Solutions to Selected Exercises in Bandit Algorithms

Tor Lattimore and Csaba Szepesvári

Draft of Saturday 16th February, 2019

Contents

Chapter 2 Foundations of Probability

2.1 Let $h = g \circ f$. Let $A \in \mathcal{H}$. We need to show that $h^{-1}(A) \in \mathcal{F}$. We claim that $h^{-1}(A) = f^{-1}(g^{-1}(A))$. Because g is $\mathcal{G}/\mathcal{H}$-measurable, $g^{-1}(A) \in \mathcal{G}$ and thus because f is $\mathcal{F}/\mathcal{G}$-measurable, $f^{-1}(g^{-1}(A))$ is $\mathcal{F}$-measurable, thus completing the proof, once we show that the claim holds. To show the claim, we show two-sided containment. For showing $h^{-1}(A) \subset f^{-1}(g^{-1}(A))$ let $x \in h^{-1}(A)$. Thus, $h(x) \in A$. By definition, $h(x) = g(f(x)) \in A$. Hence, $f(x) \in g^{-1}(A)$ and thus $x \in f^{-1}(g^{-1}(A))$. For the other direction let $x \in f^{-1}(g^{-1}(A))$. This implies that $f(x) \in g^{-1}(A)$, which implies that $h(x) = g(f(x)) \in A$.

2.3 Since $X(u) \in \mathcal{V}$ for all $u \in \mathcal{U}$ we have $X^{-1}(\mathcal{V}) = \mathcal{U}$. Therefore $\mathcal{U} \in \Sigma_X$. Suppose that $U \in \Sigma_X$, then by definition there exists a $V \in \Sigma$ such that $X^{-1}(V) = U$. Because Σ_X is a σ-algebra we have $V^c \in \Sigma$ and by definition of Σ_X we have $U^c = X^{-1}(V^c) \in \Sigma_X$. Therefore Σ_X is closed under complements. Finally let $(U_i)_i$ be a countable sequence with $U_i \in \Sigma_X$. Then $\bigcup_i U_i = X^{-1}(\cup_i X(U_i)) \in \Sigma_X$, which means that Σ_X is closed under countable unions and the proof is completed.

2.5

(a) Let $\mathcal{A}$ be the set of all σ-algebras that contain $\mathcal{G}$ and define

$$\mathcal{F}^* = \bigcap_{\mathcal{F} \in \mathcal{A}} \mathcal{F}\,.$$

We claim that $\mathcal{F}^*$ is the smallest σ-algebra containing $\mathcal{G}$. Clearly $\mathcal{F}^*$ contains $\mathcal{G}$ and is contained in all σ-algebras containing $\mathcal{G}$. Furthermore, by definition it contains exactly those A that are in every σ-algebra that contains $\mathcal{G}$. It remains to show that $\mathcal{F}^*$ is a σ-algebra. Since $\Omega \in \mathcal{F}$ for all $\mathcal{F} \in \mathcal{A}$ it follows that $\Omega \in \mathcal{F}^*$. Now suppose that $A \in \mathcal{F}^*$. Then $A \in \mathcal{F}$ for all $\mathcal{F} \in \mathcal{A}$ and $A^c \in \mathcal{F}$ for all $\mathcal{F} \in \mathcal{A}$. Therefore $A^c \in \mathcal{F}^*$. Therefore $\mathcal{F}^*$ is closed under complements. Finally, suppose that $(A_i)_i$ is a family in $\mathcal{F}^*$. Then $(A_i)_i$ are families in $\mathcal{F}$ for all $\mathcal{F} \in \mathcal{A}$ and so $\bigcup_i A_i \in \mathcal{F}$ for all $\mathcal{F} \in \mathcal{A}$ and again we have $\bigcup_i A_i \in \mathcal{F}^*$. Therefore $\mathcal{F}^*$ is a σ-algebra.

(b) Define $\mathcal{H} = \{A : X^{-1}(A) \in \mathcal{F}\}$. Then $\Omega \in \mathcal{H}$ and for $A \in \mathcal{H}$ we have $X^{-1}(A^c) = X^{-1}(A)^c$ so A^c in $\mathcal{H}$. Furthermore, for $(A_i)_i$ with $A_i \in \mathcal{H}$ we have

$$X^{-1}\left(\bigcup_i A_i\right) = \bigcup_i X^{-1}(A_i)\,.$$

Therefore $\mathcal{H}$ is a σ-algebra on Ω and by definition $\sigma(\mathcal{G}) \subseteq \mathcal{H}$. Now for any $A \in \mathcal{H}$ we have $f^{-1}(A) \in \mathcal{F}$ by definition. Therefore $f^{-1}(A) \in \mathcal{F}$ for all $A \in \sigma(\mathcal{G})$.

(c) We need to show that $\mathbb{I}\{A\}^{-1}(B) \in \mathcal{F}$ for all $B \in \mathfrak{B}(\mathbb{R})$. There are four cases. If $\{0,1\} \in B$, then $\mathbb{I}\{A\}^{-1}(B) = \Omega \in \mathcal{F}$. If $\{1\} \in B$, then $\mathbb{I}\{A\}^{-1}(B) = A \in \mathcal{F}$. If $\{0\} \in B$, then

$\mathbb{I}\{A\}^{-1}(B) = A^c \in \mathcal{F}$. Finally, if $\{0,1\} \cap B = \emptyset$, then $\mathbb{I}\{A\}^{-1}(B) = \emptyset \in \mathcal{F}$. Therefore $\mathbb{I}\{A\}$ is $\mathcal{F}$-measurable.

2.6 Trivially, $\sigma(X) = \{\emptyset, \mathbb{R}\}$. Hence Y is not $\sigma(X)/\mathfrak{B}(\mathbb{R})$-measurable because $Y^{-1}([0,1]) = [0,1] \notin \sigma(X)$.

2.7 First $\mathbb{P}(\emptyset \,|\, B) = \mathbb{P}(\emptyset \cap B)/\mathbb{P}(B) = 0$ and $\mathbb{P}(\Omega \,|\, B) = \mathbb{P}(\Omega \cap B)/\mathbb{P}(B) = 1$. Let $(E_i)_i$ be a countable collection of disjoint sets with $E_i \in \mathcal{F}$. Then

$$\begin{aligned}\mathbb{P}\left(\bigcup_i E_i \,\middle|\, B\right) &= \frac{\mathbb{P}(B \cap \bigcup_i E_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\bigcup_i (B \cap E_i))}{\mathbb{P}(B)} \\ &= \sum_i \frac{\mathbb{P}(B \cap E_i)}{\mathbb{P}(B)} = \sum_i \mathbb{P}(E_i \,|\, B)\ .\end{aligned}$$

Therefore $\mathbb{P}(\,\cdot \,|\, B)$ satisfies the countable additivity property and the proof is complete.

2.8 Using the definition of conditional probability and the assumption that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ we have:

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \,|\, A)\,\mathbb{P}(A)}{\mathbb{P}(B)}\,.$$

2.9 For part (a),

$$\mathbb{P}(X_1 < 2 \,|\, X_2 \text{ is even}) = \frac{\mathbb{P}(X_1 < 2 \text{ and } X_2 \text{ is even})}{\mathbb{P}(X_2 \text{ is even})} = \frac{3/(6^2)}{18/(6^2)} = \frac{1}{6} = \mathbb{P}(X_1 < 2)\ .$$

Therefore $\{X_1 < 2\}$ is independent from $\{X_2 \text{ is even}\}$. For part (b) note that $\sigma(X_1) = \{C \times [6] : C \in 2^{[6]}\}$ and $\sigma(X_2) = \{[6] \times C : C \in 2^{[6]}\}$. It follows that for $|A \cap B| = |A||B|/6^2$ and so

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{|A \cap B|/6^2}{|B|/6^2} = \frac{|A|}{6^2} = \mathbb{P}(A)\ .$$

Therefore A and B are independent.

2.10

(a) Let $A \in \mathcal{F}$. Then $\mathbb{P}(A \cap \Omega) = \mathbb{P}(A) = \mathbb{P}(A)\,\mathbb{P}(\Omega)$ and $\mathbb{P}(A \cap \emptyset) = 0 = \mathbb{P}(\emptyset)\,\mathbb{P}(A)$. Intuitively, Ω and $\emptyset$ happen surely/never respectively, so the occurrence or not of any other event cannot alter their likelihood.

(b) Let $A \in \mathcal{F}$ satisfy $\mathbb{P}(A) = 1$ and $B \in \mathcal{F}$ be arbitrary. Then $\mathbb{P}(B \cap A^c) \leq \mathbb{P}(A^c) = 0$. Therefore $\mathbb{P}(A \cap B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) = \mathbb{P}(B) = \mathbb{P}(A)\,\mathbb{P}(B)$. When $\mathbb{P}(A) = 0$ we have $\mathbb{P}(A \cap B) \leq \mathbb{P}(A) = 0 = \mathbb{P}(A)\,\mathbb{P}(B)$.

(c) If A and A^c are independent, then $0 = \mathbb{P}(\emptyset) = \mathbb{P}(A \cap A^c) = \mathbb{P}(A)\,\mathbb{P}(A^c) = \mathbb{P}(A)\,(1 - \mathbb{P}(A))$. Therefore $\mathbb{P}(A) \in \{0,1\}$. This makes sense because the knowledge of A provides the knowledge of A^c, so the two events can only be independent if one occurs with probability zero.

(d) If A is independent of itself, then $\mathbb{P}(A \cap A) = \mathbb{P}(A)^2$. Therefore $\mathbb{P}(A) \in \{0, 1\}$ as before. The intuition is the same as the previous part.

(e) $\Omega = \{(1,1),(1,2),(2,1),(2,2)\}$ and $\mathcal{F} = 2^\Omega$.

Ω and A for all $A \in \mathcal{F}$	(16 pairs)
$\emptyset$ and A for all $A \in \mathcal{F} - \Omega$	(15 pairs)
$\{(1,0),(1,1)\}$ and $\{(0,0),(1,0)\}$	
$\{(1,0),(1,1)\}$ and $\{(0,1),(1,1)\}$	
$\{(0,0),(0,1)\}$ and $\{(0,0),(1,0)\}$	
$\{(0,0),(0,1)\}$ and $\{(0,1),(1,1)\}$	

(f) $\mathbb{P}(X_1 \leq 2, X_1 = X_2) = \mathbb{P}(X_1 = X_2 = 1) + \mathbb{P}(X_1 = X_2 = 2) = 2/9 = \mathbb{P}(X_1 \leq 2)\,\mathbb{P}(X_1 = X_2)$ because $\mathbb{P}(X_1 \leq 2) = 2/3$ and $\mathbb{P}(X_1 = X_2) = 1/3$.

(g) If A and B are independent, then $|A \cap B|/n = \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) = |A||B|/n^2$. Rearranging shows that $n|A \cap B| = |A||B|$. All steps can be reversed showing the reverse direction.

(h) Assume that n is prime. By the previous part, $n|A \cap B| = |A||B|$ must hold if A and B are independent of each other. If $|A \cap B| = 0$, the events will be trivial. Hence, assume $|A \cap B| > 0$. Since n is prime, it follows then that n must be either a factor of $|A|$ or a factor of $|B|$. Without loss of generality, assume that it is a factor of $|A|$. This implies $n \leq |A|$. But $|A| \leq n$ also holds, hence $|A| = n$, i.e., A is a trivial event.

(i) Let X_1 and X_2 be independent Rademacher random variables and $X_3 = X_1 X_2$. Clearly these random variables are not mutually independent since X_3 takes multiple values with nonzero probability and is fully determined by X_1 and X_2. And yet X_3 and X_i are independent for $i \in \{1, 2\}$, which ensures that pairwise independence holds.

(j) No. Let $\Omega = [6]$ and $\mathcal{F} = 2^\Omega$ and P be the uniform measure. Define events $A = \{1, 3, 4\}$ and $B = \{1, 3, 5\}$ and $C = \{3, 4, 5, 6\}$. Then A and B are clearly dependent and yet

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A \cap B \,|\, C)\,\mathbb{P}(C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)\ .$$

2.11

(a) $\sigma(X) = (\Omega, \emptyset)$ is trivial. Let Y be another random variable, then X and Y are independent if and only if for all $A \in \sigma(X)$ and $B \in \sigma(Y)$ it holds that $\mathbb{P}(A \cap B) = \mathbb{P}(B)$, which is trivial when $A \in \{\Omega, \emptyset\}$.

(b) Let A be an event with $\mathbb{P}(A) > 0$. Then $\mathbb{P}(X = x \,|\, A) = \mathbb{P}(X = x \cap A)\,/\mathbb{P}(A) = 1 = \mathbb{P}(X = x)$. Similarly, $\mathbb{P}(X \neq x \,|\, A) = 0 = \mathbb{P}(X \neq x)$. Therefore X is independent of all events, including those generated by Y.

(c) Suppose that A and B are independent. Then $\mathbb{P}(A^c \,|\, B) = 1 - \mathbb{P}(A \,|\, B) = 1 - \mathbb{P}(A) = \mathbb{P}(A^c)$. Therefore A^c and B are independent and by the same argument so are A^c and B^c as well as A and B^c. The 'if' direction follows by noting that $\sigma(X) = \{\Omega, A, A^c, \emptyset\}$ and $\sigma(Y) = \{\Omega, B, B^c, \emptyset\}$ and recalling that every event is independent of Ω or the empty set. For the 'only if' note that independence of X and Y means that any pair of events taken from $\sigma(X) \times \sigma(Y)$ are independent, which by the above includes the pair A, B.

(d) Let $(A_i)_i$ be a countable family of events and $X_i(\omega) = \mathbb{I}\{\omega \in A_i\}$ be the indicator of the ith event. When the random variables/events are pairwise independent, then the above argument goes through unchanged for each pair. In the case of mutual independence the 'only if' is again the same. For the 'if', suppose that (A_i) are mutually independent. Therefore for any finite subset $K \subset \mathbb{N}$ we have

$$\mathbb{P}\left(\bigcap_{i\in K} A_i\right) = \prod_{i\in K} \mathbb{P}(A_i)$$

The same argument as the previous part shows that for any disjoint finite sets $K, J \subset \mathbb{N}$ we have

$$\mathbb{P}\left(\bigcup_{i\in K} A_i \cup \bigcup_{i\in J} A_i^c\right) = \prod_{i\in K} \mathbb{P}(A_i) \prod_{i\in J} \mathbb{P}(A_i^c) \ .$$

Therefore for any finite set $K \subset \mathbb{N}$ and $(V_i)_{i\in K}$ with $V_i \in \sigma(X_i) = \{\Omega, \emptyset, A_i, A_i^c\}$ it holds that

$$\mathbb{P}\left(\bigcap_{i\in K} V_i\right) = \prod_{i\in K} \mathbb{P}(V_i) \ ,$$

which completes the proof that $(X_i)_i$ are mutually independent.

2.12

(a) Let $A \subset \mathbb{R}$ be an open set. By definition, since f is continuous it holds that $f^{-1}(A)$ is open. But the Borel σ-algebra is generated by all open sets and so $f^{-1}(A) \in \mathfrak{B}(\mathbb{R})$ as required.

(b) Since $|\cdot| : \mathbb{R} \to \mathbb{R}$ is continuous and by definition a random variable X on measurable space $(\Omega, \mathcal{F})$ is $\mathcal{F}/\mathfrak{B}(\mathbb{R})$-measurable it follows by the previous part that $|X|$ is $\mathcal{F}/\mathfrak{B}(\mathbb{R})$-measurable and therefore a random variable.

(c) Recall that $(X)^+ = \max\{0, X\}$ and $(X)^- = -\min\{0, X\}$. Therefore $(|X|)^+ = |X| = (X)^+ + (X)^-$ and $(|X|)^- = 0$. Recall that $\mathbb{E}[X] = \mathbb{E}[(X)^+] - \mathbb{E}[(X)^-]$ exists if an only if both expectations are defined. Therefore if X is integrable, then $|X|$ is integrable. Now suppose that $|X|$ is integrable, then X is integrable by the dominated convergence theorem.

2.14 Assume without (much) loss of generality that $X_i \geq 0$ for all i. The general case follows by

considering positive and negative parts, as usual. First we claim that for any n it holds that

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]\,.$$

To show this, note the definition that

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] &= \sup\left\{\int_\Omega h\, d\mathbb{P} : h \text{ is simple and } 0 \le h \le \sum_{i=1}^{n} X_i\right\} \\ &= \sum_{i=1}^{n} \sup\left\{\int_\Omega h\, d\mathbb{P} : h \text{ is simple and } 0 \le h \le X_i\right\}.\end{aligned}$$

Next let $S_n = \sum_{i=1}^{n} X_i$ and note that by the monotone convergence theorem we have $\lim_{n\to\infty} \mathbb{E}[S_n] = \mathbb{E}[X]$, which means that

$$\mathbb{E}\left[\sum_{i=1}^{\infty} X_i\right] = \lim_{n\to\infty} \mathbb{E}[S_n] = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}[X_i]\,.$$

2.15 Suppose that $X(\omega) = \sum_{i=1}^{n} \alpha_i \mathbb{I}\{\omega \in A_i\}$ is simple and $c > 0$. Then cX is also simple and

$$\mathbb{E}[cX] = \sum_{i=1}^{n} c\alpha_i \mathbb{I}\{\omega \in A_i\} = c\sum_{i=1}^{n} \alpha_i \mathbb{I}\{\omega \in A_i\} = c\mathbb{E}[X]\,.$$

Now suppose that X is positive (but maybe not simple) and $c > 0$, then cX is also positive and

$$\begin{aligned}\mathbb{E}[cX] &= \sup\{\mathbb{E}[h] : h \text{ is simple and } h \le cX\} \\ &= \sup\{\mathbb{E}[ch] : h \text{ is simple and } h \le X\} \\ &= \sup\{c\mathbb{E}[h] : h \text{ is simple and } h \le X\} \\ &= c\mathbb{E}[X]\,.\end{aligned}$$

Finally for arbitrary random variables and $c > 0$ we have

$$\mathbb{E}[cX] = \mathbb{E}[(cX)^+] - \mathbb{E}[(cX)^-] = c\mathbb{E}[(X)^+] - c\mathbb{E}[(X)^-] = c\mathbb{E}[X]\,.$$

For negative c simply note that $(cX)^+ = -c(X)^-$ and $(cX)^- = -c(X)^+$ and repeat the above argument.

2.16 Suppose $X = \sum_{i=1}^N \alpha_i \mathbb{I}\{A_i\}$ and $Y = \sum_{i=1}^N \beta_i \mathbb{I}\{B_i\}$ are simple functions. Then

$$\begin{aligned}
\mathbb{E}[XY] &= \mathbb{E}\left[\sum_{i=1}^N \alpha_i \mathbb{I}\{A_i\} \sum_{i=1}^N \beta_i \mathbb{I}\{B_i\}\right] \\
&= \sum_{i=1}^N \sum_{j=1}^N \alpha_i \beta_j \mathbb{P}(A_i \cap B_j) \\
&= \sum_{i=1}^N \sum_{j=1}^N \alpha_i \beta_j \mathbb{P}(A_i)\,\mathbb{P}(A_j) \\
&= \mathbb{E}[X]\mathbb{E}[Y]\,.
\end{aligned}$$

Now suppose that X and Y are arbitrary nonnegative independent random variables. Then

$$\begin{aligned}
&\mathbb{E}[XY] = \sup\{\mathbb{E}[h] : h \text{ is simple and } h \le XY\} \\
&= \sup\{\mathbb{E}[hg] : h \in \sigma(X), g \in \sigma(Y) \text{ are simple and } h \le X, g \le Y\} \\
&= \sup\{\mathbb{E}[h]\mathbb{E}[g] : h \in \sigma(X), g \in \sigma(Y) \text{ are simple and } h \le X, g \le Y\} \\
&= \sup\{\mathbb{E}[h] : h \in \sigma(X) \text{ is simple and } h \le X\} \sup\{\mathbb{E}[h] : h \in \sigma(Y) \text{ is simple and } h \le Y\} \\
&= \mathbb{E}[X]\mathbb{E}[Y]\,.
\end{aligned}$$

Finally, for arbitrary random variables we have via the previous display and the linearity of expectation that

$$\begin{aligned}
\mathbb{E}[XY] &= \mathbb{E}[((X)^+ - (X)^-)((Y)^+ - (Y)^-)] \\
&= \mathbb{E}[(X)^+(Y)^+] - \mathbb{E}[(X)^+(Y)^-] - \mathbb{E}[(X)^-(Y)^+] + \mathbb{E}[(X)^-(Y)^-] \\
&= \mathbb{E}[(X)^+]\mathbb{E}[(Y)^+] - \mathbb{E}[(X)^+]\mathbb{E}[(Y)^-] - \mathbb{E}[(X)^-]\mathbb{E}[(Y)^+] + \mathbb{E}[(X)^-]\mathbb{E}[(Y)^-] \\
&= \mathbb{E}[(X)^+ - (X)^-]\mathbb{E}[(Y)^+ - (Y)^-]\,.
\end{aligned}$$

2.17 Let X be a standard Rademacher random variable and $Y = X$. Then $\mathbb{E}[X]\mathbb{E}[Y] = 0$ and $\mathbb{E}[XY] = 1$.

2.18 Using the fact that $\int_0^a 1dx = a$ for $a \ge 0$ and the nonnegativity of X we have

$$X(\omega) = \int_0^\infty \mathbb{I}\{[0, X(\omega)]\}(x)dx\,.$$

Then by Fubini's theorem,

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\int_0^\infty \mathbb{I}\{[0, X(\omega)]\}(x)dx\right] \\
&= \int_0^\infty \mathbb{E}[\mathbb{I}\{[0, X(\omega)]\}(x)]dx \\
&= \int_0^\infty \mathbb{P}(X(\omega) \ge x)\,dx\,.
\end{aligned}$$

Chapter 3 Stochastic Processes and Markov Chains

3.1

(a) We have $F_1(x) = \mathbb{I}\{x \in [1/2, 1]\}$ and $F_2(x) = \mathbb{I}\{x \in [1/4, 2/4) \cup [3/4, 4/4]\}$. More generally, $F_t(x) = \mathbb{I}_{U_t}(x)$ where

$$U_t = \{1\} \cup \bigcup_{1 \le s \le 2^{t-1}} [(2s-1)/2^t, 2s/2^t)\,.$$

Since $U_t \in \mathfrak{B}([0,1])$, F_t are random variables (see Fig. 3.1).

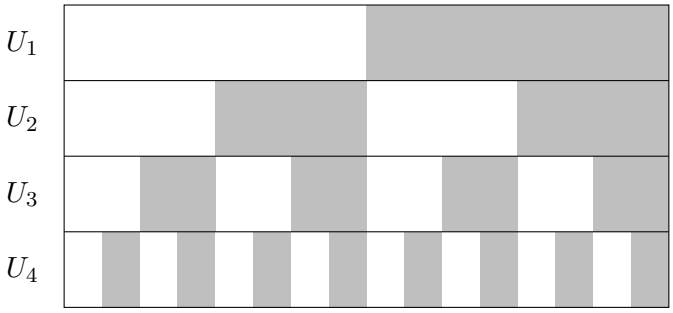


Figure 3.1: Illustration of events $(U_t)_{t=1}^{\infty}$.

