim \frac{n^2 \log^2(T)}{\Delta}.$$

Some non-linear bandit problems

Lipschitz bandit: Kleinberg, Slivkins and Upfal [2008, 2016], Bubeck, Munos, Stoltz and Szepesvari [2008, 2011];

Gaussian process bandit: Srinivas, Krause, Kakade and Seeger [2010]; and **convex bandit:**

Kleinberg 04
$R_T \lesssim n^3 T^{3/4}$

FKM 05
$R_T \lesssim \sqrt{n} T^{3/4}$

DHK/AHR 08
$R_T^{\mathcal{L}} \lesssim n^{3/2}\sqrt{T}$
$R_T^{\mathcal{L}} \gtrsim n\sqrt{T}$

ST 11
$R_T^{sm.} \lessapprox T^{2/3}$

AFHKR 11
$R_T^{i.i.d.} \lesssim n^{16}\sqrt{T}$

ADX 11
$R_T^{s.c.} \lessapprox T^{2/3}$

BCK 12
$R_T^{\mathcal{L}} \lesssim n\sqrt{T}$

S 12
$R_T^{\overset{s.c.}{sm.}} \gtrsim n\sqrt{T}$

HL 14
$R_T^{\overset{s.c.}{sm.}} \lesssim n^{3/2}\sqrt{T}$

BDKP 14
n=1
$R_T \lesssim \sqrt{T}$

DEK 15
$R_T^{sm.} \lessapprox T^{5/8}$

BE 15
$R_T \lesssim n^{11}\sqrt{T}$

HL16
$R_T \leq 2^{n^4} \log^{2n}(T)\sqrt{T}$

Contextual bandit

We now make the game-changing assumption that at the beginning of each round t a *context* $x_t \in \mathcal{X}$ is revealed to the player. The ideal notion of regret is now:

$$R_T^{\mathrm{ctx}} = \sum_{t=1}^{T} \ell_t(a_t) - \inf_{\Phi : \mathcal{X} \to \mathcal{A}} \sum_{t=1}^{T} \ell_t(\Phi(x_t)).$$

Contextual bandit

We now make the game-changing assumption that at the beginning of each round t a *context* $x_t \in \mathcal{X}$ is revealed to the player. The ideal notion of regret is now:

$$R_T^{\mathrm{ctx}} = \sum_{t=1}^{T} \ell_t(a_t) - \inf_{\Phi : \mathcal{X} \to \mathcal{A}} \sum_{t=1}^{T} \ell_t(\Phi(x_t)).$$

Contextual bandit

We now make the game-changing assumption that at the beginning of each round t a *context* $x_t \in \mathcal{X}$ is revealed to the player. The ideal notion of regret is now:

$$R_T^{\mathrm{ctx}} = \sum_{t=1}^{T} \ell_t(a_t) - \inf_{\Phi : \mathcal{X} \to \mathcal{A}} \sum_{t=1}^{T} \ell_t(\Phi(x_t)).$$

Sometimes it makes sense to restrict the mapping from contexts to actions, so that the infimum is taken over some *policy set* $\Pi \subset \mathcal{A}^{\mathcal{X}}$.

Contextual bandit

We now make the game-changing assumption that at the beginning of each round t a *context* $x_t \in \mathcal{X}$ is revealed to the player. The ideal notion of regret is now:

$$R_T^{\mathrm{ctx}} = \sum_{t=1}^{T} \ell_t(a_t) - \inf_{\Phi : \mathcal{X} \to \mathcal{A}} \sum_{t=1}^{T} \ell_t(\Phi(x_t)).$$

Sometimes it makes sense to restrict the mapping from contexts to actions, so that the infimum is taken over some *policy set* $\Pi \subset \mathcal{A}^{\mathcal{X}}$.

As far as I can tell the contextual bandit problem is an infinite playground and there is no canonical solution (or at least not yet!). Thankfully all we have learned so far can give useful guidance in this challenging problem.

Linear model after embedding

A natural assumption in several application domains is to suppose linearity in the loss after a correct embedding. Say we know mappings $(\varphi_a)_{a \in \mathcal{A}}$ such that $\mathbb{E}_t(\ell_t(a)) = \varphi_a(x_t)^\top \theta$ for some unknown $\theta \in \mathbb{R}^n$ (or in the adversarial case that $\ell_t(a) = \ell_t^\top \varphi_a(x_t)$).

Linear model after embedding

A natural assumption in several application domains is to suppose linearity in the loss after a correct embedding. Say we know mappings $(\varphi_a)_{a \in \mathcal{A}}$ such that $\mathbb{E}_t(\ell_t(a)) = \varphi_a(x_t)^\top \theta$ for some unknown $\theta \in \mathbb{R}^n$ (or in the adversarial case that $\ell_t(a) = \ell_t^\top \varphi_a(x_t)$).

