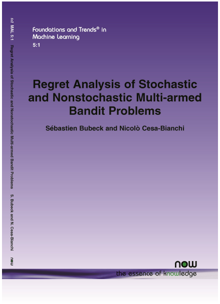
Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems, Part I

Sébastien Bubeck
Theory Group

Microsoft Research



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 $\nu_1, \dots, \nu_n$ on $[0,1]$ with mean $\mu_1, \dots, \mu_n$ (notation: $\mu^* = \max_{i \in [n]} \mu_i$).

Protocol: For each round $t = 1, 2, \dots, T$, 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 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 when $\xi = -1$.

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

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 when $\xi = -1$.

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

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,

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

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,

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

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 $t = 1, \dots, T$, 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 $\{(\nu_1(\theta), \ldots, \nu_n(\theta)), \theta \in \Theta\}$ and prior distribution π_0 over Θ. The Bayesian regret is defined as

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

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

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

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 $(\ell_1, \ldots, \ell_T) \in [0,1]^{n \times T}$, and let $\mathbb{E}_t$ denote the posterior distribution (given $\ell_1(I_1), \ldots, \ell_{t-1}(I_{t-1})$). 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):

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

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):

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

which implies:

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

Bayesian multi-armed bandit, accumulation of information

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

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.