(b) We have $\mathbb{P}(U_t) = \lambda(U_t) = \sum_{s=1}^{2^{t-1}} (1/2^t) = 1/2$.

(c) Given an index set $K \subset \mathbb{N}^+$ we need to show that $\{F_k : k \in K\}$ are independent. Or equivalently, that

$$\mathbb{P}\left(\bigcap_{k \in K} U_k\right) = \prod_{k \in K} \mathbb{P}\left(U_k\right) = 2^{-|K|}\,. \tag{3.1}$$

Let $k = \max K$. Then

$$\lambda\left(U_k \cap \bigcup_{j \in K \setminus \{k\}} U_j\right) = \frac{1}{2}\lambda\left(\bigcup_{j \in K \setminus \{k\}}\right)\,.$$

Then Eq. (3.1) follows by induction.

(d) It follows directly from the definition of independence that any subsequence of an independent sequence is also an independent sequence. That $\mathbb{P}\left(X_{m,t} = 0\right) = \mathbb{P}\left(X_{m,t} = 1\right) = 1/2$ follows from Part (b).

(e) By the previous parts $X_t = \sum_{t=1}^{\infty} X_{m,t} 2^{-t}$ is a weighted sum of an independent sequence of uniform Bernoully random variables. Therefore X_t has the same law as $Y = \sum_{t=1}^{\infty} F_t 2^{-t}$. But $Y(x) = x$ is the identity. Hence Y is uniformly distributed and so too is X_t.

(f) This follows from the definition of $(X_{m,t})_{t=1}^{\infty}$ as disjoint subsets of independent random variables $(F_t)_{t=1}^{\infty}$ and the 'grouping' result that whenever $(\mathcal{F}_t)_{t\in\mathcal{T}}$ is a collection of independent σ-algebras and $\mathcal{T}_1, \mathcal{T}_2$ are disjoint subsets of $\mathcal{T}$, then $\sigma(\cup_{t\in\mathcal{T}_1}\mathcal{F}_t)$ and $\sigma(\cup_{t\in\mathcal{T}_2}\mathcal{F}_t)$ are independent [Kallenberg, 2002, Corollary 3.7]. This latter result is a good exercise. Use a monotone class argument.

3.5 Let $A \in \mathcal{G}$ and suppose that $X(\omega) = \mathbb{I}_A(\omega)$. Then $\int_{\mathcal{X}} X(x)K(\omega, dx) = K(\omega, A)$, which is $\mathcal{F}$-measurable by the definition of a probability kernel. The result extends to simple functions by linearity. For nonnegative X let $X_n \uparrow X$ be a monotone increasing sequence of simple functions converging point-wise to X [Kallenberg, 2002, Lemma 1.11]. Then $U_n(\omega) = \int_{\mathcal{X}} X_n(x)K(\omega, dx)$ is $\mathcal{F}$-measurable. Monotone convergence ensures that $\lim_{n\to\infty} U_n(\omega) = \int_{\mathcal{X}} \lim_{n\to\infty} X_n(x)K(\omega, dx) = \int_{\mathcal{X}} X(x)K(\omega, dx) = U(\omega)$. Hence $\lim_{n\to\infty} U_n(\omega) = U(\omega)$ and a point-wise convergent sequence of measurable functions is measurable, it follows that U is $\mathcal{F}$-measurable. The result for arbitrary X follows by decomposing into positive and negative parts.

3.8 Let $(X_t)_{t=0}^n$ be $\mathbb{F} = (\mathcal{F}_t)_{t=1}^n$-adapted and $\tau = \min\{t : X_t \geq \varepsilon\}$. By the submartingale property, $\mathbb{E}[X_n \,|\, \mathcal{F}_t] \geq X_t$. Therefore

$$\mathbb{I}\{\tau = t\} X_t \leq \mathbb{I}\{\tau = t\} \mathbb{E}[X_n \,|\, \mathcal{F}_t] = \mathbb{E}[\mathbb{I}\{\tau = t\} X_n \,|\, \mathcal{F}_t] .$$

Therefore $\mathbb{E}[X_t \mathbb{I}\{\tau = t\}] \leq \mathbb{E}[\mathbb{I}\{\tau = t\} X_n]$.

$$\begin{aligned}
\mathbb{P}\left(\max_{t\in\{0,1,\dots,n\}} X_t \geq \varepsilon\right) &= \sum_{t=0}^{n} \mathbb{P}(\tau = t) \\
&= \sum_{t=0}^{n} \mathbb{P}(X_t \mathbb{I}\{\tau = t\} \geq \varepsilon) \\
&\leq \frac{1}{\varepsilon} \sum_{t=0}^{n} \mathbb{E}[X_t \mathbb{I}\{\tau = t\}] \\
&\leq \frac{1}{\varepsilon} \sum_{t=0}^{n} \mathbb{E}[X_n \mathbb{I}\{\tau = t\}] \\
&\leq \frac{\mathbb{E}[X_n]}{\varepsilon} .
\end{aligned}$$

Chapter 4 Stochastic Bandits

4.8

(a) The statement is true. Let i be a suboptimal arm. By Lemma 4.5 we have

$$0 = \lim_{n\to\infty} \frac{R_n(\pi,\nu)}{n} = \limsup_{n\to\infty} \sum_{i=1}^{k} \frac{\mathbb{E}[T_i(n)]\Delta_i}{n} \geq \limsup_{n\to\infty} \frac{\mathbb{E}[T_i(n)]}{n}\Delta_i\,.$$

Hence $\limsup_{n\to\infty} \mathbb{E}[T_i(n)]/n \leq 0 \leq \liminf_{n\to\infty} \mathbb{E}[T_i(n)]/n$ and so $\lim_{n\to\infty} \mathbb{E}[T_i(n)]/n = 0$ for suboptimal arms i. Since $\sum_{i=1}^{k} \mathbb{E}[T_i(n)]/n = 1$ it follows that $\lim_{n\to\infty} \sum_{i:\Delta_i=0} \mathbb{E}[T_i]/n = 1$.

(b) The statement is false. Consider a two-armed bandit for which the second arm is suboptimal and an algorithm that chooses the second arm in rounds $t \in \{1, 2, 4, 8, 16, \ldots\}$.

Chapter 5 Concentration of Measure

5.15 We use the Cramer-Chernoff method:

$$\begin{aligned}\mathbb{P}\left(\sum_{t=1}^{n}(X_t - \mu_t - \alpha_t) \geq \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)\right) &= \mathbb{P}\left(\exp\left(\lambda\sum_{t=1}^{n}(X_t - \mu_t - \alpha_t)\right) \geq \frac{1}{\delta}\right) \\ &\leq \delta\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{n}(X_t - \mu_t - \alpha_t)\right)\right].\end{aligned}$$

All that remains is to show that the term inside the expectation is a supermartingale. Using the fact that $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ and $1 + x \leq \exp(x)$ for all $x \in \mathbb{R}$ we have

$$\begin{aligned}\mathbb{E}_{t-1}\left[\exp\left(\lambda(X_t - \mu_t - \alpha_t)\right)\right] &= \exp(-\lambda\alpha_t)\mathbb{E}_{t-1}\left[\exp(\lambda(X_t - \mu_t))\right] \\ &\leq \exp(-\lambda\alpha_t)\left(1 + \lambda^2\mathbb{E}_{t-1}[(X_t - \mu_t)^2]\right) \\ &\leq \exp\left(\lambda\left(\lambda\mathbb{E}_{t-1}[(X_t - \mu_t)^2] - \alpha_t\right)\right) \leq 1\,.\end{aligned}$$

Therefore $\exp(\lambda\sum_{t=1}^{n}(X_t - \mu_t - \alpha_t))$ is a supermartingale, which completes the proof.

5.16 By assumption $\mathbb{P}(X_t \leq x) \leq x$, which means that for $\lambda < 1$,

$$\begin{aligned}\mathbb{E}[\exp(\lambda\log(1/X_t))] &= \int_0^\infty \mathbb{P}\left(\exp(\lambda\log(1/X_t)) \geq x\right)dx \\ &= 1 + \int_1^\infty \mathbb{P}\left(X_t \leq x^{-1/\lambda}\right)dx \leq 1 + \int_1^\infty x^{-1/\lambda}dx = \frac{1}{1-\lambda}\,.\end{aligned}$$

Applying the Cramer-Chernoff method,

$$\begin{aligned}\mathbb{P}\left(\sum_{t=1}^{n}\log(1/X_t) \geq \varepsilon\right) &= \mathbb{P}\left(\exp\left(\lambda\sum_{t=1}^{n}\log(1/X_i)\right) \geq \exp(\lambda\varepsilon)\right) \\ &\leq \exp(-\lambda\varepsilon)\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{n}\log(1/X_t)\right)\right] \leq \left(\frac{1}{1-\lambda}\right)^n\exp(-\lambda\varepsilon)\,.\end{aligned}$$

Choosing $\lambda = (\varepsilon - n)/\varepsilon$ completes the claim.

5.18 Let $\lambda > 0$. Then

$$\exp(\lambda \mathbb{E}[Z]) \leq \mathbb{E}[\exp(\lambda Z)] \leq \sum_{t=1}^{n} \mathbb{E}[\exp(\lambda X_t)] \leq n \exp(\lambda^2 \sigma^2 / 2)\,.$$

Rearranging shows that

$$\mathbb{E}[Z] \leq \frac{\log(n)}{\lambda} + \frac{\lambda \sigma^2}{2}\,.$$

Choosing $\lambda = \frac{1}{\sigma}\sqrt{2\log(n)}$ shows that $\mathbb{E}[Z] \leq \sqrt{2\sigma^2 \log(n)}$.

5.19 Let $\mathcal{P}$ be the set of measures on $([0,1], \mathfrak{B}([0,1]))$ and for $q \in \mathcal{P}$ let μ_q be its mean. The theorem will be established by induction over n. The claim is immediate when $x > n$ or $n = 1$. Assume that $n \geq 2$ and $x \in (1, n]$ and the theorem holds for $n-1$. Then

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{n} \mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}]\right) &= \mathbb{E}\left[\mathbb{P}\left(\sum_{t=2}^{n} \mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}] \geq x - \mathbb{E}[X_1 \,|\, \mathcal{F}_0] \,\middle|\, \mathcal{F}_0\right)\right] \\
&\leq \mathbb{E}\left[f_{n-1}\left(\frac{x - \mathbb{E}[X_1 \,|\, \mathcal{F}_0]}{1 - X_1}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[f_{n-1}\left(\frac{x - \mathbb{E}[X_1 \,|\, \mathcal{F}_0]}{1 - X_1}\right) \,\middle|\, \mathcal{F}_0\right]\right] \\
&\leq \sup_{q \in \mathcal{P}} \int_0^1 f_{n-1}\left(\frac{x - \mu_q}{1 - y}\right) dq(y)\,,
\end{aligned}$$

where the first inequality follows from the inductive hypothesis and the fact that $\sum_{t=2}^{n} X_t/(1-X_1) \leq 1$ almost surely. The result is completed by proving that for all $q \in \mathcal{P}$,

$$F_n(q) \doteq \int_0^1 f_{n-1}\left(\frac{x - \mu_q}{1 - y}\right) dq(y) \leq f_n(x)\,. \tag{5.1}$$

Let $q \in \mathcal{P}$ have mean μ and $y_0 = \max(0, 1 - x + \mu)$. In Lemma 5.1 below it is shown that

$$f_{n-1}\left(\frac{x-\mu}{1-y}\right) \leq \frac{1-y}{1-y_0} f_{n-1}\left(\frac{x-\mu}{1-y_0}\right)\,,$$

which after integrating implies that

$$F_n(q) \leq \frac{1-\mu}{1-y_0} f_{n-1}\left(\frac{x-\mu}{1-y_0}\right)\,.$$

Considering two cases. First, when $y_0 = 0$ the display shows that $F_n(q) \leq (1-\mu) f_{n-1}(x-\mu)$. On the other hand, if $y_0 > 0$ then $x - 1 < \mu \leq 1$ and $F_n(q) \leq (1-\mu)/(x-\mu) \leq (1-(x-1)) f_{n-1}(x-(x-1))$.

Combining the two cases we have

$$F_n(q) \le \sup_{\mu \in [0,1]} (1-\mu) f_{n-1}(1-\mu) = f_n(x)\,.$$

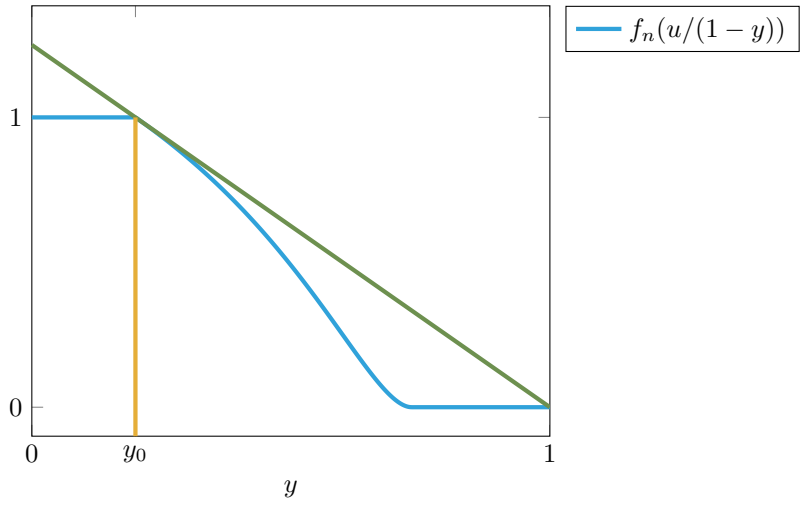
Lemma 5.1. *Suppose that $n \ge 1$ and $u \in (0, n]$ and $y_0 = \max(0, 1-u)$. Then*

$$f_n\left(\frac{u}{1-y}\right) \le \frac{1-y}{1-y_0} f_{n-1}\left(\frac{u}{1-y_0}\right) \qquad \text{for all } y \in [0,1]\,.$$

Proof. The lemma is equivalent to the claim that the line connecting $(y_0,\, f_n(u/(1-y_0)))$ and $(1,0)$ lies above $f_n(u/(1-y))$ for all $y \in [0,1]$ (see figure below). This is immediate for $n = 1$ when $f_n(u/(1-y)) = \mathbb{I}\{y \le 1-u\}$. For larger n basic calculus shows that $f_n(u/(1-y))$ is concave as a function of y on $[1-u, 1-u/n]$ and

$$\frac{\partial}{\partial y} f_n(u/(1-y))\Big|_{y=1-u} = -1/u\,.$$

Since $f_n(1) = 1$ this means that the line connecting $(1-u, 1)$ and $(1,0)$ lies above $f_n(u/(1-y))$. This completes the proof when $y_0 = 1-u$. Otherwise $y_0 \in [1-u, 1-u/n]$ and the result follows by concavity of $f_n(u/(1-y))$ on this interval. □



Chapter 6 The Explore-then-Commit Algorithm

Chapter 7 The Upper Confidence Bound Algorithm

7.1

(a) We have

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{2\log(1/\delta)}{T}}\right) &= \sum_{n=1}^{\infty} \mathbb{E}\left[\mathbb{I}\left\{T = n\right\} \mathbb{I}\left\{\sum_{t=1}^{n}(X_t - \mu) \geq \sqrt{2n\log(1/\delta)}\right\}\right] \\
&= \sum_{n=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left\{T = n\right\} \mathbb{I}\left\{\sum_{t=1}^{n}(X_t - \mu) \geq \sqrt{2n\log(1/\delta)}\right\} \middle| T\right]\right] \\
&= \sum_{n=1}^{\infty} \mathbb{E}\left[\mathbb{I}\left\{T = n\right\} \mathbb{E}\left[\mathbb{I}\left\{\sum_{t=1}^{n}(X_t - \mu) \geq \sqrt{2n\log(1/\delta)}\right\} \middle| T\right]\right] \\
&\leq \sum_{n=1}^{\infty} \mathbb{E}\left[\mathbb{I}\left\{T = n\right\} \delta\right] \\
&= \delta\,.
\end{aligned}$$

(b) Let $T = \min\left\{n : \sum_{t=1}^{n}(X_t - \mu) \geq \sqrt{2n\log(1/\delta)}\right\}$. By the law of the iterated logarithm, $T < \infty$ almost surely. The result follows.

(c) Note that

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{2\log(T(T+1)/\delta)}{T}}\right) &\leq \mathbb{P}\left(\text{exists } n : \sum_{t=1}^{n}(X_t - \mu) \geq \sqrt{2n\log(n(n+1)/\delta)}\right) \\
&\leq \sum_{n=1}^{\infty} \frac{\delta}{n(n+1)} \\
&= \delta\,.
\end{aligned}$$

Chapter 8 The Upper Confidence Bound Algorithm: Asymptotic Optimality

8.1 Following the hint, $F \leq \exp(-a)/(1-\exp(-a))$ where $a = \varepsilon^2/2$. Reordering $\exp(-a)/(1-\exp(-a)) \leq 1/a$ gives $1 + a \leq \exp(a)$ which is well known (and easy to prove). Then

$$\begin{aligned}
\sum_{t=1}^{n} \frac{1}{f(t)} &\leq \sum_{t=1}^{20} \frac{1}{f(t)} + \int_{20}^{\infty} \frac{dt}{f(t)} \leq \sum_{t=1}^{20} \frac{1}{f(t)} + \int_{20}^{\infty} \frac{dt}{t\log(t)^2} \\
&= \sum_{t=1}^{20} \frac{1}{f(t)} + \frac{1}{\log(20)} \leq \frac{5}{2}\,.
\end{aligned}$$

Chapter 9 The Upper Confidence Bound Algorithm: Minimax Optimality

9.1 Clearly (M_t) is $\mathbb{F}$-adapted. Then by Jensen's inequality and convexity of the exponential function,

$$\begin{aligned}\mathbb{E}[M_t \,|\, \mathcal{F}_{t-1}] &= \exp\left(\lambda \sum_{s=1}^{t} X_s\right) \mathbb{E}[\exp(\lambda X_t) \,|\, \mathcal{F}_{t-1}] \\ &\geq \exp\left(\lambda \sum_{s=1}^{t} X_s\right) \exp(\lambda \mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}]) \\ &= \exp\left(\lambda \sum_{s=1}^{t-1} X_s\right) \quad \text{a.s.}\end{aligned}$$

Hence M_t is a $\mathbb{F}$-submartingale.

9.4

(c) We provide the intuitive explanation only. For each $t \geq 0$ let Q_t be the measure on $\mathbb{R}$ given by

$$Q(A, t) = \mathbb{P}\left(\tau \geq t \text{ and } B_t \in A\right) .$$

Then let $q(x, t) = dQ(\cdot, t)/d\lambda(x)$ with λ the Lebesgue measure. For small dt we should expect

$$q(x, t + dt) \approx \frac{1}{\sqrt{2\pi dt}} \int_{\mathbb{R}} q(y, t) \exp(-(x - y)^2/(2dt)) dy .$$

Approximating $q(y, t) \approx q(x, t) + \partial_x q(x, t)(y - x) + \frac{1}{2}\partial_x^2 q(x, t)(y - x)^2$ by its second order Taylor series suggests that

$$q(x, t + dt) = q(x, t) + \frac{\partial_x^2 q(x, t) dt}{2} + o(dt) .$$

From this we conclude that $\partial_t q(x, t) = \frac{\partial_x^2 q(x,t)}{2}$. In other words $q(x, t)$ satisfies the heat equation. Now for $x = at + b$ we should also have $q(x, t) = 0$. So $q(x, t)$ satisfies the heat equation and boundary conditions that $q(at + b, t) = 0$ and $q(\cdot, t)$ converges to a Dirac as t tends to zero. Since $u(x, t)$ satisfies these conditions and by a uniqueness of solutions it must holds that $u = q$. Finally note that by definition

$$\mathbb{P}\left(\tau_g \geq t\right) = \int_{-\infty}^{g(t)} u(x, t) dx .$$

(d) First we evaluate the integral in (c).

$$\begin{aligned}\mathbb{P}\left(\tau_g \geq t\right) &= \int_{-\infty}^{g(t)} u(x,t)dx \\ &= \int_{-\infty}^{at+b} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right) - \frac{1}{\sqrt{2\pi t}} \exp\left(-2ab - \frac{(x-2b)^2}{2t}\right)\end{aligned}$$

Then

$$v(t) = -\frac{d}{dt}\mathbb{P}\left(\tau_g \geq t\right) = \frac{b\exp(-g(t)^2/(2t))}{\sqrt{2\pi t^3}}\,.$$

(e) Let $t > 0$ and $g(s) = f(t) + f'(t)(s-t)$ be the tangent to f at t. Consider a sample path (B_s) from a Brownian motion.

(f) Let (B_t) be a standard Brownian motion. Then $X_t = B_t - B_{t-1}$ is a sequence of independent standard Gaussian random variables. Then

$$\begin{aligned}\mathbb{P}\left(\text{exists } t \leq n : \sum_{s=1}^{t} X_s \geq f(t)\right) &\leq \mathbb{P}\left(\text{exists } t \leq n : B_t \geq f(t)\right) \\ &\leq \int_0^n \frac{\lambda_t(0)}{\sqrt{2\pi t^3}} \exp(-f(t)^2/(2t))dt\,.\end{aligned}$$

(g) By the previous part,

$$\begin{aligned}\mathbb{P}\left(\text{exists } t \leq \infty : \sum_{s=1}^{t} X_s \geq f(t)\right) &\leq \int_0^\infty \frac{\lambda_t(0)}{\sqrt{2\pi t^3}} \exp(-f(t)^2/(2t))dt \\ &\leq \int_0^\infty \frac{\sqrt{\log(h(1/\delta t)}}{h(1/\delta t)\sqrt{\pi}} \exp(-t\Delta^2/2)dt \\ &\leq \int_0^\infty \frac{c\delta}{\sqrt{\pi}} \exp(-t\Delta^2/2)dt \\ &= \frac{2c\delta}{\sqrt{\pi}\Delta^2}\,.\end{aligned}$$

(i) Consider the policy that plays each arm once and subsequently chooses

$$A_t = \text{argmax}_{i\in[k]}\, \hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)} \log\left(h\left(\frac{n}{T_i(t-1)k}\right)\right)}\,.$$

The requested experimental validation is omitted from this solution.

