$$\mathbb{E}R_n = \mathbb{E}\left[\max_{i=1,\dots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{A_t,t}\right]$$

pseudo-regret

$$\widetilde{R}_n = \max_{i=1,\dots,K} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{A_t,t}\right].$$

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

Lemma: Suppose that X is σ^2-subgaussian. Let X_1 and X_2 be independent and σ_1^2-subgaussian σ_2^2-subgaussian respectively, then:

- $\mathbb{E}[X] = 0$ and $\mathbb{V}[X] \leq \sigma^2$.
- cX is $c^2\sigma^2$-subgaussian for all $c \in \mathbb{R}$.
- $X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$-subgaussian.

Theorem: If X is σ^2-subgaussian, then

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

$$\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \text{ is } \frac{\sigma^2}{n}\text{-subgaussian.}$$

Corollary (Hoeffding's bound)

Assume that $X_i - \mu$ are independent, σ^2-subgaussian random variables. Then, their average $\hat{\mu}$ satisfies

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$
$$\mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

$$A_t = \begin{cases} i, & \text{if (t mod K)} + 1 = i \text{ and } t \leq mK; \\ \operatorname*{argmax}_i \hat{\mu}_i(mK), & t > mK, \end{cases}$$

$$
\begin{aligned}
R_n &= \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)] \\
&= m \sum_{i=1}^{K} \Delta_i + (n - mK) \sum_{i=1}^{K} \Delta_i \mathbb{P}\left(i = \operatorname*{argmax}_j \hat{\mu}_j(mK)\right) \\
&\leq m \sum_{i=1}^{K} \Delta_i + (n - mK) \sum_{i=1}^{K} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)
\end{aligned}
$$

$$
\mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \leq \delta
$$

$$
A_t = \begin{cases} \operatorname*{argmax}_i \left(\hat{\mu}_i(t-1) + \sqrt{\frac{2\log f(t)}{T_i(t-1)}}\right), & \text{if } t > K; \\ t, & \text{otherwise.} \end{cases}
$$

Corollary (Lattimore & Szepesvari): The regret of UCB is bounded by

$$
R_n \leq \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{1}{\Delta_i}\left(8\log f(n) + 8\sqrt{\pi \log f(n)} + 28\right)\right).
$$

and in particular there exists some universal constant $C > 0$ such that for all $n \geq 2$,

$$
R_n \leq \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{C\log n}{\Delta_i}\right).
$$

$$
\begin{aligned}
T_i(n) &= \sum_{t=1}^{n} \mathbb{I}\{A_t = i\} \\
&\leq \sum_{t=1}^{n} \left\{\mathbb{I}\left\{\hat{\mu}_1(t-1) + \sqrt{\frac{2\log f(t)}{T_1(t-1)}} \leq \mu_1 - \varepsilon\right\}\right\} \\
&+ \sum_{t=1}^{n} \left\{\mathbb{I}\left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2\log f(t)}{T_i(t-1)}} \geq \mu_1 - \varepsilon \text{ and } A_t = i\right\}\right\}.
\end{aligned}
$$

$$
\mathbb{E}\left[\max_i \sum_{t=1}^{n} X_{t,i} - X_{t,A_t}\right] \geq \max_i \mathbb{E}\left[\sum_{t=1}^{n} X_{t,i} - X_{t,A_t}\right].
$$

1 Adversial Bandits

$$\begin{aligned}
R_n(\pi,\nu) &= \max_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{n}(X_{t_i} - X_{t,A_t})\right] \\
&\le \mathbb{E}\left[\max_{i\in[K]} \sum_{t=1}^{n}(X_{t_i} - X_{t,A_t})\right] \\
&= \mathbb{E}\left[R_n(\pi,X)\right] \le R_n^*(\pi)
\end{aligned}$$