This is nothing but a linear bandit problem where the action set is changing over time. All the strategies we described are robust to this modification and thus in this case one can get a regret of $\sqrt{nT \log(|\mathcal{A}|)} \lesssim n\sqrt{T \log(T)}$ (and for the stochastic case one can get efficiently $n^{3/2}\sqrt{T}$).

Linear model after embedding

A natural assumption in several application domains is to suppose linearity in the loss after a correct embedding. Say we know mappings $(\varphi_a)_{a \in \mathcal{A}}$ such that $\mathbb{E}_t(\ell_t(a)) = \varphi_a(x_t)^\top \theta$ for some unknown $\theta \in \mathbb{R}^n$ (or in the adversarial case that $\ell_t(a) = \ell_t^\top \varphi_a(x_t)$).

This is nothing but a linear bandit problem where the action set is changing over time. All the strategies we described are robust to this modification and thus in this case one can get a regret of $\sqrt{nT \log(|\mathcal{A}|)} \lesssim n\sqrt{T \log(T)}$ (and for the stochastic case one can get efficiently $n^{3/2}\sqrt{T}$).

A much more challenging case is when the correct embedding $\varphi = (\varphi_a)_{a \in \mathcal{A}}$ is only known to belong to some class Φ. Without further assumptions on Φ we are basically back to the general model. Also note that a natural impulse is to run "bandits on top of bandits", that is first select some $\varphi_t \in \Phi$ and then select a_t based on the assumption that φ_t is correct. We won't get into this here, but let us investigate a related idea.

Exp4, Auer, Cesa-Bianchi, Freund and Schapire [2001]

One can play exponential weights on the set of policies with the following unbiased estimator (obvious notation: $\ell_t(\pi) = \ell_t(\pi(x_t))$, $\pi_t \sim p_t$, and $a_t = \pi_t(x_t)$)

$$\widetilde{\ell}_t(\pi) = \frac{\mathbb{1}\{\pi(x_t) = a_t\}}{\sum_{\pi' : \pi'(x_t) = a_t} p_t(\pi')} \ell_t(a_t).$$

Exp4, Auer, Cesa-Bianchi, Freund and Schapire [2001]

One can play exponential weights on the set of policies with the following unbiased estimator (obvious notation: $\ell_t(\pi) = \ell_t(\pi(x_t))$, $\pi_t \sim p_t$, and $a_t = \pi_t(x_t)$)

$$\widetilde{\ell}_t(\pi) = \frac{\mathbb{1}\{\pi(x_t) = a_t\}}{\sum_{\pi' : \pi'(x_t) = a_t} p_t(\pi')} \ell_t(a_t).$$

Easy exercise: $R_T^{\mathrm{ctx}} \leq \sqrt{2T|\mathcal{A}|\log(|\Pi|)}$ (indeed the relative entropy term is smaller than $\log(|\Pi|)$ while the variance term is exactly $|\mathcal{A}|$).

Exp4, Auer, Cesa-Bianchi, Freund and Schapire [2001]

One can play exponential weights on the set of policies with the following unbiased estimator (obvious notation: $\ell_t(\pi) = \ell_t(\pi(x_t))$, $\pi_t \sim p_t$, and $a_t = \pi_t(x_t)$)

$$\widetilde{\ell}_t(\pi) = \frac{\mathbb{1}\{\pi(x_t) = a_t\}}{\sum_{\pi' : \pi'(x_t) = a_t} p_t(\pi')} \ell_t(a_t).$$

Easy exercise: $R_T^{\mathrm{ctx}} \leq \sqrt{2T|\mathcal{A}|\log(|\Pi|)}$ (indeed the relative entropy term is smaller than $\log(|\Pi|)$ while the variance term is exactly $|\mathcal{A}|$).

The only issue of this strategy is that the computationally complexity is linear in the policy space, which might be huge. A year and half ago a major paper by Agarwal, Hsu, Kale, Langford, Li and Schapire was posted, with a strategy obtaining the same regret as Exp4 (in the i.i.d. model) but which is also computationally efficient with an oracle for the offline problem (i.e., $\min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(\pi(x_t))$). Unfortunately the algorithm is not simple enough yet to be included in these slides.

The statistician perspective, after Goldenshluger and Zeevi [2009, 2011], Perchet and Rigollet [2011]

Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{A} = [n]$, (x_t) i.i.d. from some μ absolutely continuous w.r.t. Lebesgue. The reward for playing arm a under context x is drawn from some distribution $\nu_a(x)$ on $[0, 1]$ with mean function $f_a(x)$ which is assumed to be β-Holder smooth. Let $\Delta(x)$ be the "gap" function.