(j) The proof more-or-less follows the proof of Theorem 9.1, but uses the new concentration inequalities. Define random variable

$$\Delta = \min\left\{\Delta \geq 0 : \hat{\mu}_{1s} \geq \mu_1 - \Delta \text{ for all } s \in [n]\right\}\,.$$

Let $\varepsilon > 0$. By Part (g)

$$\mathbb{P}\left(\Delta \geq \varepsilon\right) = O\left(\frac{1}{n}\right) .$$

For suboptimal arm i define

$$\kappa_i = \sum_{s=1}^{n} \mathbb{I}\left\{\hat{\mu}_{is} + \sqrt{\frac{2}{s}\log\left(h\left(\frac{n}{ks}\right)\right)} \leq \mu_1 - \varepsilon\right\} .$$

By the definition of the algorithm and κ_i and Δ,

$$T_i(n) \leq \kappa_i + n\mathbb{I}\left\{\Delta \geq \varepsilon\right\} .$$

The expectation of κ_i is bounded using the same technique as in Theorem 9.1, but the tighter confidence bound leads to an improved bound.

$$\begin{aligned}
\mathbb{E}[\kappa_i] &\leq \frac{1}{\Delta_i^2} + \sum_{s=1}^{n} \mathbb{P}\left(\hat{\mu}_{is} + \sqrt{\frac{2}{s}\log\left(h\left(\frac{n\Delta^2}{k}\right)\right)} \leq \mu_1 - \varepsilon\right) \\
&= \frac{2\log(n)}{(\Delta_i - \varepsilon)^2} + o(\log(n)) .
\end{aligned}$$

Combining the two parts shows that for all $\varepsilon > 0$,

$$\mathbb{E}[T_i(n)] \leq \frac{2\log(n)}{(\Delta_i - \varepsilon)^2} + o(\log(n)) .$$

The result follows from the fundamental regret decomposition lemma (Lemma 4.5) and by taking the limit as n tends to infinity and ε tends to zero at appropriate rates.

Chapter 10 The Upper Confidence Bound Algorithm: Bernoulli Noise

10.1 Let g be as in the hint. We have

$$g'(x) = x\left(\frac{1}{(p+x)(1-(p+x))} - 4\right) .$$

Clearly, $g'(0) = 0$. Further, since $q(1-q) \leq 1/4$ for any $q \in [0,1]$, $g'(x) \geq 0$ for $x > 0$ and $g'(x) \leq 0$ for $x < 0$. Hence, g is increasing for positive x and decreasing for negative x. Thus, $x = 0$ is a minimizer of g. Here, $g(0) = 0$, and so $g(x) \geq 0$ over $[-p, 1-p]$.

10.3 We have $g(\lambda,\mu) = -\lambda\mu + \log(1 + \mu(e^\lambda - 1))$. Taking derivatives, $\frac{d}{d\mu}g(\lambda,\mu) = -\lambda + \frac{e^\lambda - 1}{1+\mu(e^\lambda-1)}$ and $\frac{d^2}{d\mu^2}g(\lambda,\mu) = -\frac{(e^\lambda-1)^2}{(1+\mu(e^\lambda-1))^2} \leq 0$, showing that $g(\lambda,\cdot)$ is concave as suggested in the hint. Now,

let $S_p = \sum_{t=1}^{p}(X_t - \mu_t)$, $p \in [n]$ and let $S_0 = 0$. Then for $p \in [n]$,

$$\mathbb{E}\left[\exp(\lambda S_p)\right] = \mathbb{E}\left[\exp(\lambda S_{p-1})\mathbb{E}\left[\exp(\lambda(X_p - \mu_p)) \,|\, \mathcal{F}_{p-1}\right]\right] ,$$

and, by note Item 2, $\mathbb{E}\left[\exp(\lambda(X_p - \mu_p)) \,|\, \mathcal{F}_{p-1}\right] \leq \exp(g(\lambda, \mu_p))$. Hence, using that μ_n is not random,

$$\mathbb{E}\left[\exp(\lambda S_p)\right] \leq \mathbb{E}\left[\exp(\lambda S_{p-1})\right] \exp(g(\lambda, \mu_p)) .$$

Chaining this inequalities, using that $S_0 = 0$ together with that $g(\lambda, \cdot)$ is concave, we get

$$\mathbb{E}\left[\exp(\lambda S_n)\right] \leq \left(\exp\left(\tfrac{1}{n}\sum_{t=1}^{n} g(\lambda, \mu_t)\right)\right)^n \leq \exp\left(n g(\lambda, \mu)\right) .$$

Thus,

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu} - \mu \geq \varepsilon\right) &= \mathbb{P}\left(\exp(\lambda S_n) \geq \exp(\lambda n \varepsilon)\right) \\
&\leq \mathbb{E}\left[\exp(\lambda S_n)\right] \exp(-\lambda n \varepsilon) \\
&\leq \left(\mu \exp(\lambda(1 - \mu - \varepsilon)) + (1 - \mu)\exp(-\lambda(\mu + \varepsilon))\right)^n .
\end{aligned}$$

From this point, repeat the proof of Lemma 10.3 word by word.

10.4 Abbreviate $p_\theta(x) = \frac{dP_\theta}{dh}(x)$.

(a) Clearly $p_\theta(x) \geq 0$. By definition, for $B \in \mathfrak{B}(\mathbb{R})$, $P_\theta(B) = \int_B p_\theta(x) dh(x)$. Hence $P_\theta(B) \geq 0$. Furthermore,

$$P_\theta(\mathbb{R}) = \int_{\mathbb{R}} \exp(\theta T(x) - A(\theta)) dh(x) = \exp(-A(\theta)) \int_{\mathbb{R}} \exp(\theta T(x)) dh(x) = 1 .$$

Additivity is immediate since $\int_B f dh + \int_C f dh = \int_{B \cup C} f dh$ for disjoint B, C.

(b) Using the chain rule and passing the derivative under the integral yields the result:

$$\begin{aligned}
A'(\theta) &= \frac{\frac{d}{d\theta}\int_{\mathbb{R}} \exp(\theta T(x)) dh(x)}{\int_{\mathbb{R}} \exp(\theta T(x)) dh(x)} \\
&= \frac{\int_{\mathbb{R}} T(x) \exp(\theta T(x)) dh(x)}{\int_{\mathbb{R}} \exp(\theta T(x)) dh(x)} \\
&= \int_{\mathbb{R}} T(x) \exp(\theta T(x) - A(x)) dh(x) \\
&= \int_{\mathbb{R}} T(x) p_\theta(x) dh(x) \\
&= \mathbb{E}_\theta[T] .
\end{aligned}$$

In order to justify the exchange of integral and derivative use the identity that for all sufficiently

small $\varepsilon > 0$ and all $a > 0$,

$$a \leq \frac{\exp(a\varepsilon) + \exp(-a\varepsilon)}{\varepsilon}\,.$$

Hence for $\theta \in \operatorname{int}(\operatorname{dom}(A))$ there exists a neighborhood N of θ such that for all $\psi \in N$,

$$|T(x)| \exp(\psi T(x)) \leq \phi(x) = \frac{\exp((\theta+\varepsilon)T(x)) + \exp((\theta-\varepsilon)T(x))}{\varepsilon}\,.$$

Since $\phi(x)$ is integrable for sufficiently small ε it follows by the dominated convergence theorem that the derivative and integral can be exchanged.

(c) When $X \sim P_\theta$ we have

$$\begin{aligned}
\mathbb{E}[\exp(\lambda T(X))] &= \int_{\mathbb{R}} \exp(\lambda T(x)) \frac{dP_\theta}{dh} dh(x) \\
&= \int_{\mathbb{R}} \exp(\lambda T(x)) \exp(\theta T(x) - A(\theta)) dh(x) \\
&= \exp(-A(\theta)) \int_{\mathbb{R}} \exp((\theta+\lambda)T(x)) dh(x) \\
&= \exp(-A(\theta)) \exp(A(\theta+\lambda)) \\
&= \exp(A(\theta+\lambda) - A(\theta))\,.
\end{aligned}$$

(d) This is another straightforward calculation:

$$\begin{aligned}
d(\theta, \theta') &= \int_{\mathbb{R}} (\theta T(x) - A(\theta) - \theta' T(x) + A(\theta')) \exp(\theta T(x) - A(\theta)) dh(x) \\
&= A(\theta') - A(\theta) + (\theta - \theta') \int_{\mathbb{R}} T(x) \exp(\theta T(x) - A(\theta)) dh(x) \\
&= A(\theta') - A(\theta) - (\theta' - \theta) A'(\theta)\,.
\end{aligned}$$

Curiously, this is the Bregman divergence between θ' and θ induced by the convex function A. See Section 26.3 for definitions.

(e) The Crammer-Chernoff method is the solution. Let $\lambda = n(\theta' - \theta)$. Then

$$\begin{aligned}
\mathbb{P}_\theta(\hat{t} \geq \mathbb{E}_{\theta'}[T]) &= \mathbb{P}_\theta(\exp(\lambda \hat{t}) \geq \exp(\lambda \mathbb{E}_{\theta'}[T])) \\
&\leq \mathbb{E}_\theta[\exp(\lambda \hat{t})] \exp(-\lambda \mathbb{E}_{\theta'}[T]) \\
&= \prod_{t=1}^{n} \mathbb{E}_\theta[\exp(\lambda T(X_t)/n)] \exp(-\lambda A'(\theta')) \\
&= \exp(n(A(\theta + \lambda/n) - A(\theta)) - \lambda A'(\theta')) \\
&= \exp\left(n(A(\theta') - A(\theta) - (\theta' - \theta)A'(\theta'))\right) \\
&= \exp\left(-n d(\theta', \theta)\right)\,.
\end{aligned}$$

A symmetric calculation shows that for $\theta' < \theta$,

$$\mathbb{P}_\theta(\hat{t} \leq \mathbb{E}_{\theta'}[T]) \leq \exp\left(-nd(\theta', \theta)\right) .$$

(f) Let $h = \delta_0 + \delta_1$ be the sum of two Dirac measures and $T(x) = x$. Then $p_\theta(0) = 1/(1 + \exp(\theta))$ and $p_\theta(1) = 1/(1 + \exp(-\theta))$ and the domain of A is $\mathbb{R}$.

(g) Let h be Gaussian with mean 0 and variance 1 and $T(x) = x$. Then

$$\begin{aligned} A(\theta) &= \log\left(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp(x\theta - x^2/2) dx\right) \\ &= \log\left(\frac{1}{\sqrt{2\pi}} \exp(\theta^2/2) \int_{\mathbb{R}} \exp(-(x-\theta)^2/2) dx\right) \\ &= \theta^2/2 \,. \end{aligned}$$

Hence $p_\theta(x) = \exp(\theta x - \theta^2/2)$ and

$$\begin{aligned} \int_A p_\theta(x) dh(x) &= \frac{1}{\sqrt{2\pi}} \int_A \exp(\theta x - \theta^2/2) \exp(-x^2/2) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_A \exp((\theta - x)^2/2) dx \,, \end{aligned}$$

which indeed is corresponds to a Gaussian with mean θ.

10.5 When the exponential family is in canonical form the mean of P_θ is $\mu(\theta) = \mathbb{E}_\theta[T] = A'(\theta)$. Since A is strictly convex by the assumption that $\mathcal{M}$ is nonsingular it follows that $\mu(\theta)$ is strictly increasing and hence invertible. Let $\mu_{\sup} = \sup_{\theta \in \Theta} \mu(\theta)$ and $\mu_{\inf} = \inf_{\theta \in \Theta}$ and define

$$\hat{\theta}(x) = \begin{cases} \sup \Theta & \text{if } x \geq \mu_{\sup} \\ \inf \Theta & \text{if } x \leq \mu_{\inf} \\ \mu^{-1}(x) & \text{otherwise} \,. \end{cases}$$

The function $\hat{\theta}$ is the bridge between the empirical mean and the maximum likelihood estimator of θ. Precisely, let $X_1, \ldots, X_n$ be independent and identically distribution from P_θ and $\hat{\mu}_n = \frac{1}{n} \sum_{t=1}^n X_t$. Then provided that $\hat{\theta}_n = \hat{\theta}(\hat{\mu}_n) \in \Theta$, then $\hat{\theta}_n$ is the maximum likelihood estimator of θ,

$$\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} \prod_{t=1}^n \frac{dP_\theta}{dh}(X_t) \,.$$

There is an irritating edge case that $\hat{\mu}_n$ does not lie in the range of $\mu : \Theta \to \mathbb{R}$. When this occurs there is no maximum likelihood estimator.

Part I: Algorithm Then define $\underline{d}(x, y) = \mathbb{I}\{x \leq y\} \lim_{z \downarrow x} d(z, y)$ and $\bar{d}(x, y) = \mathbb{I}\{x \geq y\} \lim_{z \uparrow x} d(z, y)$. The algorithm chooses $A_t = t$ for the first k rounds and subsequently

$A_t = \operatorname{argmax}_i U_i(t)$ where

$$U_i(t) = \sup\left\{\tilde{\theta} \in \Theta : \bar{d}(\hat{\theta}_i(t-1), \tilde{\theta}) \leq \frac{\log(f(T_i(t-1)))}{T_i(t-1)}\right\}.$$

Part II: Concentration Given a fixed $\theta \in \Theta$ and independent random variables $X_1, \ldots, X_n$ sampled from $\mathbb{P}_\theta$ and $\hat{t}_s = \frac{1}{s}\sum_{u=1}^{s} T(X_u)$ and $\hat{\theta}_s = \hat{\theta}(\hat{t}_s)$. Let $\tilde{\theta} \in \Theta$ be such that $\underline{d}(\tilde{\theta}, \theta) = \varepsilon > 0$. Then

$$\mathbb{P}\left(\underline{d}(\hat{\theta}_s, \theta) \geq \varepsilon\right) \leq \mathbb{P}\left(\hat{t} \geq t(\tilde{\theta})\right) \leq \exp(-sd(\tilde{\theta}, \theta)) = \exp(-s\varepsilon)\,, \tag{10.1}$$

where the second inequality follows from Part (e) of Exercise 10.4. Similarly

$$\mathbb{P}\left(\bar{d}(\hat{\theta}_s, \theta) \geq \varepsilon\right) \leq \exp(-s\varepsilon)\,. \tag{10.2}$$

Define random variable τ by

$$\tau = \min\left\{t : \underline{d}(\hat{\theta}_s, \theta - \varepsilon) < \frac{\log(f(t))}{s} \text{ for all } s \in [n]\right\}.$$

In order to bound the expectation of τ we need a connection between $\underline{d}(\hat{\theta}_s, \theta - \varepsilon)$ and $\underline{d}(\hat{\theta}_s, \theta)$. Let $x \leq y - \varepsilon$ and $g(z) = d(x, z)$. Then

$$\begin{aligned}
g(y) &= g(y-\varepsilon) + \int_{y-\varepsilon}^{y} g'(z)dz \\
&= g(y-\varepsilon) + \int_{y-\varepsilon}^{y} (z-x)A''(z)dz \\
&\geq g(y-\varepsilon) + \inf_{z\in[y-\varepsilon,y]} A''(z) \int_{y-\varepsilon}^{y} (z-x)dz \\
&= g(y-\varepsilon) + \frac{1}{2}\inf_{z\in[y-\varepsilon,y]} A''(z)\varepsilon(2y-2x-\varepsilon) \\
&\geq g(y-\varepsilon) + \frac{\varepsilon^2 \inf_{z\in[y-\varepsilon,y]} A''(z)}{2}\,.
\end{aligned}$$

Note that $\inf_{z\in[y-\varepsilon,y]} A''(z) > 0$ is guaranteed because A'' is continuous and $[y-\varepsilon, y]$ is compact and because $\mathcal{M}$ was assumed to be nonsingular. Using this, the expectation of τ is bounded by

$$\begin{aligned}
\mathbb{E}[\tau] &= \sum_{t=1}^{n} \mathbb{P}\left(\tau \geq t\right) \\
&\leq \sum_{t=1}^{n}\sum_{s=1}^{n} \mathbb{P}\left(\underline{d}(\hat{\theta}_s, \theta-\varepsilon) \geq \frac{\log(f(t))}{s}\right) \\
&\leq \sum_{t=1}^{n}\sum_{s=1}^{n} \mathbb{P}\left(\underline{d}(\hat{\theta}_s, \theta) \geq \frac{\varepsilon^2 \inf_{z\in[\theta-\varepsilon,\theta]} A''(z)}{2} + \frac{\log(f(t))}{s}\right) \\
&\leq \sum_{t=1}^{n}\sum_{s=1}^{n} \frac{\exp(-s \inf_{z\in[\theta-\varepsilon,\theta]} A''(z)\varepsilon^2/2)}{f(t)} \\
&= O(1)\,, \qquad (10.3)
\end{aligned}$$

where the last inequality follows from Eq. (10.1) and the final inequality is the same calculation as in the proof of Lemma 10.7. Next let

$$\kappa = \min\{s \geq 1 : \hat{\theta}_u - \theta < \varepsilon \text{ for all } u \geq s\}\,.$$

The expectation of κ is easily bounded using Eq. (10.2),

$$\mathbb{E}[\kappa] \leq \sum_{s=1}^{n}\sum_{u=s}^{\infty} \left(\exp(-u d(\theta+\varepsilon, \theta))\right) = O(1)\,, \qquad (10.4)$$

where we used the fact that $\mathcal{M}$ is non-singular to ensure strict positivity of the divergences.

Part III: Bounding $\mathbb{E}[T_i(n)]$ For each arm i let $\hat{\theta}_{is} = \hat{\theta}(\hat{\mu}_{is})$. Now fix a suboptimal arm i and let $\varepsilon < (\theta_1 - \theta_i)/2$ and

$$\tau = \min\left\{t : \underline{d}(\hat{\theta}_s, \theta-\varepsilon) < \frac{\log(f(t))}{s} \text{ for all } s \in [n]\right\}\,.$$

Then define

$$\kappa = \min\{s \geq 1 : \hat{\theta}_{iu} < \theta_i + \varepsilon \text{ for all } u \geq s\}\,.$$

Then by Eq. (10.3) and Eq. (10.4), $\mathbb{E}[\tau] = O(1)$ and $\mathbb{E}[\kappa] = O(1)$. Suppose that $t \geq \tau$ and $T_i(t-1) \geq \kappa$ and $A_t = i$. Then $U_i(t) \geq U_1(t) \geq \theta_1 - \varepsilon$ and hence

$$d(\theta_i + \varepsilon, \theta_1 - \varepsilon) < \frac{\log(f(n))}{T_i(t-1)}\,.$$

From this we conclude that

$$T_i(n) \leq 1 + \tau + \kappa + \frac{\log(f(n))}{d(\theta_i + \varepsilon, \theta_1 - \varepsilon)}\,.$$

Taking expectations and limits shows that

$$\limsup_{n\to\infty} \frac{\mathbb{E}[T_i(n)]}{\log(n)} \leq \frac{1}{d(\theta_i + \varepsilon, \theta_1 - \varepsilon)}\,.$$

Since the above holds for all sufficiently small $\varepsilon > 0$ and the divergence d is continuous it follows that

$$\limsup_{n\to\infty} \frac{\mathbb{E}[T_i(n)]}{\log(n)} \leq \frac{1}{d(\theta_i, \theta_1)}$$

for all suboptimal arms i. The result follows from the fundamental regret decomposition lemma (Lemma 4.5).

10.6 For simplicity we assume the first arm is uniquely optimal. Define

$$\bar{d}(x, y) = \mathbb{I}\{x \geq y\} \lim_{z \uparrow x} d(z, y)\,, \qquad \underline{d}(x, y) = \mathbb{I}\{x \leq y\} \lim_{z \downarrow x} d(x, y)\,.$$

Let $\mu(\theta) = \int_{\mathbb{R}} x dP_\theta(x)$ and $t(\theta) = \mathbb{E}_\theta[T] = A'(\theta)$ and $\mathcal{T} = \{t(\theta) : \theta \in \Theta\}$. Define $\hat{\theta} : \mathbb{R} \to \mathrm{cl}(\Theta)$ by

$$\hat{\theta}(x) = \begin{cases} t^{-1}(x)\,, & \text{if } x \in \mathcal{T}\,; \\ \sup \Theta\,, & \text{if } x \geq \sup \mathcal{T}\,; \\ \inf \Theta\,, & \text{if } x \leq \inf \mathcal{T}\,. \end{cases}$$

The algorithm is a generalization of KL-UCB. Let

$$\hat{t}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{I}\{A_t = i\}\, T(X_t) \qquad \text{and} \qquad \hat{\theta}_i(t) = \hat{\theta}(\hat{t}_i(t))\,,$$

which is the empirical estimator of the sufficient statistic. Like UCB, the algorithm plays $A_t = t$ for $t \in [k]$ and subsequently $A_t = \mathrm{argmax}_i\, U_i(t)$, where

$$U_i(t) = \sup\left\{\mu(\theta) : d(\hat{\theta}_i(t-1), \theta) \leq \frac{\log(f(T_i(t-1))f(t))}{T_i(t-1)}\right\}$$

and ties in the argmax are broken by choosing the arm with the largest number of plays.

Part I: Concentration Given a fixed $\theta \in \Theta$ and independent random variables $X_1, \ldots, X_n$ sampled from $\mathbb{P}_\theta$ and $\hat{t}_s = \frac{1}{s}\sum_{u=1}^{s} T(X_u)$ and $\hat{\theta}_s = \hat{\theta}(\hat{t}_s)$. Let $\tilde{\theta} \in \Theta$ be such that $\bar{d}(\tilde{\theta}, \theta) = \varepsilon > 0$.

Then

$$\mathbb{P}\left(\bar{d}(\hat{\theta}_s, \theta) \geq \varepsilon\right) \leq \mathbb{P}\left(\hat{t} \geq t(\tilde{\theta})\right) \leq \exp(-sd(\tilde{\theta}, \theta)) = \exp(-s\varepsilon)\,. \tag{10.5}$$

Using an identical argument,

$$\mathbb{P}\left(\underline{d}(\hat{\theta}_s, \theta) \geq \varepsilon) \leq \exp(-s\varepsilon)\right)\,. \tag{10.6}$$

Define random variable τ by

$$\tau = \min\left\{t : d(\hat{\theta}_s, \theta) < \frac{\log(f(s)f(t))}{s} \text{ for all } s \in [n]\right\}\,.$$

Then the expectation of τ is bounded by

$$\begin{aligned}
\mathbb{E}[\tau] &= \sum_{t=1}^{n} \mathbb{P}\left(\tau \geq t\right) \\
&\leq \sum_{t=1}^{n}\sum_{s=1}^{n} \mathbb{P}\left(d(\hat{\theta}_s, \theta + \varepsilon) \geq \frac{\log(f(s)f(t))}{s}\right) \\
&\leq \sum_{t=1}^{n}\sum_{s=1}^{n} \frac{2}{f(s)f(t)} \\
&= O(1)\,,
\end{aligned} \tag{10.7}$$

where the final equality follows from Eqs. (10.5) and (10.6) and the same calculation as in the proof of Lemma 10.7. Next let

$$\kappa = \min\{s \geq 1 : |\hat{\theta}_u - \theta| < \varepsilon \text{ for all } u \geq s\}\,.$$

The expectation of κ is easily bounded Eq. (10.5) and Eq. (10.6):

$$\mathbb{E}[\kappa] \leq \sum_{s=1}^{n}\sum_{u=s}^{\infty} \left(\exp(-ud(\theta + \varepsilon, \theta) + \exp(-ud(\theta - \varepsilon, \theta))\right) = O(1)\,, \tag{10.8}$$

where we used the fact that $\mathcal{M}$ is non-singular to ensure strict positivity of the divergences.