Proof We define:

$$\Phi_t := \sum_{i=1}^{K} w_t(i),$$

In particular, $\Phi_1 = K$, since $w_t(1) = 1$ at initialization. Thus, for each $t \in \{1,\dots,T-1\}$,

$$\begin{aligned}
\Phi_{t+1} &= \sum_{i=1}^{K} w_t(i)e^{-\ell_t(i)} = \Phi_t \sum_{i=1}^{K} p_t(i)e^{-\ell_t(i)} && (1)\\
&\le \Phi_t \sum_{i=1}^{K} p_t(i)(1-\ell_t(i)+\ell_t^2(i)) && (2)\\
&= \Phi_t(1-\epsilon p_t\ell_t + \epsilon^2 p_t\ell_t^2) && (3)\\
&\le \Phi_t \dots \exp(-\epsilon p_t\ell_t + \epsilon^2 p_t\ell_t^2). && (4)
\end{aligned}$$

$$E\left[\sum_{t=1}^{n}\ell_t(i_t)\right] \le \min_i \sum_{t=1}^{n}\ell_t(i) + \epsilon \dots E\left[\sum_{t=1}^{n}\ell_t^2(i_t)\right] + \frac{\log K}{\epsilon}.$$

Explanations for lines (2)-(5) are as given below:

- 2nd equality follows from $p(t_i) = \frac{w_t(i)}{\sum_{j=1}^{K} w_t(j)}$,
- (3) follows by observing that $e^{-x} \le 1 - x + x^2$ for each $x \ge 0$,
- (4) follows by interpreting the sum as an inner product, and defining $\ell_{t,s}$ as $\ell_{t,s} := ((\ell_t(1))^2,\dots,(\ell_t(K))^2)$,
- (5) follows by observing that $1 + x \le e^x$ for each $x \in \mathbb{R}$.

By concatenating the above chain of inequalities across $t = 1,\dots,n$, we have, for each expert i,

$$w_1(i)\exp\left(\epsilon\sum_{t=1}^{n}\ell_t(i))\right) = w_n(i) \le \Phi_n \le \Phi_1 \dots \exp\left(\sum_{t=1}^{n}\left[-\epsilon p_t\ell_t + \epsilon^2 p_t\ell_t^2\right]\right).$$

Taking the logarithm on both sides gives

$$-\epsilon \sum_{t=1}^{n} \ell_t(i) \leq \log K - \epsilon \ldots \sum_{t=1}^{n} p_t \ell_t + \epsilon^2 \ldots \sum_{t=1}^{n} p_t \ell_t^2.$$

Finally, by dividing both sides by ϵ and rearranging terms, we obtain the desired theorem statement.

Algorithm 1 Hedge Algorithm

$W_1(i) \leftarrow 1$
for $t = 1, \ldots, n$ **do**
  $\ell_t(i) \leftarrow$ loss of expert $i,$ for each $i = 1, \ldots, K$
  $i_t \sim$ Expert index selected by drawing from $p_t(i) = \frac{W_t(i)}{\sum_{j=1}^{k} W_t(j)}$
  $\ell_t(i_t) \leftarrow$ Loss incurred at time t
  $W_{t+1}(i) \leftarrow W_t(i) \ldots e^{-\epsilon \ell_t(i)}$ (Weight update)
end for

$$\epsilon = \sqrt{\frac{8 \log K}{n}}$$
$$R_n \in \Theta(\sqrt{n \log K})$$

Algorithm 2 Adversarial Bandits, Setup

$\{x_t\}_{t=1}^n := \{(x_{t,1}, \ldots, x_{t,K}) \in [0,1]^K\}_{t=1}^n$ ▷ Reward vectors selected by the adversary
for time $t = 1, \ldots, n$ **do**
  $P_t(A_t|H_{t-1}) \leftarrow$ Distribution of action at time t conditioned on H_{t-1}, selected by the learner.
  $A_t \sim P_t(A_t|H_{t-1})$ ←Learner's action at time t, sampled from P_t.
  $X_t := x_{t,A_t}$ ←Reward observed by learner at time t.
end for

$$\mathrm{Var}\left[\hat{X}_{t,i}|\mathcal{H}_{t-1}\right] = \mathbb{E}\left[\frac{1_{\{A_t=i\}}}{P_{t,i}^2} \ldots X_t^2\right] - x_{t,i}^2 = x_{t,i}^2 \ldots \frac{1 - P_{t,i}}{P_{t,i}} \quad (5)$$

$$\begin{aligned} \exp(\eta \hat{S}_{t_i}) \leq \sum_{j=1}^{K} \exp(\eta \hat{S}_{t_j}) &= W_n \\ &= \frac{W_1}{W_0} \ldots \frac{W_2}{W_1} \ldots \frac{W_n}{W_{n-1}} \\ &= K \prod_{t=1}^{n} \frac{W_t}{W_{t-1}} \end{aligned}$$