The statistician perspective, after Goldenshluger and Zeevi [2009, 2011], Perchet and Rigollet [2011]

Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{A} = [n]$, (x_t) i.i.d. from some μ absolutely continuous w.r.t. Lebesgue. The reward for playing arm a under context x is drawn from some distribution $\nu_a(x)$ on $[0, 1]$ with mean function $f_a(x)$ which is assumed to be β-Holder smooth. Let $\Delta(x)$ be the "gap" function.

A key parameter is the proportion of contexts with a small gap. The margin assumption is that for some $\alpha > 0$, one has

$$\mu(\{x : \Delta(x) \in (0, \delta)\}) \leq C\delta^{\alpha}, \forall \delta \in (0, 1].$$

The statistician perspective, after Goldenshluger and Zeevi [2009, 2011], Perchet and Rigollet [2011]

Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{A} = [n]$, (x_t) i.i.d. from some μ absolutely continuous w.r.t. Lebesgue. The reward for playing arm a under context x is drawn from some distribution $\nu_a(x)$ on $[0,1]$ with mean function $f_a(x)$ which is assumed to be β-Holder smooth. Let $\Delta(x)$ be the "gap" function.

A key parameter is the proportion of contexts with a small gap. The margin assumption is that for some $\alpha > 0$, one has

$$\mu(\{x : \Delta(x) \in (0, \delta)\}) \leq C\delta^{\alpha}, \forall \delta \in (0, 1].$$

One can achieve a regret of order $T \left(\frac{n \log(n)}{T}\right)^{\frac{\beta(\alpha+1)}{2\beta+d}}$, which is optimal at least in the dependency on T. It can be achieved by running Successive Elimination on an adaptively refined partition of the space, see Perchet and Rigollet [2011] for the details.

The online multi-class classification perspective after Kakade, Shalev-Shwartz, and Tewari [2008]

Here the loss is assumed to be of the following very simple form: $\ell_t(a) = \mathbb{1}\{a \neq a_t^*\}$. In other words using the context x_t one has to predict the best action (which can be interpreted as a *class*) $a_t^* \in [n]$.

The online multi-class classification perspective after Kakade, Shalev-Shwartz, and Tewari [2008]

Here the loss is assumed to be of the following very simple form: $\ell_t(a) = \mathbb{1}\{a \neq a_t^*\}$. In other words using the context x_t one has to predict the best action (which can be interpreted as a *class*) $a_t^* \in [n]$.

KSST08 introduces the *banditron*, a bandit version of the multi-class perceptron for this problem. While with full information the online multi-class perceptron can be shown to satisfy a "regret" bound on of order $\sqrt{T}$, the banditron attains only a regret of order $T^{2/3}$. See also Chapter 4 in Bubeck and Cesa-Bianchi [2012] for more on this.

Summary of advanced results

1. The optimal regret for the linear bandit problem is $\widetilde{O}(n\sqrt{T})$. In the Bayesian context Thompson Sampling achieves this bound. In the i.i.d. case one can use an algorithm based on the optimism in face of uncertainty together with concentration properties of the least squares estimator.
2. The i.i.d. algorithm can easily be modified to be computationally efficient, or to deal with sparsity in the unknown vector θ.
3. Extensions/variants: semi-bandit model, non-linear bandit (Lipschitz, Gaussian process, convex).
4. Contextual bandit is still a very active subfield of bandit theory.
5. Many important things were omitted. Example: knapsack bandit, see Badanidiyuru, Kleinberg and Slivkins [2013].

Some open problems we discussed

1. Prove the lower bound $\mathbb{E}R_T = \Omega(\sqrt{Tn\log(n)})$ for the adversarial n-armed bandit with adaptive adversary.
2. Guha and Munagala [2014] conjecture: for product priors, TS is a 2-approximation to the optimal Bayesian strategy for the objective of minimizing the number of pulls on suboptimal arms.
3. Find a "simple" strategy achieving the Bubeck and Slivkins [2012] best of both worlds result.
4. For the combinatorial bandit problem, find a strategy with regret at most $n^{3/2}\sqrt{T}$ (current best is $n^2\sqrt{T}$).
5. Is there a computationally efficient strategy for i.i.d. linear bandit with optimal $n\sqrt{T}$ gap-free regret and with $\log(T)$ gap-based regret?
6. Is there a natural framework to think about "bandits on top of bandits" (while keeping $\sqrt{T}$-regret)?