Part II: Bounding $\mathbb{E}[T_i(n)]$ Choose $\varepsilon > 0$ sufficiently small that for all suboptimal arms i,

$$\sup_{\phi \in [\theta_i - \varepsilon, \theta_i + \varepsilon]} \mu(\phi) < \mu^*$$

and define

$$\begin{aligned}
d_{i,\min}(\varepsilon) &= \min_{\phi \in \Theta} \{d(\theta_i + x, \phi) : \mu(\phi) = \mu^*, x = \pm\varepsilon\} \\
d_{i,\inf}(\varepsilon) &= \inf_{\phi \in \Theta} \{d(\theta_i + x, \phi) : \mu(\phi) > \mu^*, x = \pm\varepsilon\}\,.
\end{aligned}$$

Let $\hat{\theta}_{is}$ be the empirical estimate of θ_i based on the first s samples of arm i, which means that $\hat{\theta}_i(t) = \hat{\theta}_{iT_i(t)}$. Let τ be the smallest t such that

$$d(\hat{\theta}_{1s}, \theta_1) < \frac{\log(f(s)f(t))}{s} \qquad \text{for all } s \in [n]\,,$$

which means that $U_1(t) \geq \mu^*$ for all $t \geq \tau$. For suboptimal arms i let κ_i be the random variable

$$\kappa_i = \min\{s : |\hat{\theta}_{iu} - \theta_i| < \varepsilon \text{ for all } u \geq s\}\,.$$

Now suppose that $t \geq \tau$ and $T_i(t-1) \geq \kappa_i$ and $A_t = i$. Then $U_i(t) \geq U_1(t) \geq \mu^*$, which implies that

$$d_{i,\min}(\varepsilon) \geq \frac{\log(f(T_i(t-1))f(t))}{T_i(t-1)}\,.$$

This means that

$$\sum_{i>1} T_i(t) \leq \tau + \sum_{i>1}\left(1 + \max\left\{\kappa_i,\, \frac{\log(t^4)}{d_{i,\min}(\varepsilon)}\right\}\right)\,. \tag{10.9}$$

Then let

$$\Lambda = \max\left\{t : T_1(t-1) \leq \max_{i>1} T_i(t-1)\right\}\,,$$

which by Eq. (10.9) and Eq. (10.7) and Eq. (10.8) satisfies $\mathbb{E}[\Lambda] = O(1)$. Suppose now that $t \geq \Lambda$. Then $T_1(t-1) > \max_{i>1} T_i(t-1)$ and by the definition of the algorithm $A_t = i$ implies that $U_i(t) > \mu^*$ and so

$$T_i(t-1) \leq \frac{\log(T_i(t-1)^2 f(t))}{d_{i,\inf}(\varepsilon)}\,.$$

Hence

$$T_i(n) \leq 1 + \Lambda + \frac{\log(f(T_i(n))f(t))}{d_{i,\inf}(\varepsilon)}\,.$$

Since $\mathbb{E}[\Lambda] = O(1)$ we conclude that

$$\limsup_{n\to\infty} \frac{\mathbb{E}[T_i(n)]}{\log(n)} \leq \frac{1}{d_{i,\inf}(\varepsilon)}\,.$$

The result because $\lim_{\varepsilon \to 0} d_{i,\inf}(\varepsilon) = d_{i,\inf}$ and by the fundamental regret decomposition (Lemma 4.5).

Chapter 11 The Exp3 Algorithm

11.2 Let π be a deterministic policy. Therefore A_t is a function of $x_{1A_1}, \dots, x_{t-1,A_{t-1}}$. We define $(x_t)_{t=1}^n$ inductively by

$$x_{ti} = \begin{cases} 0 & \text{if } A_t = i \\ 1 & \text{otherwise}\,. \end{cases}$$

Clearly the policy will collect zero reward and yet

$$\max_{i \in [k]} \sum_{t=1}^{n} x_{ti} \geq \frac{1}{k} \sum_{t=1}^{n} \sum_{i=1}^{k} x_{ti} = \frac{n(k-1)}{k}\,.$$

Therefore the regret is at least $n(k-1)/k = n(1-1/k)$ as required.

11.5 The first two parts are purely algebraic and are omitted.

(c) Let $G_0, \dots, G_s$ be a sequence of independent geometrically distributed random variables with $\mathbb{E}[G_u] = 1/q_u(\alpha)$. Then

$$\begin{aligned}
\mathbb{P}\left(T_2(n/2) \geq s+1\right) &= \mathbb{P}\left(\sum_{u=0}^{s} G_u \leq n/2\right) \\
&\geq \mathbb{P}\left(\sum_{u=0}^{s-1} G_u \leq n/4\right) \mathbb{P}\left(G_s \leq n/4\right) \\
&= \left(1 - \mathbb{P}\left(\sum_{u=0}^{s-1} G_u > n/4\right)\right)\left(1 - \left(1 - \frac{1}{8n}\right)^{n/4}\right) \\
&\geq \frac{1}{2}(1 - \exp(-1/32)) \\
&\geq \frac{1}{65}\,.
\end{aligned}$$

(d) Suppose that $T_2(n/2) \geq s+1$. Then for all $t > n/2$,

$$\hat{L}_{t2} = \sum_{u=1}^{t-1} \hat{Y}_{u2} \geq 8\alpha n \geq 2n\,.$$

On the other hand,

$$\hat{L}_{t1} = \sum_{u=1}^{t-1} \hat{Y}_{u1} = \sum_{u=n/2+1}^{t-1} \frac{1}{P_{u1}}\,.$$

Using induction and the fact that $P_{t1} \geq 1/2$ as long as $\hat{L}_{t1} \leq \hat{L}_{t2}$ it follows that on the event $E = \{T_t(n/2) \geq s+1\}$ that $P_{t1} \geq 1/2$ for all t. Therefore

$$\begin{aligned} P_{t2} &\leq \exp\left(-\eta \sum_{s=1}^{t-1} (\hat{Y}_{s2} - \hat{Y}_{s1})\right) \\ &\leq \exp\left(\eta \left(\sum_{s=1}^{t-1} \hat{Y}_{s1} - 2n\right)\right) \\ &\leq \exp\left(-n\eta\right) , \end{aligned}$$

which combined with the previous part means that

$$\mathbb{P}\left(A_t = 1 \text{ for all } t > n/2\right) \geq \frac{1}{65}\left(1 - \frac{n}{2}\exp\left(-n\eta\right)\right) .$$

The result follows because on the event $\{A_t = 1 \text{ for all } t > n/2\}$, the regret satisfies

$$\hat{R}_n \geq \frac{n}{2} - \frac{\alpha n n}{2} \geq \frac{n}{4} .$$

(e) Markov's inequality need not hold for negative random variables and $\hat{R}_n$ can be negative. For this problem it even holds that $\mathbb{E}[\hat{R}_n] < 0$.

11.6 The result follows from a long sequence of equalities:

$$\begin{aligned} \mathbb{P}\left(\log a_i + G_i \geq \max_{j \in [k]} \log a_j + G_j\right) &= \mathbb{E}\left[\prod_{j \neq i} \mathbb{P}\left(\log a_j + G_j \leq \log a_i + G_i\right)\right] \\ &= \mathbb{E}\left[\prod_{j \neq i} \exp\left(-\frac{a_j}{a_i}\exp(-G_i)\right)\right] \\ &= \mathbb{E}\left[U_i^{\sum_{j \neq i} \frac{a_j}{a_i}}\right] \\ &= \frac{1}{1 + \sum_{j \neq i} \frac{a_j}{a_i}} \\ &= \frac{a_i}{\sum_{j=1}^k a_j} . \end{aligned}$$

Chapter 12 The Exp3-IX Algorithm

12.2 We proceed in five steps.

Step 1: Decomposition Using that $\sum_a P_{ta} = 1$ and some algebra we get

$$\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}(Z_{ta} - Z_{tA^*}) \\ = \underbrace{\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}(\tilde{Z}_{ta} - \tilde{Z}_{tA^*})}_{\text{(A)}} + \underbrace{\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}(Z_{ta} - \tilde{Z}_{ta})}_{\text{(B)}} + \underbrace{\sum_{t=1}^{n}(\tilde{Z}_{tA^*} - Z_{tA^*})}_{\text{(C)}} .$$

Step 2: Bounding (A) By assumption *(c)* we have $\beta_{ta} \geq 0$, which by assumption *(a)* means that $\eta\tilde{Z}_{ta} \leq \eta\hat{Z}_{ta} \leq \eta|\hat{Z}_{ta}| \leq 1$ for all a. A straightforward modification of the analysis in the last chapter shows that (A) is bounded by

$$\begin{aligned}
\text{(A)} &\leq \frac{\log(k)}{\eta} + \eta\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}\tilde{Z}_{ta}^2 \\
&= \frac{\log(k)}{\eta} + \eta\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}(\hat{Z}_{ta}^2 + \beta_{ta}^2) - 2\eta\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}\hat{Z}_{ta}\beta_{ta} \\
&\leq \frac{\log(k)}{\eta} + \eta\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}\hat{Z}_{ta}^2 + 3\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}\beta_{ta} ,
\end{aligned}$$

where in the last two line we used the assumptions that $\eta\beta_{ta} \leq 1$ and $\eta|\hat{Z}_{ta}| \leq 1$.

Step 3: Bounding (B) For (B) we have

$$\text{(B)} = \sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}(Z_{ta} - \tilde{Z}_{ta}) = \sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}(Z_{ta} - \hat{Z}_{ta} + \beta_{ta}) .$$

We prepare to use Exercise 5.15. By assumptions *(c)* and *(d)* respectively we have $\eta\mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq \beta_{ta}$ and $\mathbb{E}_{t-1}[\hat{Z}_{ta}] = Z_{ta}$. By Jensen's inequality,

$$\eta\mathbb{E}_{t-1}\left[\left(\sum_{a=1}^{k} P_{ta}(Z_{ta} - \hat{Z}_{ta})\right)^2\right] \leq \eta\sum_{a=1}^{k} P_{ta}\mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq \sum_{a=1}^{k} P_{ta}\beta_{ta} .$$

Therefore by Exercise 5.15, with probability at least $1 - \delta$

$$\text{(B)} \leq 2\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}\beta_{ta} + \frac{\log(1/\delta)}{\eta} .$$

Step 4: Bounding (C) For (C) we have

$$\text{(C)} = \sum_{t=1}^{n}(\tilde{Z}_{tA^*} - Z_{tA^*}) = \sum_{t=1}^{n}\left(\hat{Z}_{tA^*} - Z_{tA^*} - \beta_{tA^*}\right) .$$

Because A^* is random we cannot directly apply Exercise 5.15, but need a union bound over all actions. Let a be fixed. Then by Exercise 5.15 and the assumption that $\eta|\hat{Z}_{ta}| \leq 1$ and $\mathbb{E}_{t-1}[\hat{Z}_{ta}] = Z_{ta}$ and $\eta\mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq \beta_{ta}$, with probability at least $1 - \delta$.

$$\sum_{t=1}^{n} \left(\hat{Z}_{ta} - Z_{ta} - \beta_{ta}\right) \leq \frac{\log(1/\delta)}{\eta} .$$

Therefore by a union bound we have with probability at most $1 - k\delta$,

$$\text{(C)} \leq \frac{\log(1/\delta)}{\eta} .$$

Step 5: Putting it together Combining the bounds on (A), (B) and (C) in the last three steps with the decomposition in the first step shows that with probability at least $1 - (k+1)\delta$,

$$R_n \leq \frac{3\log(1/\delta)}{\eta} + \eta \sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}\hat{Z}_{ta}^2 + 5\sum_{t=1}^{n}\sum_{a=1}^{k} P_{ta}\beta_{ta} .$$

where we used the assumption that $\delta \leq 1/k$.

Chapter 13 Lower Bounds: Basic Ideas

13.2 Notice that a policy has zero regret on all bandits for which the first arm is optimal if and only if $\mathbb{P}_{\nu\pi}(A_t = 1) = 1$ for all $t \in [n]$. Hence the policy that always plays the first arm is optimal.

Chapter 14 Foundations of Information Theory

14.4 Let $\mu = P - Q$, which is a signed measure on $(\Omega, \mathcal{F})$. By the Hahn decomposition theorem there exist disjoint sets $A, B \subset \Omega$ such that $A \cup B = \Omega$ and $\mu(E) \geq 0$ for all measurable $E \subseteq A$ and $\mu(E) \leq 0$ for all measurable $E \subseteq B$. Then

$$\begin{aligned}
\int_\Omega X dP - \int_\Omega X dQ &= \int_A X d\mu + \int_B X d\mu \\
&\leq b\mu(A) + a\mu(B) \\
&= (b-a)\mu(A) \\
&\leq (b-a)\delta(P,Q) ,
\end{aligned}$$

where we used the fact that $\mu(B) = P(B) - Q(B) = Q(A) - P(A) = -\mu(A)$.

14.10 Dobrushin's theorem says that for any ω,

$$\mathrm{D}(P(\cdot\,|\,\omega), Q(\cdot\,|\,\omega)) = \sup_{(A_i)} \sum_i P(A_i\,|\,\omega) \log\left(\frac{P(A_i\,|\,\omega)}{Q(A_i\,|\,\omega)}\right) ,$$

where the supremum is taken over all finite partitions of $\mathbb{R}$ with rational-valued end-points. By the definition of a Markov kernel it follows that the quantity inside the supremum on the right-hand side is $\mathcal{F}$-measurable as a function of ω for any finite partition. Since the supremum is over a countable set, the whole right-hand side is $\mathcal{F}$-measurable as required.

14.11 First assume that $P \ll Q$. Then let P^t and Q^t be the restrictions of P and Q to $(\mathbb{R}^t, \mathfrak{B}(\mathbb{R}^t))$ given by

$$P^t(A) = P(A \times \Omega^{n-t}) \qquad \text{and} \qquad Q^t(A) = Q(A \times \Omega^{n-t})\,.$$

You should check that $P \ll Q$ implies that $P^t \ll Q^t$ and hence there exists a Radon-Nikodym derivative dP^t/dQ^t. Define

$$F(x_t \,|\, x_1, \ldots, x_{t-1}) = \frac{dP^t}{dQ^t}(x_1, \ldots, x_t) \Big/ \frac{dP^{t-1}}{dQ^{t-1}}(x_1, \ldots, x_{t-1})\,,$$

which is well defined for all $x_1, \ldots, x_{t-1} \in \mathbb{R}^{t-1}$ except for a set of P^{t-1}-measure zero. Then for any $A \in \mathfrak{B}(\mathbb{R}^{t-1})$ and $B \in \mathfrak{B}(\mathbb{R})$,

$$\begin{aligned}
\int_A \int_B F(x_t \,|\, \omega) Q_t(dx_t \,|\, \omega) P^{t-1}(d\omega) &= \int_A \int_B \frac{dP^t}{dQ^t}(x_t, \omega) Q_t(dx_t \,|\, \omega Q^{t-1}(d\omega) \\
&= \int_{A \times B} \frac{dP^t}{dQ^t} dQ^t \\
&= P(A \times B)\,.
\end{aligned}$$

A monotone class argument shows that $F(x_t \,|\, \omega)$ is P^{t-1}-almost surely the Radon-Nikodym derivative of $P_t(\cdot \,|\, \omega)$ with respect to $Q_t(\cdot \,|\, \omega)$. Hence

$$\begin{aligned}
\mathrm{D}(P, Q) &= \mathbb{E}_P \left[\log \left(\frac{dP}{dQ} \right) \right] \\
&= \sum_{t=1}^n \mathbb{E}_P \left[\log \left(F(X_t \,|\, X_1, \ldots, X_{t-1}) \right) \right] \\
&= \sum_{t=1}^n \mathbb{E}_P \left[\mathrm{D}(P_t(\cdot \,|\, X_1, \ldots, X_{t-1}), Q_t(\cdot \,|\, X_1, \ldots, X_{t-1})) \right]\,.
\end{aligned}$$

Now suppose that $P \not\ll Q$. Then by definition $\mathrm{D}(P, Q) = \infty$. We need to show this implies there exists a $t \in [n]$ such that $\mathrm{D}(P_t(\cdot \,|\, \omega), Q_t(\cdot \,|\, \omega)) = \infty$ with nonzero probability. Proving the contrapositive, let

$$U_t = \{\omega : \mathrm{D}(P_t(\cdot \,|\, \omega), Q_t(\cdot \,|\, \omega)) < \infty\}$$

and assume that $P(U_t = 1)$ for all t. Then $U = \cap_{t=1}^n U_t$ satisfies $P(U) = 1$. On U_t let $F(x_t \,|\, x_1, \ldots, x_{t-1}) = dP_t(\cdot \,|\, x_1, \ldots, x_{t-1})/dQ_t(\cdot \,|\, x_1, \ldots, x_{t-1})(x_t)$ and otherwise let $F(x_t \,|\, x_1, \ldots, x_{t-1}) = 0$. Iterating applications of Fubini's theorem shows that for any $(A_t)_{t=1}^n$

with $A_t \in \mathfrak{B}(\mathbb{R})$ it holds that

$$\int_{A_1 \times \cdots \times A_n} \prod_{t=1}^{n} F(x_t \,|\, x_1, \ldots, x_{t-1}) Q(dx_1, \ldots, dx_n) = P(A_1 \times \cdots \times A_n)\,.$$

Hence $\prod_{t=1}^{n} F(x_t \,|\, x_1, \ldots, x_{t-1})$ behaves like the Radon-Nikodym derivative of P with respect to Q on rectangles. Another monotone class argument extends this to all measurable sets and the existence of dP/dQ guarantees that $P \ll Q$.

Chapter 15 Minimax Lower Bounds

15.1 Abbreviate $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ and let $R(P) = \mathbb{E}_P[d(\hat{\theta}, P)]$. By the triangle inequality

$$d(\hat{\theta}, P_0) + d(\hat{\theta}, P_1) \geq d(P_0, P_1) = \Delta\,.$$

Let $E = \{d(\hat{\theta}, P_0) \leq \Delta/2\}$. On E^c it holds that $d(\hat{\theta}, P_0) \geq \Delta/2$ and on E it holds that $d(\hat{\theta}, P_1) \geq \Delta - d(\hat{\theta}, P_0) \geq \Delta/2$.

$$R(P_0) + R(P_1) \geq \frac{\Delta}{2}(P_0(E^c) + P_1(E)) \geq \frac{\Delta}{4} \exp(-\operatorname{D}(P_0, P_1))\,.$$

The result follows because $\max\{a, b\} \geq (a + b)/2$.

Chapter 16 Instance Dependent Lower Bounds

16.2

(a) Suppose that $\mu \neq \mu'$. Then $\operatorname{D}(\mathcal{R}(\mu), \mathcal{R}(\mu')) = \infty$, since $\mathcal{R}(\mu)$ and $\mathcal{R}(\mu')$ are not absolutely continuous. Therefore $d_{\inf}(\mathcal{R}(\mu), \mu^*, \mathcal{M}) = \infty$.

(b) Notice that each arm returns exactly two possible rewards and once these have been observed, then the mean is known. Consider the algorithm that plays each arm until it has observed both possible rewards from that arm and subsequently plays optimally. The expected number of trials before both rewards from an arm are observed is $\sum_{i=2}^{\infty} i 2^{1-i} = 3$. Hence

$$R_n \leq 3 \sum_{i=1}^{k} \Delta_i\,.$$

(c) Let P_μ be the shifted Rademacher distribution with mean μ. Then $D(P_\mu, P_{\mu+\Delta})$ is not differentiable as a function of Δ.

Chapter 17 High Probability Lower Bounds

17.1

Proof of Claim 17.5. We have

$$\delta \le \mathbb{P}_Q(\hat{R}_n \ge u) = \int_{[0,1]^{n\times k}} \mathbb{P}_{\delta_x}(\hat{R}_n \ge u) dQ(x)\,.$$

Therefore there exists an x with $\mathbb{P}_{\delta_x}(\hat{R}_n \ge u) \ge \delta$. □

Proof of Claim 17.6. Abbreviate $\mathbb{P}_i$ to be the law of $A_1, X_1, \ldots, A_n, X_n$ induced by $\mathbb{P}_{Q_i}$ and let $\mathbb{E}_i$ be the corresponding expectation operator. Following the standard argument, let j be the arm that minimizes $\mathbb{E}_1[T_j(n)]$, which satisfies

$$\mathbb{E}_1[T_j(n)] \le \frac{n}{k-1}\,.$$

Therefore by Theorem 14.2 and Lemma 15.1,

$$\begin{aligned}
\max\left\{\mathbb{P}_1(T_1(n) \le n/2), \mathbb{P}_j(T_j(n) \le n/2)\right\} &\ge \mathbb{P}_1(T_1(n) \le n/2) + \mathbb{P}_j(T_1(n) > n/2) \\
&\ge \frac{1}{2}\exp\left(-\mathbb{E}_1[T_i(n)]2\Delta^2\right) \\
&\ge \frac{1}{2}\exp\left(-\mathbb{E}_1[T_i(n)]\frac{k-1}{n}\log\left(\frac{1}{8\delta}\right)\right) \\
&\ge 4\delta\,.
\end{aligned}$$

Therefore there exists an i such that $\mathbb{P}_i(T_i(n) \le n/2) \ge 2\delta$. □

Proof of Claim 17.7. Notice that if $\eta_t + 2\Delta < 1$ and $\eta_t > 0$, then $X_{tj} \in (0,1)$ for all $j \in [k]$. Now

$$\begin{aligned}
\mathbb{P}_i\left(\eta_t + 2\Delta \ge 1 \text{ or } \eta_t \le 0\right) &\le \exp\left(-\frac{(1/2 - 2\Delta)^2}{2\sigma^2}\right) + \exp\left(-\frac{(1/2)^2}{2\sigma^2}\right) \\
&\le \exp\left(-\frac{100}{32}\right) + \exp\left(-\frac{25}{2}\right) \\
&\le \frac{1}{8}\,.
\end{aligned}$$

Let $M = \sum_{t=1}^n \mathbb{I}\left\{\eta_t \le 0 \text{ or } \eta_t + 2\Delta \ge 1\right\}$, which is an upper bound on the number of rounds where clipping occurs. By Hoeffding's bound,

$$\mathbb{P}_i\left(M \ge \mathbb{E}_i[M] + \sqrt{\frac{n\log(1/\delta)}{2}}\right) \le \delta\,,$$

which means that with probability at least $1-\delta$,

$$M \le \frac{n}{8} + \sqrt{\frac{n\log(1/\delta)}{2}} \le \frac{n}{4}\,,$$

where we used the fact that $n \ge 32\log(1/\delta)$. □

Chapter 18 Contextual Bandits

18.1

(a) By Jensen's inequality,

$$\begin{aligned}\sum_{c\in\mathcal{C}}\sqrt{\sum_{t=1}^{n}\mathbb{I}\{c_t=c\}} &= |\mathcal{C}|\sum_{c\in\mathcal{C}}\frac{1}{|\mathcal{C}|}\sqrt{\sum_{t=1}^{n}\mathbb{I}\{c_t=c\}} \\ &\le |\mathcal{C}|\sqrt{\sum_{c\in\mathcal{C}}\frac{1}{|\mathcal{C}|}\sum_{t=1}^{n}\mathbb{I}\{c_t=c\}} \\ &= \sqrt{|\mathcal{C}|n}\,,\end{aligned}$$

where the inequality follows from Jensen's inequality and the concavity of $\sqrt{\cdot}$ and the last equality follows since $\sum_{c\in\mathcal{C}}\sum_{t=1}^{n}\mathbb{I}\{c_t=c\}=n$.