Algorithm 3 EXP3 Algorithm

Input: horizon n, number of arms K, learning rate η;
Set $\hat{S}_{0,i} = 0$ for all $1 \leq i \leq K$;
for $t = 1, 2, \ldots, n$ **do**
  $P_{t,i} = \frac{\exp(\eta \hat{S}_{t-1,i})}{\sum_{j \in [k]} \exp(\eta \hat{S}_{t-1,j})}$;
  Sample $A_t \sim P_t$, receive X_t;
  Update $\hat{S}_{t,i} = \hat{S}_{t-1,i} + 1 - \frac{\mathbb{I}(A_t = i)(1 - X_t)}{P_{t,i}} = \hat{S}_{t-1,i} + \hat{X}_{t,i}$;
end for

$$\begin{aligned}
\frac{W_t}{W_{t-1}} &= \sum_j \frac{\exp(\eta \hat{S}_{t-1,j})}{W_{t-1}} \exp(\eta \hat{X}_{tj}) = \sum_j P_{tj} \exp(\eta \hat{X}_{tj}) && (6) \\
&= \sum_j P_{tj} \exp(\eta) \exp(\eta(\hat{X}_{tj} - 1)) && (7) \\
&\leq \exp(\eta)(1 + \eta \sum_j P_{tj} \hat{X}_{tj} + \eta^2 \sum_j P_{tj} \hat{X}_{tj}^2) && (8) \\
&\leq \exp\left(\eta \sum_j P_{tj} \hat{X}_{tj} + \frac{\eta^2}{2} \sum_j P_{tj} (\hat{X}_{tj} - 1)^2 \right) && (9)
\end{aligned}$$

Theorem 1 (Lattimore. Theorem 11.2). For rewards $x_{t,i} \in [0,1]$, and the learning rate tuned to $\eta = \sqrt{2\log(K)/(nK)}$, we have for any arm i

$$R_{n,i} \leq \sqrt{2nK\log(K)}$$

$$R_n^* \geq c\sqrt{nK}$$

$$\hat{Y}_{ti} = 1 - \hat{X}_{ti}$$

$$\hat{Y}_{ti} = \frac{\mathbb{I}\{A_t = i\} Y_t}{P_{ti} + \gamma} \qquad \text{where } \hat{Y}_{ti} = 1 - \hat{X}_{ti}$$

$$\eta_1 = \sqrt{\frac{2\log(K+1)}{nK}}$$

$$\eta_2 = \sqrt{\frac{\log(K) + \log(\frac{K+1}{\delta})}{nK}}$$

The following hold:

1 If Exp3-IX is run with parameters $\eta = \eta_1$ and $\gamma = \eta/2$, then

$$\mathbb{P}\left(\hat{R}_n \geq \sqrt{8.5nK\log(K+1)} + \left(\sqrt{\frac{nK}{2\log(K+1)}} + 1 \right) \log\left(\frac{1}{\delta} \right) \right) \leq \delta.$$

2 If Exp3-IX is run with parameters $\eta = \eta_2$ and $\gamma = \eta/2$, then

$$\mathbb{P}\left(\hat{R}_n \geq 2\sqrt{(2\log(K+1) + \log(1/\delta))nK} + \log\left(\frac{K+1}{\delta} \right) \right) \leq \delta.$$