(b) When each context occurs $n/|\mathcal{C}|$ times we have

$$\sum_{c\in\mathcal{C}}\sqrt{\sum_{t=1}^{n}\mathbb{I}\{c_t=c\}} = \sqrt{n|\mathcal{C}|}\,,$$

which matches the upper bound proven in the previous part.

18.6

(a) This follows because $X_i \le \max_j X_j$ for any family of random variables $(X_i)_i$. Hence $\mathbb{E}[X_i] \le \mathbb{E}[\max_j X_j]$.

(b) Modify the proof by proving a bound on

$$\mathbb{E}\left[\sum_{t=1}^{n}E_{m^*}^{(t)}x_t - \sum_{t=1}^{n}X_t\right]$$

for an arbitrarily fixed m^*. By the definition of learning experts, $\mathbb{E}_t[\hat{X}_t]=x_t$ and so Eq. (18.10) also remains valid. Note this would not be true in general if $E^{(t)}$ were allowed to depend on A_t. The rest follows the same way as in the oblivious case.

18.7 The inequality $E_n^* \le nk$ is trivial since $\max_m E_{mi}^{(t)} \le 1$. To prove $E_n^* \le nM$, let $m_{ti}^* = \operatorname{argmax}_m E_{mi}^{(t)}$. Then

$$E_n^* = \sum_{t=1}^{n} \sum_{i=1}^{k} \sum_{m=1}^{M} E_{mi}^{(t)} \mathbb{I}\{m = m_{ti}^*\} \le \sum_{t=1}^{n} \sum_{m=1}^{M} \sum_{i=1}^{k} E_{mi}^{(t)} = nM\,,$$

where in the last step we used the fact that $\sum_{i=1}^{k} E_{mi}^{(t)} = 1$.

18.8 Let $\hat{X}_{t\phi} = k\mathbb{I}\{A_t = \phi(C_t)\}$. Assume that $t \le m$. Since A_t is chosen uniformly at random,

$$\begin{aligned}
\mathbb{E}[\hat{X}_{t\phi}] &= \mathbb{E}\left[k\mathbb{I}\{A_t = \phi(C_t)\} X_t\right] \\
&= \mathbb{E}\left[k\mathbb{I}\{A_t = \phi(C_t)\} \mathbb{E}\left[X_t \mid A_t, C_t\right]\right] \\
&= \mathbb{E}\left[k\mathbb{I}\{A_t = \phi(C_t)\} \mu(C_t, A_t)\right] \\
&= \mathbb{E}\left[\mu(C_t, \phi(C_t))\right] \\
&= \mu(\phi)\,.
\end{aligned}$$

The variance satisfies

$$\mathbb{E}[(\hat{X}_{t\phi} - \mu(\phi))^2] \le \mathbb{E}[\hat{X}_{t\phi}^2] \le k^2 \mathbb{E}[\mathbb{I}\{A_t = \phi(C_t)\}] = k\,.$$

Finally note that $\hat{X}_{t\phi} \in [0, k]$. Using the technique from Exercise 5.14,

$$\mathbb{E}[\exp(\lambda(\hat{\mu}(\phi) - \mu(\phi)))] \le \exp\left(\frac{k\lambda^2 g(\lambda k/m)}{m}\right)\,,$$

where $g(x) = (\exp(x) - 1 - x)/x^2$, which for $x \in (0, 1]$ satisfies $g(x) \le 1/2 + x/4$. Suppose that $m \ge 2k \log(2|\Phi|)$. Then by following the argument in Exercise 5.18,

$$\begin{aligned}
\mathbb{E}\left[\max_{\phi} |\hat{\mu}(\phi) - \mu(\phi)|\right] &\le \inf_{\lambda > 0} \left(\frac{\log(2|\Phi|)}{\lambda} + \frac{k\lambda}{2m} + \frac{k^2\lambda^2}{4m^2}\right) \\
&\le \sqrt{\frac{2k \log(2|\Phi|)}{m}} + \frac{k \log(2|\Phi|)}{2m}\,,
\end{aligned}$$

where the second inequality follows by choosing

$$\lambda = \sqrt{\frac{2m \log(2|\Phi|)}{k}} \le \frac{m}{k}\,.$$

Therefore the regret is bounded by

$$R_n \le m + 2n\sqrt{\frac{2k \log(2|\Phi|)}{m}} + \frac{nk \log(2|\Phi|)}{m}\,.$$

By tuning m it follows that

$$R_n = O\left(n^{2/3}(k\log(|\Phi|))^{1/3}\right)$$

Chapter 19 Stochastic Linear Bandits

19.2 Let $\mathcal{T}$ be the set of rounds t when $\|x_t\|_{V_{t-1}^{-1}} \geq 1$ and $G_t = V_0 + \sum_{s=1}^{t} \mathbb{I}_{\mathcal{T}}(s) x_s x_s^\top$. Then

$$\begin{aligned}
\left(\frac{d\lambda + |\mathcal{T}|L^2}{d}\right)^d &\geq \left(\frac{\operatorname{trace}(G_n)}{d}\right)^d \\
&\geq \det(G_n) \\
&= \det(V_0) \prod_{t\in T} (1 + \|x_t\|^2_{G_{t-1}^{-1}}) \\
&\geq \det(V_0) \prod_{t\in T} (1 + \|x_t\|^2_{V_{t-1}^{-1}}) \\
&\geq \lambda^d 2^{|\mathcal{T}|}\,.
\end{aligned}$$

Rearranging and taking the logarithm shows that

$$|\mathcal{T}| \leq \frac{d}{\log(2)} \log\left(1 + \frac{|\mathcal{T}|L^2}{d\lambda}\right).$$

Abbreviate $a = d/\log(2)$ and $b = L^2/d\lambda$, which are both positive. Then

$$a\log\left(1 + b(3a\log(1+ab))\right) \leq a\log\left(1 + 3a^2b^2\right) \leq a\log(1+ab)^3 = 3a\log(1+ab)\,.$$

Since $x - a\log(1+bx)$ is decreasing for $x \geq 3a\log(1+ab)$ it follows that

$$|\mathcal{T}| \leq 3a\log(1+ab) = \frac{3d}{\log(2)} \log\left(1 + \frac{L^2}{\lambda\log(2)}\right).$$

19.3 We only present in detail the solution for the first part.

(a) Partition $\mathcal{C}$ into m equal length subintervals, call these $\mathcal{C}_1, \ldots, \mathcal{C}_m$. Associate a bandit algorithm with each of these subintervals. In round t, upon seeing $C_t \in \mathcal{C}$, play the bandit algorithm associated with the unique subinterval that C_t belongs to. For example, one could use a Exp3

as in the previous chapter, or UCB. The regret is

$$\begin{aligned} R_n &= \mathbb{E}\left[\sum_{t=1}^{n} \max_{a\in[k]} r(C_t, a) - \sum_{t=1}^{n} X_t\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \max_{a\in[k]} r(C_t, a) - \max_{a} \tilde{r}([C_t], a)\right]}_{\text{(I)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \max_{a\in[k]} \tilde{r}([C_t], a) - X_t\right]}_{\text{(II)}}, \end{aligned}$$

where for $c \in \mathcal{C}$, $[c]$ is the index of the unique part $\mathcal{C}_i$ that c belongs to and for $i \in [m]$,

$$\tilde{r}(i, a) = \begin{cases} \mathbb{E}\left[r(C_t, a) \,|\, [C_t] = i\right] & \text{if } \mathbb{P}\left([C_t] = i\right) > 0 \\ 0 & \text{otherwise}\,. \end{cases}$$

The first term in the regret decomposition is called the approximation error, and the second is the error due to learning. The approximation error is bounded using the Lipschitz assumption and the definition of the discretization:

$$\begin{aligned} \max_{a\in[k]} r(c, a) - \max_{a\in[k]} \tilde{r}([c], a) &\le \max_{a\in[k]} (r(c, a) - \tilde{r}([c], a)) \\ &= \max_{a\in[k]} \mathbb{E}[r(c, a) - r(C_t, a) \,|\, [C_t] = [c]]) \\ &\le \max_{a\in[k]} \mathbb{E}[L|c - C_t| \,|\, [C_t] = [c]]) \\ &\le L/m\,, \end{aligned}$$

where in the second last inequality we used the assumption that r is Lipschitz and in the last that when $[C_t] = [c]$ it holds that $|C_t - c| \le 1/m$. It remains to bound the error due to learning. Note that $\mathbb{E}[X_t \,|\, [C_t] = i] = \tilde{r}(i, A_t)$. As a result, the data experienced by the bandit associated with $\mathcal{C}_i$ satisfies the conditions of a stochastic bandit environment. Consider the case when Exp3 is used with the adaptive learning rate as described in Exercise 28.11. For $i \in [m]$, let $T_i = \{t \in [n] \,|\, [C_t] = i\}$, $N_i = |T_i|$ and

$$R_{ni} = \sum_{t\in T_i} \left(\max_{a\in[k]} \tilde{r}([C_t], a) - X_t\right).$$

Then, by Eq. (11.2), $\mathbb{E}[R_{ni}] \le C\mathbb{E}[\sqrt{k \log(k) N_i}]$ and thus

$$\text{(II)} \le \sum_{i=1}^{m} \mathbb{E}\left[R_{ni}\right] \le C\mathbb{E}\left[(k \log(k))^{1/2} \sum_{i=1}^{m} N_i^{1/2}\right] \le C\sqrt{k \log(k) m n}\,,$$

where the last inequality follows from Cauchy-Schwarz. Hence,

$$R_n \le \frac{Ln}{m} + C\sqrt{k \log(k) m n}\,.$$

Optimizing m gives $m = (L/C)^{2/3}(n/(k\log(k)))^{1/3}$ and

$$R_n \leq C' n^{2/3}(Lk\log(k))^{1/3}$$

for some constant C' that depends only on C. The same argument works with no change if the bandit algorithm is switched to UCB, just a little extra work is needed to deal with the fact that UCB will be run for a random number of rounds. Luckily, the number of rounds is independent of the rewards experienced and the actions taken.

(b) Consider a lower bound that hides a 'spike' at one of m positions.

(c) Make an argument that $\mathcal{C}$ can be partitioned into $m = (3dL/\varepsilon)^d$ partitions to guarantee that for fixed a the function $r(c, a)$ varies by at most ε within each partition. Then bound the regret by

$$R_n \leq n\varepsilon + \sqrt{2nk\left(\frac{3dL}{\varepsilon}\right)^d \log(k)}\,.$$

By optimizing ε you should arrive at a bound that depends on the horizon like $O(n^{(d+1)/(d+2)})$, which is really quite bad, but also not improvable without further assumptions. You might find the results of Exercise 20.3 to be useful.

Chapter 20 Confidence Bounds for Least Squares Estimators

20.1 From the definition of the design $V_n = mI$ and the ith coordinate of $\hat{\theta}_n - \theta_*$ is the sum of m independent standard Gaussian random variables. Hence

$$\mathbb{E}\left[\|\hat{\theta}_n - \theta_*\|_{V_n}^2\right] = m\sum_{i=1}^{d}\frac{1}{m} = d\,.$$

20.2 Let $m = n/d$ and assume for simplicity that m is a whole number. The actions $(A_t)_{t=1}^n$ are chosen in blocks of size m with the ith block starting in round $t_i = (i-1)m + 1$. For all $i \in \{1, \ldots, d\}$ we set $A_{t_i} = e_i$. For $t \in \{t_i, \ldots, t_i + m - 1\}$ we set $A_{t_i} = e_i$ if $\eta_{t_i} > 0$ and $A_t = \mathbf{0}$ otherwise. Clearly $V_n^{-1} \geq I$ and hence $\|\mathbf{1}\|_{V_n^{-1}} \leq d$. The choice of $(A_t)_{t=1}^n$ ensures that $\mathbb{E}[\hat{\theta}_{n,i}]$ is independent of i. Furthermore,

$$\mathbb{E}[\hat{\theta}_{n,1}] = \mathbb{E}\left[\mathbb{I}\left\{\eta > 0\right\}\frac{\sum_{t=1}^m \eta_t}{m} + \mathbb{I}\left\{\eta_1 < 0\right\}\eta_1\right] = \frac{1}{\sqrt{2\pi}}\left(\frac{1}{m} - 1\right)\,.$$

Therefore

$$\mathbb{E}\left[\frac{\langle\hat{\theta}_n, \mathbf{1}\rangle^2}{\|\mathbf{1}\|_{V_n^{-1}}^2}\right] \geq \frac{1}{d}\mathbb{E}\left[\langle\hat{\theta}_n, \mathbf{1}\rangle^2\right] \geq \frac{1}{d}\mathbb{E}\left[\langle\hat{\theta}_n, \mathbf{1}\rangle\right]^2 = \frac{d}{2\pi}\left(\frac{m-1}{m}\right)^2\,.$$

20.3

(a) If $C \subset \mathcal{A}$ is an ε-covering then it is also an ε'-covering with any $\varepsilon' \geq \varepsilon$. Hence, $\varepsilon \to N(\varepsilon)$ is a decreasing function of ε.

(b) The inequality $M(2\varepsilon) \leq N(\varepsilon)$ amounts to showing that any 2ε packing has a cardinality at most the cardinality of any ε covering. Assume this does not hold, that is, there is a 2ε packing $P \subset \mathcal{A}$ and an ε-covering $C \subset \mathcal{A}$ such that $|P| \geq |C|+1$. By the pigeonhole principle, there is $c \in C$ such that there are distinct $x, y \in P$ such that $x, y \in B(c, \varepsilon)$. Then $\|x - y\| \leq \|x - c\| + \|c - y\| \leq 2\varepsilon$, which contradicts that P is a 2ε-packing.

If $M(\varepsilon) = \infty$, the inequality $N(\varepsilon) \leq M(\varepsilon)$ is trivially true. Otherwise take a maximum ε-packing P of $\mathcal{A}$. This packing is automatically an ε-covering as well (otherwise P would not be a maximum packing), hence, the result.

(c) We show the inequalities going left to right. For the first inequality, if $N \doteq N(\varepsilon) = \infty$ then there is nothing to be shown. Otherwise let C be a minimum cardinality ε-cover of $\mathcal{A}$. Then from the definition of cover and the additivity of volume, $\mathrm{vol}(\mathcal{A}) \leq \sum_{x \in \mathcal{A}'} \mathrm{vol}(B(x, \varepsilon)) = N\varepsilon^d \mathrm{vol}(B)$. Reordering gives the inequality.

The next inequality, namely that $N(\varepsilon) \leq M(\varepsilon)$ has already been shown.

Consider now the inequality bounding $M \doteq M(\varepsilon)$. Let P be a maximum cardinality ε-packing of $\mathcal{A}$. Then, for any $x, y \in P$ distinct, $B(x, \varepsilon/2) \cap B(y, \varepsilon/2) = \emptyset$. Further, for $x \in P$, $B(x, \varepsilon/2) \subset \mathcal{A} + \frac{\varepsilon}{2}B$ and thus $\cup_{x \in P} B(x, \varepsilon/2) \subset \mathcal{A} + \frac{\varepsilon}{2}B$, hence, by the additivity of volume, $M\mathrm{vol}(\frac{\varepsilon}{2}B) \leq \mathrm{vol}(\mathcal{A} + \frac{\varepsilon}{2}B)$.

For the next inequality note that $\varepsilon B \subset \mathcal{A}$ immediately implies that $\mathcal{A} + \frac{\varepsilon}{2}B \subset \mathcal{A} + \frac{1}{2}\mathcal{A}$ (check the containment using the definitions), while the convexity of $\mathcal{A}$ implies that $\mathcal{A} + \frac{1}{2}\mathcal{A} \subset \frac{3}{2}\mathcal{A}$. For this second claim let $u \in \mathcal{A} + \frac{1}{2}\mathcal{A}$. Then $u = x + \frac{1}{2}y$ for some $x, y \in \mathcal{A}$. By the convexity of $\mathcal{A}$, $\frac{2}{3}u = \frac{2}{3}x + \frac{1}{3}y \in \mathcal{A}$ and hence $u = \frac{3}{2}(\frac{2}{3}u) \in \frac{3}{2}\mathcal{A}$. For the final inequality note that for measurable X and $c > 0$ we have $\mathrm{vol}(cX) = c^d \mathrm{vol}(X)$. This is true because cX is the image of X under the linear mapping represented by a diagonal matrix with c on the diagonal and this matrix has determinant c^d.

(d) Let $\mathcal{A}$ be bounded, and say, $\mathcal{A} \subset rB$ for some $r > 0$. Then $\mathrm{vol}(\mathcal{A} + \varepsilon/2B) \leq \mathrm{vol}(rB + \varepsilon/2B) = \mathrm{vol}((r + \varepsilon/2)B) < +\infty$, hence, the previous part gives that $N(\varepsilon) \leq M(\varepsilon) < +\infty$. Now assume that $N(\varepsilon) < \infty$ and let C be a minimum cover of $\mathcal{A}$. Then $\mathcal{A} \subset \cup_{x \in C} B(x, \varepsilon) \subset \cup_{x \in C}(\|x\| + \varepsilon)B \subset \max_{x \in C}(\|x\| + \varepsilon)B$ hence, $\mathcal{A}$ is bounded.

20.6 Proving that $\bar{M}_t$ is $\mathcal{F}_t$-measurable is actually not trivial. It follows because $M_t(\cdot)$ is measurable and by the 'sections' lemma [Kallenberg, 2002, Lem 1.26]. It remains to show that $\mathbb{E}[\bar{M}_t \,|\, \mathcal{F}_{t-1}] \leq \bar{M}_{t-1}$ almost surely. Proceeding by contradiction, suppose that $\mathbb{P}(\mathbb{E}[\bar{M}_t \,|\, \mathcal{F}_{t-1}] - \bar{M}_{t-1} > 0) > 0$. Then there exists an $\varepsilon > 0$ such that the set $A = \{\omega : \mathbb{E}[\bar{M}_t \,|\, \mathcal{F}_{t-1}](\omega) - \bar{M}_{t-1}(\omega) > \varepsilon\} \in \mathcal{F}_{t-1}$

satisfies $\mathbb{P}(A) > 0$. Then

$$\begin{aligned} 0 < \int_A (\mathbb{E}[\bar{M}_t \,|\, \mathcal{F}_{t-1}] - \bar{M}_{t-1}) d\mathbb{P} &= \int_A (\bar{M}_t - \bar{M}_{t-1})) d\mathbb{P} \\ &= \int_A \int_{\mathbb{R}^d} (M_t(x) - M_{t-1}(x)) dh(x) d\mathbb{P} \\ &= \int_{\mathbb{R}^d} \int_A (M_t(x) - M_{t-1}(x)) d\mathbb{P} dh(x) \\ &\leq 0\,, \end{aligned}$$

where the first equality follows from the definition of conditional expectation, the second by substituting the definition of $\bar{M}_t$ and the third from Fubini-Tonelli's theorem. The last follows from Lemma 20.2 and the definition of conditional expectation again. The proof is completed by noting the deep result that $0 \not< 0$. In this proof it is necessary to be careful to avoid integrating over conditional $\mathbb{E}[M_t(x) \,|\, \mathcal{F}_{t-1}]$, which are only defined for each x almost surely and need not be measurable as a function of x.

20.9 Let $f(\lambda) = \frac{1}{\sqrt{2\pi}} \exp(-\lambda^2/2)$ be the density of the standard Gaussian and define supermartingale M_t by

$$M_t = \int_{\mathbb{R}} f(\lambda) \exp\left(\lambda S_t - \frac{t\sigma^2\lambda^2}{2}\right) d\lambda = \frac{1}{\sqrt{t\sigma^2 + 1}} \exp\left(\frac{S_t^2}{2\sigma^2(t+1)}\right).$$

Since $\mathbb{E}[M_\tau] = M_0 = 1$, the maximal inequality shows that $\mathbb{P}(\sup_t M_t \geq 1/\delta) \leq \delta$, which after rearranging the previous display completes the result.

20.10

(a) This follows from straightforward calculus.

(b) The result is trivial for $\Lambda < 0$. For $\Lambda \geq 0$ we have

$$\begin{aligned} M_n &= \int_{\mathbb{R}} f(\lambda) \exp\left(\lambda S_n - \frac{\lambda^2 n}{2}\right) d\lambda \\ &\geq \int_{\Lambda}^{\Lambda(1+\varepsilon)} f(\lambda) \exp\left(\lambda S_n - \frac{\lambda^2 n}{2}\right) d\lambda \\ &\geq \varepsilon\Lambda f(\Lambda(1+\varepsilon)) \exp\left(\Lambda(1+\varepsilon)S_n - \frac{\Lambda^2(1+\varepsilon)^2 n}{2}\right) \\ &= \varepsilon\Lambda f(\Lambda(1+\varepsilon)) \exp\left(\frac{(1-\varepsilon^2)S_n^2}{2n}\right). \end{aligned}$$

(c) Let $n \in \mathbb{N}$. Since M_t is a supermartingale with $M_0 = 1$ it follows that

$$P_n = \mathbb{P}(\text{exists } t \leq n : M_t \geq 1/\delta) \leq \delta\,.$$

Hence $\mathbb{P}\left(\text{exists } t : M_t \geq 1/\delta\right) \leq \delta$. Substituting the result from the previous part and rearranging completes the proof.

(d) A suitable choice of f is $f(\lambda) = \dfrac{\mathbb{I}\left\{\lambda \leq e^{-e}\right\}}{\lambda \log\left(\frac{1}{\lambda}\right)\left(\log\log\left(\frac{1}{\lambda}\right)\right)^2}$.

(e) Let $\varepsilon_n = \min\{1/2, 1/\log\log(n)\}$ and $\delta \in [0,1]$ be the largest (random) value such that S_n never exceeds

$$\sqrt{\frac{2n}{1-\varepsilon_n^2}\left(\log\left(\frac{1}{\delta}\right) + \log\left(\frac{1}{\varepsilon_n \Lambda_n f(\Lambda_n(1+\varepsilon_n))}\right)\right)}\,.$$

By Part (c) we have $\mathbb{P}\left(\delta > 0\right) = 1$. Furthermore, $\limsup_{n\to\infty} S_n/n = 0$ almost surely by the strong law of large numbers, so that $\Lambda_n \to 0$ almost surely. On the intersection of these almost sure events we have

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log(n)}} \leq 1\,.$$

20.11 We first show a bound on the right tail of S_t. A symmetric argument suffices for the left tail. Let $Y_s = X_s - \mu_s|X_s|$ and $M_t(\lambda) = \exp(\sum_{s=1}^t(\lambda Y_s - \lambda^2|X_s|/2))$. Define filtration $\mathcal{G}_1 \subset \cdots \subset \mathcal{G}_n$ by $\mathcal{G}_t = \sigma(\mathcal{F}_{t-1}, |X_t|)$. Using the fact that $X_s \in \{-1,0,1\}$ we have for any $\lambda > 0$ that

$$\mathbb{E}[\exp(\lambda Y_s - \lambda^2|X_s|/2)\,|\,\mathcal{G}_s] \leq 1\,.$$

Therefore $M_t(\lambda)$ is a supermartingale for any $\lambda > 0$. The next step is to use the method of mixtures with a uniform distribution on $[0,2]$. Let $M_t = \int_0^2 M_t(\lambda)d\lambda$. Then Markov's inequality shows that for any $\mathcal{G}_t$-measurable stopping time τ with $\tau \leq n$ almost surely, $\mathbb{P}\left(M_\tau \geq 1/\delta\right) \leq \delta$. Next we need a bound on M_τ. The following holds whenever $S_t \geq 0$.

$$\begin{aligned}
M_t &= \frac{1}{2}\int_0^2 M_t(\lambda)d\lambda \\
&= \frac{1}{2}\sqrt{\frac{\pi}{2N_t}}\left(\operatorname{erf}\left(\frac{S_t}{\sqrt{2N_t}}\right) + \operatorname{erf}\left(\frac{2N_t - S_t}{\sqrt{2N_t}}\right)\right)\exp\left(\frac{S_t^2}{2N_t}\right) \\
&\geq \frac{\operatorname{erf}(\sqrt{2})}{2}\sqrt{\frac{\pi}{2N_t}}\exp\left(\frac{S_t^2}{2N_t}\right)\,.
\end{aligned}$$

The bound on the upper tail completed via a stopping time, which shows that

$$\mathbb{P}\left(\text{exists } t \leq n : S_t \geq \sqrt{2N_t\log\left(\frac{2}{\delta\operatorname{erf}(\sqrt{2})}\sqrt{\frac{2N_t}{\pi}}\right)} \text{ and } N_t > 0\right) \leq \delta\,.$$

The result follows by symmetry and union bound.

20.12

(a) Following the hint, we show that $\exp(L_t(\theta_*))$ is a martingale. Indeed, letting $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$,

$$\begin{aligned}
\mathbb{E}\left[\exp(L_t(\theta_*))|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[p_{\hat{\theta}_{t-1}}(X_t)/p_{\theta_*}(X_t)\right]\exp(L_{t-1}(\theta_*)) \\
&= \exp(L_{t-1}(\theta_*)) \int \cancel{p_{\theta_*}(x)} \frac{p_{\hat{\theta}_{t-1}}(x)}{\cancel{p_{\theta_*}(x)}} d\mu(x) \\
&= \exp(L_{t-1}(\theta_*)) \int p_{\hat{\theta}_{t-1}}(x) d\mu(x) \\
&= \exp(L_{t-1}(\theta_*))\,.
\end{aligned}$$

Then, applying the Cramer-Chernoff trick,

$$\begin{aligned}
\mathbb{P}\left(L_t(\theta_*) \geq \log(1/\delta) \text{ for some } t \geq 1\right) &= \mathbb{P}\left(\sup_{t\in\mathbb{N}} \exp(L_t(\theta_*)) \geq 1/\delta\right) \\
&\leq \delta \mathbb{E}\left[\exp(L_0(\theta_*))\right] = \delta\,,
\end{aligned}$$

where the inequality is due to Theorem 3.9, the maximal inequality of nonnegative supermartingales.

(b) This follows from the definition of $\mathcal{C}_t$ and Part (a).

Chapter 21 Optimal Design for Least Squares Estimators

21.1 Following the hint,

$$\nabla f(\pi)_a = \frac{\operatorname{trace}(\operatorname{adj}(V(\pi))aa^\top)}{\det(V(\pi))} = \frac{a^\top \operatorname{adj}(V(\pi))a}{\det(V(\pi))} = a^\top V(\pi)^{-1} a = \|a\|^2_{V(\pi)^{-1}}\,,$$

where in the third equality we used that $\operatorname{adj}(V(\pi))$ is symmetric since $V(\pi)$ is symmetric, hence, following the hint, $\operatorname{adj}(V(\pi)/\det(V(\pi)) = V(\pi)^{-1}$.

21.2 By the determinant product rule,

$$\begin{aligned}
\log\det(H + tZ) &= \log\det(H^{1/2}(I + tH^{-1/2}ZH^{-1/2})H^{1/2}) \\
&= \log\det(H) + \log\det(I + tH^{-1/2}ZH^{-1/2}) \\
&= \log\det(H) + \sum_i \log(1 + t\lambda_i)\,,
\end{aligned}$$

where λ_i are the eigenvalues of $H^{-1/2}ZH^{-1/2}$. Since $\log(1 + t\lambda_i)$ is concave, their sum is also concave, proving that $t \mapsto \log\det(H + tZ)$ is concave.

21.3 Let $\mathcal{A}$ be a compact subset of $\mathbb{R}^d$ and $(\mathcal{A}_n)_n$ be a sequence of finite subsets with $\mathcal{A}_n \subset \mathcal{A}_{n+1}$ and $\operatorname{span}(\mathcal{A}_n) = \mathbb{R}^d$ and $\lim_{n\to\infty} d(\mathcal{A}, \mathcal{A}_n) = 0$ where d is the Hausdorff metric. Then let π_n be a G-optimal design for $\mathcal{A}_n$ with support of size at most $d(d+1)/2$ and $V_n = V(\pi_n)$. Given any $a \in \mathcal{A}$

we have

$$\|a\|_{V_n^{-1}} \le \min_{b\in\mathcal{A}_n} \left(\|a-b\|_{V_n^{-1}} + \|b\|_{V_n^{-1}} \right) \le \sqrt{d} + \min_{b\in\mathcal{A}_n} \|a-b\|_{V_n^{-1}} .$$

Let $W \in \mathbb{R}^{d\times d}$ be matrix with columns $w_1, \ldots, w_d$ in $\mathcal{A}_1$ that span $\mathbb{R}^d$. The operator norm of $V_n^{-1/2}$ is bounded by

$$\begin{aligned}
\|V_n^{-1/2}\| &= \|W^{-1} W V_n^{-1/2}\| \\
&\le \|W^{-1}\| \|V_n^{-1/2} W\| \\
&= \|W^{-1}\| \sup\left\{ \|Wx\|_{V_n^{-1}} : \|x\|_2 = 1 \right\} \\
&\le \|W^{-1}\| \sup\left\{ \sum_{i=1}^d |x_i| \|w_i\|_{V_n^{-1}} : \|x\|_2 = 1 \right\} \\
&\le \|W^{-1}\| \sqrt{d} \sup_{x:\|x\|_2=1} \sum_{i=1}^d x_i \\
&\le d\|W^{-1}\| .
\end{aligned}$$

Taking the limit as n tends to infinity shows that

$$\begin{aligned}
\limsup_{n\to\infty} \|a\|_{V_n^{-1}} &\le \sqrt{d} + \limsup_{n\to\infty} \min_{b\in\mathcal{A}} \|a-b\|_{V_n^{-1}} \\
&\le \sqrt{d} + d\|W^{-1}\| \limsup_{n\to\infty} \min_{b\in\mathcal{A}} \|a-b\|_2 \\
&\le \sqrt{d} .
\end{aligned}$$

Since $\|\cdot\|_{V_n^{-1}} : \mathcal{A} \to \mathbb{R}$ is continuous and $\mathcal{A}$ is compact it follows that

$$\limsup_{n\to\infty} \sup_{a\in\mathcal{A}} \|a\|^2_{V_n^{-1}} \le d .$$

Notice that π_n may be represented as a tuple of vector/probability pairs with at most $d(d+1)/2$ entries and where the vectors lie in $\mathcal{A}$. Since the set of all such tuples with the obvious topology forms a compact set it follows that (π_n) has a cluster point π^*, which represents a distribution on $\mathcal{A}$ with support at most $d(d+1)/2$. The previous display shows that $g(\pi^*) \le d$. The fact that $g(\pi^*) \ge d$ follows from the same argument as the proof of Theorem 21.1.

21.5 Let π be a Dirac at a and $\pi(t) = \pi^* + t(\pi^* - \pi)$. Since $\pi^*(a) > 0$ it follows for sufficiently small $t > 0$ that $\pi(t)$ is a distribution over $\mathcal{A}$. Because π^* is a minimizer of f,

$$0 \ge \frac{d}{dt} f(\pi(t))|_{t=0} = \langle \nabla f(\pi^*), \pi^* - \pi \rangle = d - \|a\|^2_{V(\pi)^{-1}} .$$

Rearranging shows that $\|a\|^2_{V(\pi)^{-1}} \geq d$. The other direction follows by Theorem 21.1.

Chapter 22 Stochastic Linear Bandits for Finitely Many Arms

Chapter 23 Stochastic Linear Bandits with Sparsity

23.2 The usual idea does the trick. Recall that $R_n = \sum_{i=1}^d R_{ni}$ where

$$R_{ni} = n|\theta_i| - \mathbb{E}\left[\sum_{t=1}^n A_{ti}\theta_i\right].$$

We proved that

$$R_{ni} \leq 3|\theta_i| + \frac{C\log(n)}{|\theta_i|}.$$

Clearly $R_{ni} \leq 2n|\theta_i|$. Let $\Delta > 0$ be a constant to be tuned later. Then

$$\begin{aligned} R_{ni} &\leq \sum_{i:|\theta_i|>\Delta} \left(3|\theta_i| + \frac{C\log(n)}{|\theta_i|}\right) + \sum_{i:|\theta_i|\in(0,\Delta)} 2n\Delta \\ &\leq 3\|\theta\|_1 + \frac{C\|\theta\|_0\log(n)}{\Delta} + 2\|\theta\|_0 n\Delta\,. \end{aligned}$$

Choosing $\Delta = \sqrt{\log(n)/n}$ completes the result.

Chapter 24 Minimax Lower Bounds for Stochastic Linear Bandits

24.1 Assume without loss of generality that $i = 1$ and let $\theta^{(-1)} \in \Theta^{p-1}$. The objective is to prove that

$$\frac{1}{|\Theta|}\sum_{\theta^{(1)}\in\Theta} R_{n1}(\theta) \geq \frac{\sqrt{kn}}{8}.$$

For $j \in [k]$ let $T_j(n) = \sum_{t=1}^n \mathbb{I}\{B_{t1} = j\}$ be the number of times base action j is played in the first bandit. Define $\psi_0 \in \mathbb{R}^d$ to be the vector with $\psi_0^{(-1)} = \theta^{(-1)}$ and $\psi_0^{(1)} = 0$. For $j \in [k]$ let $\psi_j \in \mathbb{R}^d$ be given by $\psi_j^{(-1)} = \theta^{(-1)}$ and $\psi_j^{(1)} = \Delta e_j$. Abbreviate $\mathbb{P}_j = \mathbb{P}_{\psi_j}$ and $\mathbb{E}_j[\cdot] = \mathbb{E}_{\mathbb{P}_j}[\cdot]$. With this notation, we have

$$\frac{1}{|\Theta|}\sum_{\theta^{(1)}\in\Theta} R_{n1}(\theta) = \frac{1}{k}\sum_{j=1}^k \Delta(n - \mathbb{E}_j[T_j(n)])\,. \tag{24.1}$$

Lemma 15.1 gives that

$$\mathrm{D}(\mathbb{P}_0, \mathbb{P}_j) = \frac{1}{2}\mathbb{E}_0\left[\sum_{t=1}^{n}\langle A_t, \psi_0 - \psi_j\rangle^2\right] = \frac{\Delta^2}{2}\mathbb{E}_0\left[T_j(n)\right] .$$

Choosing $\Delta = \sqrt{k/n}/2$ and applying Pinsker's inequality yields

$$\begin{aligned}
\sum_{j=1}^{k}\mathbb{E}_j[T_j(n)] &\le \sum_{j=1}^{k}\mathbb{E}_0[T_j(n)] + n\sum_{j=1}^{k}\sqrt{\frac{1}{2}\mathrm{D}(\mathbb{P}_0, \mathbb{P}_j)} \\
&= n + n\sum_{j=1}^{k}\sqrt{\frac{\Delta^2}{4}\mathbb{E}_0[T_j(n)]} \\
&\le n + n\sqrt{\frac{k\Delta^2}{4}\sum_{j=1}^{k}\mathbb{E}_0[T_j(n)]} && \text{(Cauchy-Schwarz)} \\
&= n + n\sqrt{\frac{k\Delta^2 n}{4}} \\
&\le 3nk/4 . && \text{(since } k \ge 2\text{)}
\end{aligned}$$

Combining the above display with Eq. (24.1) completes the proof:

$$\frac{1}{|\Theta|}\sum_{\theta^{(1)}\in\Theta} R_{n1}(\theta) = \frac{1}{k}\sum_{j=1}^{k}\Delta(n - \mathbb{E}_j[T_j(n)]) \ge \frac{n\Delta}{4} = \frac{1}{8}\sqrt{kn} .$$

Chapter 25 Asymptotic Lower Bounds for Stochastic Linear Bandits

25.3 For (a) let $\theta_1 = \Delta$ and $\theta_i = 0$ for $i > 1$ and let $\mathcal{A} = \{e_1, \ldots, e_{d-1}\}$. Then adding e_d increases the asymptotic regret. For (b) let $\theta_1 = \Delta$ and $\theta_i = 0$ for $1 < i < d$ and $\theta_d = 1$ and $\mathcal{A} = \{e_1, \ldots, e_{d-1}\}$. Then for small values of Δ adding e_d decreases the asymptotic regret.

Chapter 26 Foundations of Convex Analysis

26.1 Let $\mathbb{P}$ be the on the space on which X is defined. Following the hint, let $x_0 = \mathbb{E}[X] \in \mathbb{R}^d$. Then let $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ be such that $\langle a, x_0\rangle + b = f(x_0)$ and $\langle a, x\rangle + b \le f(x)$ for all $x \in \mathbb{R}^d$. The hyperplane $\{x : \langle a, x\rangle + b - f(x_0) = 0\}$ is guaranteed to exist by the supporting hyperplane theorem. Then

$$\int f(X)d\mathbb{P} \ge \int (\langle a, X\rangle + b)d\mathbb{P} = \langle a, x_0\rangle + b = f(x_0) = f(\mathbb{E}[X]) .$$

An alternative is of course to follow the ideas next to the picture in the main text. As you may recall that proof is given for the case when X is discrete. To extend the proof to the general case, one can use the 'standard machinery' of building up the integral from simple functions, but the resulting proof, originally due to Needham [1993], is much longer than what was given above.

26.2

(a) Using the definition,

$$\begin{aligned} f^{**}(x) &= \sup_{u\in\mathbb{R}^d} \langle x,u\rangle - f^*(u) \\ &= \sup_{u\in\mathbb{R}^d} \langle x,u\rangle - (\sup_{y\in\mathbb{R}^d} \langle y,u\rangle - f(y)) \\ &\leq \sup_{u\in\mathbb{R}^d} \langle x,u\rangle - (\langle x,u\rangle - f(x)) \\ &= f(x)\,. \end{aligned}$$

(b) We only need to show that $f^{**}(x) \geq f(x)$.

$$\begin{aligned} f^{**}(x) &= \sup_{u\in\mathbb{R}^d} \langle x,u\rangle - f^*(u) \\ &\geq \langle x,\nabla f(x)\rangle - f^*(\nabla f(x)) \\ &= \langle x,\nabla f(x)\rangle - (\sup_{y\in\mathbb{R}^d} \langle y,\nabla f(x)\rangle - f(y)) \\ &\geq \langle x,\nabla f(x)\rangle - (\sup_{y\in\mathbb{R}^d} \langle y,\nabla f(x)\rangle - f(x) - \langle y-x,\nabla f(x)\rangle) \\ &= f(x)\,, \end{aligned}$$

where in the second inequality we used the definition of convexity to ensure that $f(y) \geq f(x) + \langle y-x,\nabla f(x)\rangle$.

26.8

(a) Fix $u \in \mathbb{R}^d$. By definition $f^*(u) = \sup_x \langle x,u\rangle - f(x)$. To find this value we solve for x where the derivative of $\langle x,u\rangle - f(x)$ in x is equal to zero. As calculated before, $\nabla f(x) = \log(x)$. Thus, we need to find the solution to $u = \log(x)$, giving $x = \exp(u)$. Plugging this value, we get $f^*(u) = \langle \exp(u),u\rangle - f(\exp(u))$. Now, $f(\exp(u)) = \langle \exp(u), \log(\exp(u))\rangle - \langle \exp(u),\mathbf{1}\rangle = \langle \exp(u),u\rangle - \langle \exp(u),\mathbf{1}\rangle$. Hence, $f^*(u) = \langle \exp(u),\mathbf{1}\rangle$ and $\nabla f^*(u) = \exp(u)$.

(b) From our calculation, $\mathrm{dom}(\nabla f^*) = \mathbb{R}^d$.

(c) $D_{f^*}(u,v) = f^*(u) - f^*(v) - \langle \nabla f^*(v), u-v\rangle = \langle \exp(u) - \exp(v),\mathbf{1}\rangle - \langle \exp(v),u-v\rangle$.

(d) To check Part (a) of Theorem 26.6 note that $\nabla f(x) = \log(x)$ and $\nabla f^*(u) = \exp(u)$, which are indeed inverses of each other and their respective domains match that of $\mathrm{int}(\mathrm{dom}(f))$ and

int(dom(f^*)), respectively. To check Part (b) of Theorem 26.6, we calculate $D_{f^*}(\nabla f(y), \nabla f(x))$:

$$\begin{aligned} &D_{f^*}(\nabla f(y), \nabla f(x)) \\ &\quad = \langle \exp(\log(y)) - \exp(\log(x)), \mathbf{1} \rangle - \langle \exp(\log(x)), \log(y) - \log(x) \rangle \\ &\quad = \langle y - x, \mathbf{1} \rangle - \langle x, \log(y) - \log(x) \rangle \,, \end{aligned}$$

which is indeed equal to $D_f(x, y)$.

Chapter 27 Exp3 for Adversarial Linear Bandits

27.4 Let $x \in \mathbb{R}^d$. Then by the Cauchy-Schwarz inequality,

$$\|x\|_{A^{-1}}^2 = \langle x, A^{-1}x \rangle \leq \|x\|_{B^{-1}} \|A^{-1}x\|_B \leq \|x\|_{B^{-1}} \|A^{-1}x\|_A = \|x\|_{B^{-1}} \|x\|_{A^{-1}} \,.$$

Hence $\|x\|_{A^{-1}} \leq \|x\|_{B^{-1}}$ for all x, which completes the claim.

27.6

(a) A straightforward calculation shows that $\mathcal{L} = \left\{ y \in \mathbb{R}^d : \|y\|_V \leq 1 \right\}$. Let $Tx = V^{1/2}x$ and note that $T^{-1}\mathcal{L} = T\mathcal{A} = B = \left\{ u \in \mathbb{R}^d : \|u\|_2 \leq 1 \right\}$. Then let $\mathcal{U}$ be an ε-cover of B with respect to $\|\cdot\|_2$ with $|\mathcal{U}| \leq (3/\varepsilon)^d$ and $\mathcal{C} = T^{-1}\mathcal{U}$. Given $x \in \mathcal{A}$ let $u = Tx$ and $u' \in \mathcal{U}$ be such that $\|u - u'\|_2 \leq \varepsilon$ and $x' = T^{-1}u'$. Then

$$\|x - x'\| = \sup_{y \in \mathcal{L}} \langle x - x', y \rangle = \sup_{y \in \mathcal{L}} \langle T^{-1}u - T^{-1}u', y \rangle = \sup_{y \in \mathcal{L}} \langle u - u', T^{-1}y \rangle \leq \varepsilon \,.$$

(b) Notice that $\mathcal{L}$ is convex, symmetric, bounded and $\operatorname{span}(\mathcal{L}) = \mathbb{R}^d$. Let $\mathcal{E} = \left\{ y \in \mathbb{R}^d : \|y\|_V \leq 1 \right\}$ be the ellipsoid of maximum volume contained by $\operatorname{cl}(\mathcal{L})$. Then let

$$\mathcal{E}_* = \left\{ y \in \mathbb{R}^d : \|y\|_{\mathcal{E}} \leq 1 \right\} = \left\{ y \in \mathbb{R}^d : \|y\|_{V^{-1}} \leq 1 \right\} \,,$$

which satisfies $\mathcal{A} \subseteq \mathcal{E}_*$. Since $\operatorname{span}(\mathcal{L}) = \mathbb{R}^d$ the matrix V^{-1} is positive definite. By the previous result there exists a $\bar{\mathcal{C}} \subset \mathcal{E}_*$ of size at most $(3d/\varepsilon)^d$ such that

$$\sup_{x \in \mathcal{E}_*} \inf_{x' \in \bar{\mathcal{C}}} \|x - x'\|_{\mathcal{E}} \leq \varepsilon / d \,.$$

Using the fact that $\mathcal{L} \subseteq d\mathcal{E}$ we have

$$\sup_{x \in \mathcal{E}_*} \inf_{x' \in \bar{\mathcal{C}}} \|x - x'\|_{\mathcal{L}} \leq \varepsilon \,.$$

We are nearly done. The problem is that $\bar{\mathcal{C}}$ may contain elements not in $\mathcal{A}$. To resolve this issue

let $\mathcal{C} = \{\Pi(x) : x \in \bar{\mathcal{C}}\}$ where $\Pi(x) \in \operatorname{argmin}_{x' \in \mathcal{A}} \|x - x'\|_{\mathcal{E}}$. Then note that

$$\|x - \Pi(x')\|_{\mathcal{L}} \leq d\|x - \Pi(x')\|_{\mathcal{E}} \leq d\|x - x'\|_{\mathcal{E}}\,.$$

(c) Let $\bar{\mathcal{C}}$ be an $\varepsilon/2$-cover of $\mathrm{cl}(\mathrm{co}(\mathcal{A}))$, which by the previous part has size at most $(6d/\varepsilon)^d$. The result follows by choosing $\mathcal{C} = \left\{\Pi(x) : x \in \bar{\mathcal{C}}\right\}$ where Π is the projection onto $\mathcal{A}$ with respect to $\|\cdot\|_{\mathcal{E}}$ where $\mathcal{E}$ is the maximum volume ellipsoid contained by $\mathrm{cl}(\mathrm{co}(\mathcal{A}))$.