2 Contextual Bandits

$$
\begin{aligned}
S_n &= \sum_{c \in C} \max_{k \in [K]} \sum_{t : c_t = c} x_{t,k} \\
&= \max_{\phi : C \to [K]} \sum_{t=1}^{n} x_{t,\phi(c_t)}.
\end{aligned}
$$

$$
\begin{aligned}
R_n = S_n - \sum_t X_t &= \sum_{c \in C} \mathbb{E}\left[\max_{k \in |K|} \sum_{t : c_t = c} (x_{t,k} - X_t)\right] \\
R_n &= \sum_{c \in C} \mathbb{E}\left[R^c\left(T^c(n)\right)\right].
\end{aligned}
$$

$$
\eta_s = \sqrt{\frac{\log(K)}{sK}}, \text{ then one can show that } R^c(s) \leq 2\sqrt{sK\log(K)}
$$

$$
\begin{aligned}
R_n &= \sum_{c \in C} \mathbb{E}\left[R^c\left(T^c(n)\right)\right] \\
&\leq \sqrt{2|C|nK\ln K}
\end{aligned}
$$

By Jensen's inequality, since $f(x) = x^2$ is convex when $x \geq 0$

$$
\begin{aligned}
R_n &= \sum_{c \in C} \mathbb{E}\left[R^c\left(T^c(n)\right)\right] \\
&\leq \sum_{c \in C} \sqrt{2T^c(n)K\ln K} \\
&= \sqrt{2K\ln K} \sum_{c \in C} \sqrt{T^c(n)} \\
&\leq \sqrt{2K\ln K} \sqrt{|C| \sum_{c \in C} (\sqrt{T^c(n)})^2} \\
&= \sqrt{2|C|nK\ln K}
\end{aligned}
$$

$$
\text{since } \sum_c T^c(n) = n
$$

$$
\begin{aligned}
S_n &= \sum_{P \in \mathcal{P}} \max_{k \in |K|} \sum_{t : c_t \in P} x_{t,k} \\
&= \max_{\phi \in \Phi(P)} \sum_{t=1}^{n} x_{t,\phi(c_t)}
\end{aligned}
$$

2.1 Contextual Bandits with Expert Advice

$$R_n = \mathbb{E}\left[\max_m \sum_{t=1}^{n} E_m^{(t)} x_t - \sum_{t=1}^{n} X_t\right] \tag{10}$$

for $t = 1$ **to** n **do**
  Receive the advice $E^{(t)}$
  Choose the action $A_t \sim P_{t,\cdot}$ at random, where
  Receive the reward $X_t = x_{t,A_t}$
  Estimate the rewards of all the actions; say: $\hat{X}_{ti} = 1 - \frac{\mathbb{I}A_t \neq i}{P_{ti}+\gamma}(1 - X_t)$
  Propagate the rewards to the experts: $\tilde{X}_t = E^{(t)}\hat{X}_t$
end for

Algorithm 4 EXP4 Algorithm

Input: horizon n, number of arms K, learning rate η, number of experts M;
$\hat{S}_{1,m} = 0$, $Q_{1,m} = 1/M$ for all $1 \leq m \leq M$
for $t = 1, 2, \ldots, n$ **do**
  Receive the advice $E^{(t)}$, calculate $P_t = Q_t E^{(t)}$
  Sample $A_t \sim P_t$ where , receive $X_t = x_{t,A_t}$;
  Estimate the rewards of all arm:

$$\hat{X}_{t,i} = 1 - \frac{\mathbb{I}\{A_t = i\}}{P_{ti}}(1 - X_t), \quad i \in [K]$$

  Estimate the reward vectors of all expert: $\tilde{X}_t = E^{(t)}\hat{X}_t$
  Update the estimated cumulative reward for each expert

$$\hat{S}_{t+1,m} = \hat{S}_{t,m} + \tilde{X}_{t,m}$$

  Update the distribution Q_t using exponential weighting:

$$Q_{t+1,m} = \frac{\exp(\eta \hat{S}_{t+1,m})}{\sum_{m'} \exp(\eta \hat{S}_{t+1,m'})}, \quad m \in [M]$$

end for

Definition	Matrix	Size
Expert advice	$E^{(t)}$	M x K
Expert weights	Q_t	1 x M
Probability of choosing arm	P_t	1 x K
Estimated reward of all arms	$\hat{X}_t$	K x 1
Estimated reward of all experts	$\tilde{X}_t$	M x 1

$$\sum_{t=1}^{n} \hat{X}_{ti} - \sum_{t=1}^{n}\sum_{j=1}^{K} P_{tj}\hat{X}_{tj} \leq \frac{\log(M)}{\eta} + \frac{\eta}{2}\sum_{t,j} P_{tj}(1 - \hat{X}_{tj})^2. \tag{11}$$