27.9 Define $(W_t)_{t=0}^n$ by

$$W_t = \int_{\mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t} \hat{Y}_s(a)\right) da\,,$$

which means that $W_0 = \mathrm{vol}(\mathcal{A})$ and

$$W_n = \mathrm{vol}(\mathcal{A}) \prod_{t=0}^{n-1} \frac{W_{t+1}}{W_t}\,.$$

Following the proof of Theorem 11.1,

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \int_{\mathcal{A}} \exp(-\eta \hat{Y}_t(a)) dP_t(a) \\
&\leq \int_{\mathcal{A}} \left(1 - \eta \hat{Y}_t(a) + \eta^2 \hat{Y}_t(a)^2\right) dP_t(a) \\
&\leq \exp\left(-\eta \int_{\mathcal{A}} \hat{Y}_t(a) dP_t(a) + \eta^2 \int_{\mathcal{A}} \hat{Y}_t(a)^2 dP_t(a)\right)\,.
\end{aligned}$$

Therefore

$$W_n \leq \log(\mathrm{vol}(\mathcal{A})) - \eta \sum_{t=1}^{n} \int_{\mathcal{A}} \hat{Y}_t(a) dP_t(a) + \eta^2 \sum_{t=1}^{n} \int_{\mathcal{A}} \hat{Y}_t(a)^2 dP_t(a)\,.$$

Taking the expectation, substituting the definition of W_n and rearranging shows that

$$\begin{aligned}
R_n \leq 2n\gamma + \frac{1}{\eta}\mathbb{E}\left[\log\left(\frac{\mathrm{vol}(\mathcal{A})}{\int \exp\left(-\eta \sum_{t=1}^n (\hat{Y}_t(a) - \hat{Y}_t(a^*))\right) da}\right)\right] \\
+ \eta \mathbb{E}\left[\sum_{t=1}^{n} \int_{\mathcal{A}} \hat{Y}_t(a)^2 dP_t(a)\right]\,,
\end{aligned}$$

where we used the fact that

$$-\eta \sum_{t=1}^{n} \hat{Y}_t(a^*) = \log\left(\frac{1}{\exp\left(\eta \sum_{t=1}^n \hat{Y}_t(a^*)\right)}\right)\,.$$

The result is completed by the same argument as in the proof of Theorem 27.1 to bound

$$\eta \mathbb{E}\left[\sum_{t=1}^{n} \int_{\mathcal{A}} \hat{Y}_t(a)^2 dP_t(a)\right] \leq \eta d n \,.$$

27.11 Throughout $\|\cdot\| = \|\cdot\|_2$ is the standard euclidean norm. By translating $\mathcal{K}$ we may assume that $x^* = \mathbf{0}$. Furthermore, suppose we can prove the result for u with $\|u\| = 1$, then

$$\begin{aligned}
\frac{\mathrm{vol}(\mathcal{K})}{\int_{\mathcal{K}} \exp(-\langle x, u\rangle) dx} &= \frac{\mathrm{vol}(\mathcal{K})}{\int_{\mathcal{K}} \exp(-\langle x\|u\|, u/\|u\|\rangle) dx} \\
&= \frac{\|u\|^d \mathrm{vol}(\mathcal{K})}{\int_{\|u\|\mathcal{K}} \exp(-\langle y, u/\|u\|\rangle) dy} \\
&= \frac{\mathrm{vol}(\|u\|\mathcal{K})}{\int_{\|u\|\mathcal{K}} \exp(-\langle y, u/\|u\|\rangle) dy} \\
&\leq 1 + \max\left\{0, d \log\left(\sup_{x,y \in \|u\|\mathcal{K}} \langle x - y, u/\|u\|\rangle\right)\right\} \\
&= 1 + \max\left\{0, d \log\left(\sup_{x,y \in \mathcal{K}} \langle x - y, u\rangle\right)\right\} .
\end{aligned}$$

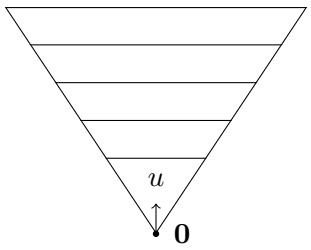
From now on we assume that $\|u\| = 1$ and let $f(t)$ be the volume with respect to the $(d-1)$-form of $\mathcal{K}_t = \{x \in \mathcal{K} : \langle x, u\rangle = t\}$. Let $\alpha = \sup_{x \in \mathcal{K}} \langle x, u\rangle$. Then

$$\frac{\mathrm{vol}(\mathcal{K})}{\int_{\mathcal{K}} \exp(-\langle x, u\rangle) dx} = \frac{\int_0^\alpha f(t) dt}{\int_0^\alpha f(t) \exp(-t) dt} \,.$$

A rearrangement argument shows that the convex body maximizing this quantity is $\mathcal{K} = \{\beta x : x \in \mathcal{K}_\alpha, \beta \in [0,1]\}$ where $f(\alpha) > 0$. Therefore

$$\begin{aligned}
\frac{\mathrm{vol}(\mathcal{K})}{\int_{\mathcal{K}} \exp(-\langle x, u\rangle) dx} &\leq \frac{\int_0^\alpha t^{d-1}}{\int_0^\alpha t^{d-1} \exp(-t) dt} \\
&= \frac{\alpha^d}{d(\Gamma(d) - \Gamma(d, \alpha))} \\
&\leq \max\left\{e, e\alpha^d\right\} ,
\end{aligned}$$

where we used the fact that for $\mathcal{K}$ the $(d-1)$-volume of the slices is $f(t) = (t f(\alpha))^{d-1}$.



The whole triangle is $\mathcal{K}$ with x^* at the bottom corner. The thin lines represent $\mathcal{K}_t$ for different values of t, which $(d-1)$-dimensional subsets of $\mathcal{K}$ that lie in affine spaces with normal vector u.

Chapter 28 Follow the Regularized Leader and Mirror Descent

28.3 The first step is the same as the proof of Theorem 28.4.

$$R_n(a) = \sum_{t=1}^{n} \langle a_t - a, y_t \rangle = \sum_{t=1}^{n} \langle a_t - a_{t+1}, y_t \rangle + \sum_{t=1}^{n} \langle a_{t+1} - a, y_t \rangle \,.$$

Next let $\Phi_t(a) = F(a)/\eta + \sum_{s=1}^{t} \langle a, y_s \rangle$ so that

$$\begin{aligned}
\sum_{t=1}^{n} \langle a_{t+1} - a, y_t \rangle &= \sum_{t=1}^{n} \langle a_{t+1}, y_t \rangle - \Phi_n(a) + \frac{F(a)}{\eta} \\
&= \sum_{t=1}^{n} \left(\Phi_t(a_{t+1}) - \Phi_{t-1}(a_{t+1}) \right) - \Phi_n(a) + \frac{F(a)}{\eta} \\
&= -\Phi_0(a_1) + \sum_{t=0}^{n-1} (\Phi_t(a_{t+1}) - \Phi_t(a_{t+2})) + \Phi_n(a_{n+1}) - \Phi_n(a) + \frac{F(a)}{\eta} \\
&\leq \frac{F(a) - F(a_1)}{\eta} + \sum_{t=0}^{n-1} (\Phi_t(a_{t+1}) - \Phi_t(a_{t+2})) \,. \qquad (28.1)
\end{aligned}$$

Now $D_{\Phi_t}(a,b) = \frac{1}{\eta} D_F(a,b)$. Therefore

$$\begin{aligned}
\Phi_t(a_{t+1}) - \Phi_t(a_{t+2}) &= \langle \nabla \Phi_t(a_{t+1}), a_{t+1} - a_{t+2} \rangle - \frac{1}{\eta} D_F(a_{t+2}, a_{t+1}) \\
&\leq -\frac{1}{\eta} D_F(a_{t+2}, a_{t+1}) \,.
\end{aligned}$$

Substituting this into Eq. (28.1) completes the proof.

28.8

(a) For the first relation we have $\nabla F^*(x)_i = \exp(x)_i$ and $F^*(x) = \sum_{i=1}^{k} \exp(x_i)$. Then $D_F(P_t, \tilde{P}_{t+1}) = D_{F^*}(\nabla F^*(\tilde{P}_{t+1}), \nabla F^*(P_t)) = \sum_{i=1}^{k} P_{ti}(\exp(-\eta \hat{Y}_{ti}) - 1 + \eta \hat{Y}_{ti})$. The second relation follows from the inequality $\exp(x) \leq 1 + x + x^2/2$ for $x \leq 0$.

(b) Using part (a) we have

$$\begin{aligned}\frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^{n} D_F(P_t, \tilde{P}_{t+1})\right] &\leq \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k} P_{ti}\hat{Y}_{ti}^2\right] \\ &\leq \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k} \frac{\mathbb{I}\{A_t = i\}}{P_{ti}}\right] \\ &= \frac{\eta n k}{2}\,.\end{aligned}$$

(c) Simple calculus shows that for $p \in \mathcal{P}_{k-1}$, $F(p) \geq -\log(k) - 1$ and $F(p) \leq -1$ is obvious. Therefore $\operatorname{diam}_F(\mathcal{A}) = \max_{p,q\in\mathcal{A}} F(p) - F(q) \leq \log(k)$.

(d) By the previous exercise Exp3 chooses A_t sampled from P_t. Then applying Theorem 28.4 and parts (b) and (c) and choosing $\eta = \sqrt{\log(k)/(2nk)}$ yields the result.

28.9 Abbreviate $D(x, y) = D_F(x, y)$. By the definition of $\tilde{a}_{t+1}$ and the first-order optimality conditions we have $\eta_t y_t = \nabla F(a_t) - \nabla F(\tilde{a}_{t+1})$. Therefore

$$\begin{aligned}\langle a_t - a, y_t\rangle &= \frac{1}{\eta_t}\langle a_t - a, \nabla F(a_t) - \nabla F(\tilde{a}_{t+1})\rangle \\ &= \frac{1}{\eta_t}\left(-\langle a - a_t, \nabla F(a_t)\rangle - \langle a_t - \tilde{a}_{t+1}, \nabla F(\tilde{a}_{t+1})\rangle + \langle a - \tilde{a}_{t+1}, \nabla F(\tilde{a}_{t+1})\rangle\right) \\ &= \frac{1}{\eta_t}\left(D(a, a_t) - D(a, \tilde{a}_{t+1}) + D(a_t, \tilde{a}_{t+1})\right)\,.\end{aligned}$$

Summing completes the proof. For the second part use the fact that $D(a, \tilde{a}_{t+1}) \geq D(a, a_{t+1})$.

28.10

(a) We use the same argument as the solution to Exercise 28.3. First,

$$R_n(a) = \sum_{t=1}^{n}\langle a_t - a, y_t\rangle = \sum_{t=1}^{n}\langle a_t - a_{t+1}, y_t\rangle + \sum_{t=1}^{n}\langle a_{t+1} - a, y_t\rangle\,.$$

The next step also mirrors that in Exercise 28.3, but now we have to keep track of the changing

potentials:

$$\begin{aligned}
&\sum_{t=1}^{n}\langle a_{t+1}-a,y_t\rangle = \sum_{t=1}^{n}\langle a_{t+1},y_t\rangle - \Phi_{n+1}(a) + F_{n+1}(a)\\
&= \sum_{t=1}^{n}\left(\Phi_{t+1}(a_{t+1}) - \Phi_t(a_{t+1})\right) + \sum_{t=1}^{n}\left(F_t(a_{t+1}) - F_{t+1}(a_{t+1})\right) - \Phi_{n+1}(a) + F_{n+1}(a)\\
&= -\Phi_1(a_1) + \sum_{t=0}^{n-1}(\Phi_{t+1}(a_{t+1}) - \Phi_{t+1}(a_{t+2})) + \Phi_{n+1}(a_{n+1}) - \Phi_{n+1}(a)\\
&\qquad + F_{n+1}(a) + \sum_{t=1}^{n}\left(F_t(a_{t+1}) - F_{t+1}(a_{t+1})\right)\\
&\leq F_{n+1}(a) - F_1(a_1) + \sum_{t=0}^{n-1}(\Phi_{t+1}(a_{t+1}) - \Phi_{t+1}(a_{t+2})) + \sum_{t=1}^{n}\left(F_t(a_{t+1}) - F_{t+1}(a_{t+1})\right)\,.
\end{aligned}$$

Now $D_{\Phi_t}(a,b) = D_{F_t}(a,b)$. Therefore

$$\begin{aligned}
\Phi_{t+1}(a_{t+1}) - \Phi_{t+1}(a_{t+2}) &= \langle \nabla\Phi_{t+1}(a_{t+1}), a_{t+1}-a_{t+2}\rangle - D_{F_{t+1}}(a_{t+2},a_{t+1})\\
&\leq -D_{F_{t+1}}(a_{t+2},a_{t+1})\,,
\end{aligned}$$

which combined with the previous big display completes the proof.

(b) Note that adding a constant to the potential does not change the policy or the Bregman divergence. Applying the previous part with $F_t(a) = (F(a) - \min_{b\in\mathcal{A}} F(b))/\eta_t$ leads immediately to the result.

28.11

(a) Apply your solution to Exercise 26.10.

(b) Since $\hat{Y}_t$ is unbiased we have

$$R_n = \max_{i\in[k]} \mathbb{E}\left[\sum_{t=1}^{n}(y_{tA_t} - y_{ti})\right] = \max_{i\in[k]} \mathbb{E}\left[\sum_{t=1}^{n}\langle P_t - P, \hat{Y}_t\rangle\right]\,.$$

Then apply the result from Exercise 28.10 combined with the fact that $\mathrm{diam}_F(\mathcal{A}) = \log(k)$.

(c) Consider two cases. First, if $P_{t+1,A_t} \geq P_{tA_t}$, then

$$\langle P_t - P_{t+1}, \hat{Y}_t\rangle = (P_{tA_t} - P_{t+1,A_t})\hat{Y}_{tA_t} \leq 0\,.$$

On the other hand, if $P_{t+1,A_t} \leq P_{tA_t}$, then by Theorem 26.12 with $H = \nabla^2 f(q) = \mathrm{diag}(1/q)$ for some $q \in [P_t, P_{t+1}]$ we have

$$\langle P_t - P_{t+1}, \hat{Y}_t\rangle - \frac{D_F(P_{t+1},P_t)}{\eta_t} \leq \frac{\eta_t}{2}\|\hat{Y}_t\|^2_{H^{-1}}\,,$$

Since $P_{t+1,A_t} \le P_{tA_t}$ we have

$$\frac{\eta_t}{2}\|\hat{Y}_t\|^2_{H^{-1}} = \frac{\eta_t q_{A_t}\hat{Y}^2_{tA_t}}{2} \le \frac{\eta_t P_{tA_t}\hat{Y}^2_{tA_t}}{2} \le \frac{\eta_t y^2_{tA_t}}{2P_{tA_t}}\,.$$

Therefore

$$\langle P_t - P_{t+1}, \hat{Y}_t\rangle - \frac{D_F(P_{t+1}, P_t)}{\eta_t} \le \frac{\eta_t y^2_{tA_t}}{2P_{tA_t}} \le \frac{\eta_t}{2P_{tA_t}}\,.$$

(d) Continuing from the previous part and using the fact that $\mathbb{E}[1/P_{tA_t}] = k$ shows that

$$R_n \le \mathbb{E}\left[\frac{\log(k)}{\eta_n} + \frac{1}{2}\sum_{t=1}^n \frac{\eta_t}{P_{tA_t}}\right] = \frac{\log(k)}{\eta_n} + \frac{k}{2}\sum_{t=1}^n \eta_t\,.$$

(e) Choose $\eta_t = \sqrt{\frac{\log(k)}{kt}}$ and use the fact that $\sum_{t=1}^n \sqrt{1/t} \le 2\sqrt{n}$.

28.12

(a) The result is obvious for any algorithm when $n < k$. Assume for the remainder that $n \ge k$. The learning rate is chosen to be

$$\eta_t = \sqrt{\frac{k\log(n/k)}{1 + \sum_{s=1}^{t-1} y^2_{tA_t}}}\,,$$

which is obviously decreasing. By noting that $\int f'(x)/\sqrt{f(x)}\,dx = 2\sqrt{f(x)}$ and making a simple approximation,

$$\sum_{t=1}^n \eta_t y^2_{tA_t} \le 2\sqrt{k\sum_{t=1}^n y^2_{tA_t}\log(n/k)}\,. \tag{28.2}$$

Define $R_n(p) = \sum_{t=1}^n \langle P_t - p, \hat{Y}_t\rangle$. Then

$$R_n = \sup_{p\in\mathcal{P}_{k-1}} \mathbb{E}[R_n(p)] \le k + \sup_{p\in[1/n,1]^k\cap\mathcal{P}_{k-1}} \mathbb{E}[R_n(p)]\,. \tag{28.3}$$

For the remainder of the proof let $p \in [1/n, 1] \cap \mathcal{P}_{k-1}$ be arbitrary. Notice that $F(p) - \min_{q\in\mathcal{P}_{k-1}} F(q) \le k\log(n/k)$. By the result in Exercise 28.10,

$$R_n(p) \le \frac{k\log(n/k)}{\eta_n} + \sum_{t=1}^n \langle P_t - P_{t+1}, \hat{Y}_t\rangle - \frac{D_F(P_{t+1}, P_t)}{\eta_t}\,, \tag{28.4}$$

If $P_{t+1,A_t} \ge P_{tA_t}$, then $\langle P_t - P_{t+1}, \hat{Y}_t\rangle \le 0$. Now suppose that $P_{t+1,A_t} \le P_{tA_t}$. By Theorem 26.11

there exists a $\xi \in [P_t, P_{t+1}]$ such that

$$\begin{aligned}\langle P_t - P_{t+1}, \hat{Y}_t \rangle - D_{F_{t-1}}(P_{t+1}, P_t) &\leq \frac{\eta_t}{2} \|\hat{Y}_t\|^2_{\nabla^2 F(\xi)^{-1}} \\ &= \frac{\eta_t}{2} \xi^2_{A_t} \hat{Y}^2_{tA_t} \leq \frac{\eta_t}{2} P^2_{tA_t} \hat{Y}^2_{tA_t} = \frac{\eta_t y^2_{tA_t}}{2}\,.\end{aligned}$$

By the definition of η_n and Eq. (28.2),

$$\begin{aligned}R_n(p) &\leq \frac{k \log(n/k)}{\eta_n} + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^n \eta_t y^2_{tA_t}\right] \\ &\leq 2\sqrt{k\left(1 + \sum_{t=1}^{n-1} y^2_{tA_t}\right)\log(n/k)}\,.\end{aligned}$$

Therefore by Eq. (28.3),

$$\begin{aligned}R_n &\leq k + 2\mathbb{E}\left[\sqrt{k\left(1 + \sum_{t=1}^{n-1} y^2_{tA_t}\right)\log(n/k)}\right] \\ &\leq k + 2\sqrt{k\left(1 + \mathbb{E}\left[\sum_{t=1}^{n-1} y^2_{tA_t}\right]\right)\log(n/k)}\,,\end{aligned}$$

where the second line follows from Jensen's inequality.

(b) Combining the previous result with the fact that $y_t \in [0,1]^k$ shows that

$$\begin{aligned}R_n &\leq k + 2\sqrt{k\left(1 + \mathbb{E}\left[\sum_{t=1}^n y_{tA_t}\right]\right)\log(n/k)} \\ &= k + 2\sqrt{k\left(1 + R_n + \min_{a \in [k]} \sum_{t=1}^n y_{ta}\right)\log(n/k)}\,.\end{aligned}$$

Solving the quadratic shows that for a suitably large universal constant C,

$$R_n \leq k + C\sqrt{k\left(1 + \min_{a \in [k]} \sum_{t=1}^n y_{ta}\right)\log(n/k)}\,.$$

28.13 The first parts are mechanical and are skipped.

(e) By Part (c), Theorem 28.5 and Theorem 26.12,

$$R_n \leq \frac{2\sqrt{k}}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^n \|\hat{Y}_t\|^2_{\nabla^2 F(Z_t)^{-1}}\right]\,, \tag{28.5}$$

where $Z_t \in \mathcal{P}_{k-1} = \alpha P_t + (1-\alpha)P_{t+1}$ for some $\alpha \in [0,1]$. Then

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k} \|\hat{Y}_t\|^2_{\nabla^2 F(Z_t)^{-1}}\right] &= 2\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k} \frac{A_{ti}y_{ti}^2}{P_{ti}^2} Z_{ti}^{3/2}\right] \\
&\le 2\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k} \frac{A_{ti}}{P_{ti}^{1/2}}\right] \\
&= 2\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k} P_{ti}^{1/2}\right] \\
&\le 2n\sqrt{k}\,,
\end{aligned}$$

where the first inequality follows from Part (b) and the second from Cauchy-Schwarz inequality. The result follows by substituting the above display into Eq. (28.5) and choosing $\eta = \sqrt{2/n}$.

28.14

(a) Following the suggestion in the hint let F be the negentropy potential and

$$\begin{aligned}
x_t &= \operatorname{argmin}_{x\in X}\left(F(x) + \eta\sum_{s=1}^{t-1} f(x, y_s)\right) \\
y_t &= \operatorname{argmin}_{y\in Y}\left(F(y) - \eta\sum_{s=1}^{t-1} f(x_s, y)\right)\,.
\end{aligned}$$

Then let $\varepsilon_d(n) = \sqrt{\frac{2\log(d)}{n}}$. By Proposition 28.7,

$$\begin{aligned}
\min_{x\in X}\max_{y\in Y} f(x,y) &\le \max_{y\in Y} f(\bar{x}_n, y) \\
&= \max_{y\in Y}\frac{1}{n}\sum_{t=1}^{n} f(x_t, y) \\
&\le \frac{1}{n}\sum_{t=1}^{n} f(x_t, y_t) + \varepsilon_j(n) \\
&\le \frac{1}{n}\min_{x\in X}\sum_{t=1}^{n} f(x, y_t) + \varepsilon_j(n) + \varepsilon_k(n) \\
&= \min_{x\in X} f(x, \bar{y}_n) + \varepsilon_j(n) + \varepsilon_k(n) \\
&\le \max_{y\in Y}\min_{x\in X} f(x,y) + \varepsilon_j(n) + \varepsilon_k(n)\,.
\end{aligned}$$

Taking the limit as n tends to infinity shows that

$$\min_{x\in X}\max_{y\in Y} f(x,y) \le \max_{y\in Y}\min_{x\in X} f(x,y)\,.$$

Since the other direction holds trivially, equality holds.