Lemma: Let $\hat{X}_{t,i}$ and $P_{t,i}$ satisfy, for all $t \in [n]$ and $i \in [K]$, the relations $\hat{X}_{ti} \leq 1$ and:

$$P_{ti} = \frac{\exp\left(\eta \sum_{s=1}^{t} \hat{X}_{ti}\right)}{\sum_j \exp\left(\eta \sum_{s=1}^{t} \hat{X}_{tj}\right)}.$$

Then, for any $i \in [K]$,

$$\sum_{t=1}^{n} \hat{X}_{t,i} - \sum_{t=1}^{n} \sum_{j=1}^{K} P_{t,j} \hat{X}_{t,j} \leq \frac{\log(K)}{\eta} + \frac{\eta}{2} \sum_{t,j} P_{t,j} (1 - \hat{X}_{t,j})^2.$$

Let $\hat{X}_{t,i}$ and $P_{t,i}$ satisfy, for all $t \in [n]$ and $i \in [K]$, the relations $\hat{X}_{ti} \leq 1$ and:

$$P_{ti} = \frac{\exp\left(\eta \sum_{s=1}^{t} \hat{X}_{ti}\right)}{\sum_j \exp\left(\eta \sum_{s=1}^{t} \hat{X}_{tj}\right)}.$$

Then, for any $i \in [K]$,

$$\sum_{t=1}^{n} \tilde{X}_{t,m} - \sum_{t=1}^{n} \sum_{m'} Q_{t,m'} \tilde{X}_{t,m'} \leq \frac{\log(M)}{\eta} + \frac{\eta}{2} \sum_{t,m'} Q_{t,m'} (1 - \tilde{X}_{t,m'})^2.$$

$$\mathbb{E}\left[\sum_{t,m'} Q_{t,m'} (1 - \tilde{X}_{t,m'})^2\right]$$

$$\tilde{Y}_{t,m} = 1 - \tilde{X}_{t,m}$$

$$\hat{Y}_{t,m} = 1 - \hat{X}_{t,m}$$

$$\tilde{Y}_t = E^{(t)} \hat{Y}_t$$

$$\tilde{Y}_{t,m} = E^{(t)}_m \hat{Y}_t$$

$$\begin{aligned}
\mathbb{E}\left[\tilde{Y}_{t,m}^2\right] &= \mathbb{E}\left[\left(E^{(t)}_m \hat{Y}_t\right)^2\right] = \mathbb{E}\left[\left(\sum_i \frac{E^{(t)}_{m,i} \mathbb{I}\{A_t = i\} y_{t,i}}{P_{t,i}}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{E^{(t)}_{m,A_t} y_{t,A_t}}{P_{t,A_t}}\right)^2\right] = \sum_i \frac{(E^{(t)}_{m,i})^2 y_{t,i}^2}{P_{t,i}} \leq \sum_i \frac{(E^{(t)}_{m,i})^2}{P_{t,i}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{m'} Q_{t,m'}(1-\tilde{X}_{t,m'})^2\right] &= \mathbb{E}\left[\sum_m Q_{t,m}\tilde{Y}_{t,m}^2\right] \\
&\le \sum_m Q_{t,m}\sum_i \frac{(E_{m,i}^{(t)})^2}{P_{t,i}} \\
&\le \sum_i \left(\max_{m'} E_{m',i}^{(t)}\right)^2 \sum_m \frac{Q_{t,m}E_{m,i}^{(t)}}{P_{t,i}}.
\end{aligned}$$

Defining

$$E_n^* = \sum_{t,i}\left(\max_{m'} E_{m',i}^{(t)}\right),$$

Theorem (Regret of Exp4): If $\eta = \sqrt{\frac{2\log(M)}{E_n^*}}$, the regret of Exp4 satisfies

$$R_n \le \mathbb{E}\left[\sqrt{2\log(M)E_n^*}\right]$$

Since $\sum_i E_{m,i}^{(t)} = 1$ for all $m \in [M]$,

$$\sum_i \left(\max_m E_{m,i}^{(t)}\right) \ge 1$$

$$E_n^* = \sum_t\sum_i\left(\max_{m'} E_{m',i}^{(t)}\right) \ge \sum_t 1 \ge n$$

$$R_n \le \mathbb{E}\left[\sqrt{2n\log(M)}\right]$$

Since $E_{m,i}^{(t)} \le 1$ for all $m \in [M]$,

$$E_n^* = \sum_t\sum_i\left(\max_m E_{m,i}^{(t)}\right) \le \sum_t\sum_i 1 \le nK$$

$$R_n \le \mathbb{E}\left[\sqrt{2nK\log(M)}\right]$$

Since $\max_m E_{m,i}^{(t)} \le \sum_m E_{m,i}^{(t)}$ for all $m \in [M]$,

$$E_n^* = \sum_t\sum_i\left(\max_m E_{m,i}^{(t)}\right) \le \sum_t\sum_i\sum_m E_{m,i}^{(t)} \le \sum_t\sum_m\sum_i E_{m,i}^{(t)} \le nM$$

$$R_n \le \mathbb{E}\left[\sqrt{2nM\log(M)}\right]$$

$$E_n^* = \min(M,K)n$$

3 Bayesian Bandits

$$X_i\text{: Rewards of arm i, a random variable}$$
$$P(X_i;\theta)\text{: Unknown reward distribution, parameterized by } \theta_i$$
$$P(\theta)\text{: Prior belief about the distribution of } \theta_i$$

We are about to compute the posterior distribution after observing the reward of arm i at time step 1 to t $\mathbb{P}(\theta|x_{1,i}, x_{2,i}, \dots, x_{n,i})$, by Bayes' Theorem:

$$P(\theta_i|x_{1,i}, x_{2,i}, \dots, x_{t,i}) \propto P(\theta_i)P(x_{1,i}, x_{2,i}, \dots, x_{t,i}|\theta_i)$$
$$P(\theta_i|x_{1,i}, x_{2,i}, \dots, x_{t,i}) = P(R(i)|x_{1,i}, x_{2,i}, \dots, x_{t,i})$$

If we have the posterior distribution, we can compute

- Distribution of tyhe next reward of arm i $X_{t+1,i}$

$$P(X_{t+1,i}|x_{1,i}, \dots, x_{t,i}) = \int_\theta P(X_{t+1,i};\theta_i)P(\theta_i|x_{1,i}, \dots, x_{t,i})d\theta$$

- $\hat{R}_{t,i} = \frac{1}{k}\sum_{j=1}^{k} R_{j,i}$
- $R_{j,i} \sim P(R_i|x_{1,i}, \dots, x_{t,i})$
- $x_{t,i} \sim P(x_i;\theta)$
- $\hat{X}_{t,i} = \frac{1}{t}\sum_{j=1}^{t} x_{j,i}$
- $A_t = \arg\max_i \hat{X}_{t,i}$

Algorithm 5 Thompson Samplings

Input: horizon n, number of arms K, parameters α and β;
for $t = 1, \dots, n$ **do**
    for $i = 1, \dots, K$ **do**
        Sample $\hat{\theta}_i \sim \text{Beta}(\alpha_i, \beta_i)$
    end for
    $A_t \leftarrow \arg\max_i \hat{\theta}_i$
    Take action A_t and observe x_{t,A_t}
    $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{A_t} + x_{t,A_t}, \beta_{A_t} + 1 - x_{t,A_t})$
end for

$$\begin{aligned} P(\theta_i|x_{t,i}) &\propto P(x_{t,i}|\theta_i)P(\theta_i) \\ &\propto \theta_i^{x_{t,i}}(1-\theta_i)^{1-x_{t,i}}\theta_i^{\alpha_i-1}(1-\theta_i)^{\beta_i-1} \\ &= \theta_i^{\alpha_i-1+x_{t,i}}(1-\theta_i)^{\beta_i-1+1-x_{t,i}} \\ &\propto \text{Beta}(\alpha_i + x_{t,i}, \beta_i + 1 - x_{t,i}) \end{aligned}$$

$$\max_{\theta'} \mathbb{E}[\text{Regret}_n|\theta=\theta'] = O\left(\sqrt{Kn\log(n)}\right)$$

Algorithm 6 Greedy Method of Bayesian

Input: horizon n, number of arms K, parameters α and β;
for $t = 1, \ldots, n$ **do**
  for $i = 1, \ldots, K$ **do**
    $\hat{\theta}_i \leftarrow \alpha_i/(\alpha_i + \beta_i)$
  end for
  $A_t \leftarrow \arg\max_i \hat{\theta}_i$
  Take action A_t and observe x_{t,A_t}
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{A_t} + x_{t,A_t}, \beta_{A_t} + 1 - x_{t,A_t})$
end for

4 Gittins Index

$$V(p;\alpha,\beta,\gamma) = \max\left\{\frac{p}{1-\gamma}, \frac{\alpha}{\alpha+\beta} + \gamma\left(\frac{\alpha}{\alpha+\beta}V(p;\alpha+1,\beta,\gamma) + \frac{\beta}{\alpha+\beta}V(p;\alpha,\beta+1,\theta)\right)\right\}$$

$$\frac{\alpha}{\alpha+\beta} + \gamma\left(\frac{\alpha}{\alpha+\beta}V(p;\alpha+1,\beta,\gamma) + \frac{\beta}{\alpha+\beta}V(p;\alpha,\beta+1,\theta)\right)$$

4.1 General Case

$$J_\pi(\vec{s}) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_{i_t}(s_{i_t}(t)) \middle| \vec{s}(0) = \vec{s}\right]$$

$$\lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_{i_t}(s_{i_t}(t))\right]$$

$$J(s_i) = \sup_{\tau>0} \mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t (r_i(s_i(t) - \lambda) \middle| s_i(0) = s_i\right] = 0$$

$$G_i(s_i) = \sup_{\tau>0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t r_i(s_i(t)) \middle| s_i(0) = s_i\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t \middle| s_i(0) = s_i\right]}$$

$$E_\pi\left[\sum_{t=0}^{\infty} \gamma^t \left(r_{i_t}(x_{i_t}(t)) - \lambda_{i_t}(x_{it}(t))\right) \middle| x(0)\right] \le 0$$

$$E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_{i_t}(x_{i_t}) \middle| x(0)\right] \le E_\pi\left[\sum_{t=0}^{\infty} \gamma^t \lambda_{i_t}(x_{i_t}) \middle| x(0)\right]$$

$$E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_{i_t}(x_{i_t}) \middle| x(0)\right] \le E_\pi\left[\sum_{t=0}^{\infty} \gamma^t \lambda_{i_t}(x_{i_t}) \middle| x(0)\right] \le E_{\pi^*}\left[\sum_{t=0}^{\infty} \gamma^t \lambda_{i_t}(x_{i_t}) \middle| x(0)\right]$$

$$
\begin{aligned}
\text{maximize} \quad & \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \sum_{i=1}^{n} r_i(s_i, u_i)\right] \\
\text{subject to} \quad & \sum_{i=1}^{n} u_i(t) = m, \quad \forall t, \\
& u_i(t) \in \{0, 1\}, \quad \forall i.
\end{aligned}
$$

4.2 Relaxed Constraints

$$
\begin{aligned}
\text{maximize} \quad & \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \sum_{i=1}^{n} r_i(s_i, u_i)\right] \\
\text{subject to} \quad & \mathbb{E}\sum_{t=0}^{T-1} \gamma^t \sum_{i=1}^{n} u_i(t) = m/(1-\gamma), \\
& u_i(t) \in \{0, 1\}, \quad \forall i.
\end{aligned}
$$

$$
\mathbb{E}\sum_{t=0}^{\infty} \gamma^t \sum_{i=1}^{n} u_i(t) = \sum_{t=0}^{\infty} \gamma^t m = m/(1-\gamma)
$$

$$
\text{maximize} \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \sum_{i=1}^{n} \left(r_i(s_i, u_i) - \lambda u_i(t)\right)\right] + \lambda\left(m/(1-\gamma)\right)
$$

4.3 Decoupled

$$
\text{maximize} \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \left(r_i(s_i, u_i) - \lambda u_i(t)\right)\right]
$$