(b) Following a similar plan. Let $F(x) = \frac{1}{2}\|x\|_2^2$ and $g_s(x) = f(x, y_s)$ and $h_s(y) = f(x_s, y)$. Then define

$$x_t = \operatorname{argmin}_{x\in X}\left(F(x) + \eta\sum_{s=1}^{t-1}\langle x, \nabla g_s(x_s)\rangle\right)$$
$$y_t = \operatorname{argmin}_{y\in Y}\left(F(y) - \eta\sum_{s=1}^{t-1}\langle y, \nabla h_s(y_s)\rangle\right).$$

Let $G = \sup_{x\in X, y\in Y}\|\nabla f(x,y)\|_2$ and $B = \sup_{z\in X\cup Y}\|z\|_2$. Then let $\varepsilon(n) = GB\sqrt{1/n}$. A straightforward generalization of the above argument and the analysis in Proposition 28.6 shows that

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} f(x_t, y_t) &= \frac{1}{n}\sum_{t=1}^{n} g_t(x_t)\\
&= \min_{x\in X}\frac{1}{n}\left(\sum_{t=1}^{n} g_t(x) + \sum_{t=1}^{n}(g_t(x_t) - g_t(x))\right)\\
&\le \min_{x\in X}\frac{1}{n}\left(\sum_{t=1}^{n} g_t(x) + \frac{1}{n}\sum_{t=1}^{n}\langle x_t - x, \nabla g_t(x_t)\rangle\right)\\
&\le \min_{x\in X}\frac{1}{n}\sum_{t=1}^{n} g_t(x) + \varepsilon(n)\\
&= \min_{x\in X}\frac{1}{n}\sum_{t=1}^{n} f(x, y_t) + \varepsilon(n)\\
&\le \min_{x\in X} f(x, \bar{y}_n) + \varepsilon(n).
\end{aligned}$$

In the same manner,

$$\max_{y\in Y} f(\bar{x}_n, y) \le \frac{1}{n}\sum_{t=1}^{n} f(x_t, y_t) + \varepsilon(n).$$

Hence

$$\begin{aligned}
\min_{x\in X}\max_{y\in Y} f(x,y) &\le \max_{y\in Y} f(\bar{x}_n, y)\\
&\le \min_{x\in X} f(x, \bar{y}_n) + 2\varepsilon(n)\\
&\le \max_{y\in Y}\min_{x\in X} f(x,y) + 2\varepsilon(n),.
\end{aligned}$$

And the result is again completed by taking the limit as n tends to infinity.

In both cases the pair of average iterates $(\bar{x}_n, \bar{y}_n)$ has a cluster point that is a saddle point of $f(\cdot,\cdot)$. In general the iterates (x_n, y_n) do not have a cluster point that is a saddle point.

28.15 Let $X = Y = \mathbb{R}$ and $f(x, y) = x + y$. Clearly X and Y are convex topological vector spaces and f is linear and linear in both arguments. Then $\inf_{x \in X} \sup_{y \in Y} f(x, y) = \infty$ and $\sup_{y \in Y} \inf_{x \in X} f(x, y) = -\infty$.

Chapter 29 The Relation Between Adversarial and Stochastic Linear Bandits

29.2 First we check that $\hat{\theta}_t = dE_t A_t Y_t / (1 - \|\bar{A}_t\|_2)$ is appropriately bounded.

$$\eta \|\hat{\theta}_t\|_2 = \frac{\eta d E_t \|A_t\|_2 |Y_t|}{1 - \|\bar{A}_t\|_2} \leq \frac{\eta d}{1 - \|\bar{A}_t\|_2} \leq \frac{1}{2}\,,$$

where the last step holds by choosing $1 - r = 2\eta d$. All of the steps in the proof of Theorem 28.11 are the same until the expectation of the dual norm of $\hat{\theta}_t$. Then

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\theta}_t\|^2_{\nabla F(Z_t)^{-1}}\right] &\leq \mathbb{E}\left[(1 - \|Z_t\|_2)\|\hat{\theta}_t\|^2\right] \\
&= d^2 \mathbb{E}\left[\frac{(1 - \|Z_t\|_2) E_t Y_t^2}{(1 - \|\bar{A}_t\|_2)^2}\right] \\
&\leq 2d^2\,.
\end{aligned}$$

This last inequality is where things have changed, with the d becoming a d^2. From this we conclude that

$$R_n \leq \frac{1}{\eta} \log\left(\frac{1}{2\eta d}\right) + (1 - r)n + \eta n d^2 \leq \frac{1}{\eta} \log\left(\frac{1}{2\eta d}\right) + 2\eta n d + \eta n d^2\,.$$

And the result follows by tuning η.

29.4

(a) Let $a^* = \operatorname{argmin}_{a \in \mathcal{A}} \ell(a)$ be the optimal action. Then

$$\begin{aligned}
R_n &= \mathbb{E}\left[\sum_{t=1}^n \ell(A_t) - \ell(a^*)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^n \langle A_t - a^*, \theta \rangle\right] + 2n\varepsilon \\
&= \mathbb{E}\left[\sum_{t=1}^n \langle \bar{A}_t - a^*, \theta \rangle\right] + 2n\varepsilon\,. \qquad (29.1)
\end{aligned}$$

The estimator is $\hat{\theta}_t = dE_tA_tY_t/(1 - \|\bar{A}_t\|_2)$, which is no longer unbiased. Then

$$\begin{aligned}\mathbb{E}[\hat{\theta}_t \,|\, \mathcal{F}_{t-1}] &= \mathbb{E}\left[\frac{dE_tA_tY_t}{1 - \|\bar{A}_t\|_2} \,\middle|\, \mathcal{F}_{t-1}\right] \\ &= d\mathbb{E}\left[\frac{E_tA_t(\langle A_t, \theta\rangle + \eta_t + \varepsilon(A_t))}{1 - \|\bar{A}_t\|_2} \,\middle|\, \mathcal{F}_{t-1}\right] \\ &= \theta + \sum_{i=1}^{d} \varepsilon(e_i)e_i\,,\end{aligned}$$

which means that

$$\begin{aligned}\mathbb{E}[\langle \bar{A}_t - a^*, \theta\rangle] &\le \mathbb{E}[\langle \bar{A}_t - a^*, \hat{\theta}_t\rangle] + \varepsilon\mathbb{E}[\|\bar{A}_t - a^*\|_1\|\mathbf{1}\|_\infty] \\ &\le \mathbb{E}[\langle \bar{A}_t - a^*, \hat{\theta}_t\rangle] + 2\varepsilon\,.\end{aligned}$$

Combining with Eq. (29.1) shows that

$$R_n \le \mathbb{E}\left[\sum_{t=1}^{n}\langle \bar{A}_t - a^*, \hat{\theta}_t\rangle\right] + 2\varepsilon n\,.$$

Then we need to check that $\eta\|Y_t\|_2 = \eta\|dE_tA_tY_t/(1 - \|\bar{A}_t\|_2)\|_2 \le \eta d/(1-r) \le 1/2$. Now proceed as in Exercise 29.2.

(b) When $\varepsilon(a) = 0$ for all a, the lower bound is $\Omega(d\sqrt{n})$. Now add a spike $\varepsilon(a) = -\varepsilon$ in the vicinity of the optimal arm. Since $\mathcal{A}$ is continuous, the learner will almost surely never identify the 'needle' and hence its regret is $\Omega(d\sqrt{n} + \varepsilon n)$.

Chapter 30 Combinatorial Bandits

30.2

(b) Using independence of $(X_j)_{j=1}^d$ shows that almost surely,

$$\begin{aligned}\mathbb{E}[M_j \,|\, M_{j-1}] &= M_{j-1}\int_0^{M_{j-1}} \exp(-x)dx + \int_{M_{j-1}}^{\infty} x\exp(-x)dx \\ &= M_{j-1} + \exp(-M_{j-1})\,.\end{aligned}$$

Taking the expectation of both sides yields the result.

(c) The base case when $j = 1$ is immediate. For $j \ge 2$,

$$\mathbb{E}[\exp(-aM_j)] = \mathbb{E}[\exp(-aM_{j-1})] - \frac{a}{a+1}\mathbb{E}[\exp(-(a+1)M_{j-1}]\,.$$

Therefore by induction it follows that

$$\mathbb{E}[\exp(-aM_j)] = \frac{a!}{\prod_{b=1}^{a}(j+b)}\,.$$

(d) Combining (b) and (c) shows that for $j \geq 2$,

$$\mathbb{E}[M_j] = \mathbb{E}[M_{j-1}] + \mathbb{E}[\exp(-M_{j-1})] = \mathbb{E}[M_{j-1}] + \frac{1}{j}\,.$$

The result follow by induction.

Chapter 31 Non-Stationary Bandits

31.1 As suggested, Exp4 is used with each element of Γ_{nm} identified with one expert. Consider an arbitrary enumeration of $\Gamma_{nm} = \{a^{(1)}, \ldots, a^{(G)}\}$ where $G = |\Gamma_{nm}|$. The predictions of expert $g \in [G]$ for round $t \in [n]$ encoded as a probability distribution over $[k]$ (as required by the prediction-with-expert-advice framework) is $E_{g,j}^t = \mathbb{I}\left\{a_t^{(g)} = j\right\}$, $j \in [k]$. The expected regret of Exp4 when used with these experts is

$$R_n^{\text{experts}} = \mathbb{E}\left[\sum_{t=1}^{n} y_{tA_t} - \min_{g\in[G]} \sum_{t=1}^{n} E_g^{(t)} y_t\right]\,,$$

where compared to Chapter 18 we switched to losses. By definition,

$$\sum_{t=1}^{n} E_g^{(t)} y_t = \sum_{t=1}^{n} y_{t,a_t^{(g)}}$$

and hence

$$R_n^{\text{experts}} = R_{nm}\,.$$

Thus, Theorem 18.1 indeed proves (31.1). To prove (31.2) it remains to show that $G = |\Gamma_{nm}| \leq Cm\log(kn/m)$. For this note that $G = \sum_{s=1}^{m} G_{n,s}^*$ where $G_{n,s}^*$ is the number of sequences from $[k]^n$ that switch exactly $s-1$ times. When $m - 1 \leq n/2$, a crude upper bound on G is mG_{nm}^*. For $s = 1$, $G_{n,s}^* = k$. For $s > 1$, a sequence with $s-1$ switches is determined by the location of the switches, and the identity of the action taken in each segment where the action does not change. The possible switch locations are of the form $(t, t+1)$ with $t = 1, \ldots, n-1$. Thus the number of these locations is $n-1$, of which, we need to choose $s-1$. There are $\binom{n-1}{s-1}$ ways of doing this. Since there are s segments and for the first segment we can choose any action and for the others we can choose any other action than the one chosen for the previous segments, there are kk^{s-1} valid ways of assigning actions to segments. Thus, $G_{n,s}^* = kk^{s-1}\binom{n-1}{s-1}$. Define $\Phi_m(n) = \sum_{i=0}^{m}\binom{n}{i}$. Hence,

$G \le k^m \sum_{s=0}^{m-1} \binom{n-1}{s} = k^m \Phi_{m-1}(n-1) \le k^m \Phi_m(n)$. Now note that for $n \ge m$, $0 \le m/n \le 1$, hence

$$\left(\frac{m}{n}\right)^m \Phi_m(n) \le \sum_{i=0}^{m} \left(\frac{m}{n}\right)^i \binom{n}{i} \le \sum_{i=0}^{n} \left(\frac{m}{n}\right)^i \binom{n}{i} = \left(1 + \frac{m}{n}\right)^n \le e^m .$$

Reordering gives $\Phi_m(n) \le \left(\frac{en}{m}\right)^m$. Hence, $\log(G) \le m \log(ekn/m)$. Plugging this into (31.1) gives (31.2).

31.3 Use the construction and analysis in Exercise 11.5 and note that when $m = 2$ the random version of the regret is nonnegative on the bandit constructed there.

Chapter 32 Ranking

32.2 The argument is half-convincing. The heart of the argument is that under the criterion that at least one item should attract the user, it may be suboptimal to present the list composed of the fittest items. The example with the query 'jaguar' is clear: Assume half of the users will mean 'jaguar' as the big cat, while the other half will mean it as the car. Presenting items that are relevant for both meanings may have a better chance to satisfy a randomly picked user than going with the top m list, which may happen to support only one of the meanings. This shows that there is indeed an issue with 'linearizing' the problem by just considering individual item fitness values.

However, the argument is confusing in other ways. First, it treats conditions (for example, independence of attractiveness) that are sufficient but not necessary to validate the probabilistic ranking principle (PRP) as if they were also necessary. In fact, in click model studied here, the mentioned independence assumption is not needed. To clarify, the strong assumption in the stochastic click model, is that the optimal list is indeed optimal. Under this assumption, the independence assumption is not needed.

Next, that the same document can have different relevance to different users fits even the cascade model, where the vector of attraction values are different each time they are sampled from the model. So this alone would not undermine the PRP.

Finally, the last sentence confuses relevance and 'usefulness'. Again, in the cascade model, the relevance (attractiveness) of a document (item) does not depend on the relevance of any other document. Yet in the reward in the cascade model is exactly one if and only if at least one document presented is relevant (attractive).

32.6 Following the proof of Theorem 32.2 the first part until Eq. (32.5) we have

$$R_n \le nm\mathbb{P}(F_n) + \sum_{j=1}^{\ell} \sum_{i=1}^{\min\{m, j-1\}} \mathbb{E}\left[\mathbb{I}\left\{F_n^c\right\} \sum_{t=1}^{n} U_{tij}\right] .$$

As before the first term is bounded using Lemma 32.4. Then using the first part of the proof of

Lemma 32.7 shows that

$$\mathbb{I}\{F_n^c\}\sum_{t=1}^{n} U_{tij} \le 1 + \sqrt{2N_{nij}\log\left(\frac{c\sqrt{n}}{\delta}\right)}\,.$$

Substituting into the previous display and applying Cauchy-Schwarz shows that

$$R_n \le nm\mathbb{P}(F_n) + m\ell + \sqrt{2m\ell\mathbb{E}\left[\sum_{j=1}^{\ell}\sum_{i=1}^{\min\{m,j-1\}} N_{nij}\right]\log\left(\frac{c\sqrt{n}}{\delta}\right)}\,.$$

Writing out the definition of N_{nij} reveals that we need to bound

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^{\ell}\sum_{i=1}^{\min\{m,j-1\}} N_{nij}\right] &\le \sum_{t=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\sum_{d=1}^{M_t}\sum_{j\in\mathcal{P}_{td}}\sum_{i\in\mathcal{P}_{td}\cap[m]} U_{tij}\middle|\mathcal{F}_{t-1}\right]\right] \\ &\le \sum_{t=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\sum_{d=1}^{M_t}\sum_{j\in\mathcal{P}_{td}}\sum_{i\in\mathcal{P}_{td}\cap[m]} (C_{ti}+C_{tj})\middle|\mathcal{F}_{t-1}\right]\right] = (\mathrm{A})\,.\end{aligned}$$

Expanding the two terms in the inner sum and bounding each separately leads to

$$\begin{aligned}\mathbb{E}\left[\sum_{d=1}^{M_t}\sum_{j\in\mathcal{P}_{td}}\sum_{i\in\mathcal{P}_{td}\cap[m]} C_{ti}\,\middle|\,\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\sum_{d=1}^{M_t}|\mathcal{P}_{td}|\sum_{i\in\mathcal{P}_{td}\cap[m]} C_{ti}\,\middle|\,\mathcal{F}_{t-1}\right] \\ &\le \sum_{d=1}^{M_t}|\mathcal{I}_{td}\cap[m]||\mathcal{P}_{td}\cap[m]| \le m^2\,,\end{aligned}$$

where the inequality follows from the fact that for $i\in\mathcal{P}_{td}$,

$$\mathbb{E}[C_{ti}\,|\,\mathcal{F}_{t-1}] \le \mathbb{P}\left(A_t^{-1}(i)\in[m]\,|\,\mathcal{F}_{t-1}\right) = \frac{|\mathcal{I}_{td}\cap[m]|}{|\mathcal{I}_{td}|} = \frac{|\mathcal{I}_{td}\cap[m]|}{|\mathcal{P}_{td}|}\,.$$

For the second term that makes up (A),

$$\begin{aligned}\mathbb{E}_{t-1}\left[\sum_{d=1}^{M_t}\sum_{j\in\mathcal{P}_{td}}\sum_{i\in\mathcal{P}_{td}\cap[m]} C_{tj}\,\middle|\,\mathcal{F}_{t-1}\right] &= \mathbb{E}_{t-1}\left[\sum_{d=1}^{M_t}|\mathcal{P}_{td}\cap[m]|\sum_{j\in\mathcal{P}_{td}} C_{tj}\,\middle|\,\mathcal{F}_{t-1}\right] \\ &\le \sum_{d=1}^{M_t}|\mathcal{P}_{td}\cap[m]||\mathcal{I}_{td}\cap[m]| \le m^2\,.\end{aligned}$$

Hence $(\mathrm{A}) \le nm^2$ and $R_n \le nm\mathbb{P}(F_n) + m\ell + \sqrt{4m^3\ell n\log\left(\frac{c\sqrt{n}}{\delta}\right)}$ and the result follows from

Lemma 32.4.

Chapter 33 Pure Exploration

33.3 Abbreviate $f(\alpha) = \inf_{d \in D} \langle \alpha, d \rangle$, which is clearly positively homogeneous: $f(c\alpha) = cf(\alpha)$ for any $c \geq 0$. Because D is nonempty, $f(\mathbf{0}) = 0$. Hence we can ignore $\alpha = \mathbf{0}$ in both optimization problems and so

$$\begin{aligned}
\left(\sup_{\alpha \in \mathcal{P}_{k-1}} f(\alpha) \right)^{-1} L &= \inf_{\alpha \in \mathcal{P}_{k-1}} \frac{L}{f(\alpha)} \\
&= \inf_{\alpha \geq 0 : \|\alpha\|_1 > 0} \frac{L \|\alpha\|_1}{f(\alpha)} \\
&= \inf_{\alpha \geq 0 : \|\alpha\|_1 > 0} \|L\alpha / f(\alpha)\|_1 \\
&= \inf \left\{ \|\alpha\|_1 : f(\alpha) \geq L \right\} ,
\end{aligned}$$

where we used the positive homogeneity of $f(\alpha)$ and the ℓ_1 norm.

33.4

(a) For each $i > 1$ define

$$\mathcal{E}_i = \{ \tilde{\nu} \in \mathcal{E} : \mu_1(\tilde{\nu}) = \mu_i(\tilde{\nu}) \text{ and } \mu_j(\tilde{\nu}) = \mu_j(\nu) \text{ for } j \notin \{1, i\} \} .$$

You can easily show that

$$\begin{aligned}
\inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^{k} \alpha_i \, \mathrm{D}(\nu_i, \tilde{\nu}_i) &= \min_{i>1} \inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}(\nu)} \left(\alpha_1 \, \mathrm{D}(\nu_1, \tilde{\nu}_1) + \alpha_i \, \mathrm{D}(\nu_i, \tilde{\nu}_i) \right) \\
&= \min_{i>1} \inf_{\tilde{\mu} \in \mathbb{R}} \left(\frac{\alpha_1 (\mu_1(\nu) - \tilde{\mu})^2}{2\sigma_1^2} + \frac{\alpha_i (\mu_i(\nu) - \tilde{\mu})^2}{2\sigma_i^2} \right) \\
&= \frac{1}{2} \min_{i>1} \frac{\alpha_1 \alpha_i \Delta_i^2}{\alpha_1 \sigma_i^2 + \alpha_i \sigma_1^2} .
\end{aligned}$$

(b) Let $\alpha_1 = \alpha$ so that $\alpha_2 = 1 - \alpha$. By the previous part

$$\begin{aligned}
(c^*(\nu))^{-1} = \max_{\alpha \in [0,1]} \inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^{k} \alpha_i \, \mathrm{D}(\nu_i, \tilde{\nu}_i) &= \max_{\alpha \in [0,1]} \frac{\alpha_1 \alpha_2 \Delta_2^2}{\alpha_1 \sigma_2^2 + \alpha_i \sigma_1^2} \\
&= \max_{\alpha \in [0,1]} \frac{\alpha(1-\alpha)\Delta_2^2}{\alpha \sigma_2^2 + (1-\alpha)\sigma_1^2} \\
&= \frac{\Delta_2^2}{2(\sigma_1^2 + \sigma_2^2)} .
\end{aligned}$$

(c) By the result in Exercise 33.3 and Part (a) of this exercise,

$$\begin{aligned} c^*(\nu) &= \inf\left\{ \|\alpha\|_1 : \alpha \in [0,\infty)^k,\ \inf_{\tilde{\nu}\in\mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^k \alpha_i \,\mathrm{D}(\nu_i,\tilde{\nu}_i) = 1 \right\} \\ &= \inf\left\{ \|\alpha\|_1 : \alpha \in [0,\infty)^k,\ \min_{i>1} \frac{\alpha_1\alpha_i\Delta_i^2}{2\alpha_1\sigma_i^2 + 2\alpha_i\sigma_1^2} = 1 \right\}. \end{aligned}$$

Let $\alpha_1 = 2a\sigma_1^2/\Delta_{\min}^2$, which by the constraint that $\alpha \geq 0$ must satisfy $a > 1$. Then

$$\begin{aligned} c^*(\nu) &= \inf_{\alpha_1 > 2\sigma_1^2/\Delta_{\min}^2} \alpha_1 + \sum_{i=2}^k \frac{2\alpha_1\sigma_i^2}{\alpha_1\Delta_i^2 - 2\sigma_1^2} \\ &\leq \inf_{a>1} \frac{2a\sigma_1^2}{\Delta_{\min}^2} + \frac{a}{a-1}\sum_{i=2}^k \frac{2\sigma_i^2/\Delta_i^2}{a-1} \\ &= \left(\sqrt{\frac{2\sigma_1^2}{\Delta_{\min}^2}} + \sqrt{\sum_{i=2}^k \frac{2\sigma_i^2}{\Delta_i^2}}\right)^2 \\ &= \frac{2\sigma_1^2}{\Delta_{\min}^2} + \sum_{i=2}^k \frac{2\sigma_i^2}{\Delta_i^2} + \frac{4\sigma_1}{\Delta_{\min}}\sqrt{\sum_{i=2}^k \frac{2\sigma_i^2}{\Delta_i^2}}. \end{aligned}$$

(d) From the previous part

$$c^*(\nu) = \inf\left\{ \|\alpha\|_1 : \alpha \in [0,\infty)^k,\ \min_{i>1} \frac{\alpha_1\alpha_i\Delta_i^2}{2\alpha_1\sigma_i^2 + 2\alpha_i\sigma_1^2} = 1 \right\},$$

Let $\alpha_1 = 2a\sigma_1^2/\Delta_{\min}^2$ with $a > 1$. Then

$$\begin{aligned} c^*(\nu) &= \inf_{\alpha_1 > 2\sigma_1^2/\Delta_{\min}^2} \left(\alpha_1 + \sum_{i=2}^k \frac{2\alpha_1\sigma_i^2}{\alpha_1\Delta_i^2 - 2\sigma_1^2}\right) \\ &\leq \inf_{a>1} \left(\frac{2a\sigma_1^2}{\Delta_{\min}^2} + \frac{a}{a-1}\sum_{i=2}^k \frac{2\sigma_i^2}{\Delta_i^2}\right) \\ &= \left(\sqrt{\frac{2\sigma_1^2}{\Delta_{\min}^2}} + \sqrt{\sum_{i=2}^k \frac{2\sigma_i^2}{\Delta_i^2}}\right)^2 \\ &= \frac{2\sigma_1^2}{\Delta_{\min}^2} + \sum_{i=2}^k \frac{2\sigma_i^2}{\Delta_i^2} + \frac{4\sigma_1}{\Delta_{\min}}\sqrt{\sum_{i=2}^k \frac{2\sigma_i^2}{\Delta_i^2}}. \end{aligned}$$

(e) Notice that the inequality in the previous part is now an equality.

33.6

(a) Let $\nu \in \mathcal{E}$ be an arbitrary Gaussian bandit with $\mu_1(\nu) > \max_{i>1} \mu_i(\nu)$ and assume that

$$\liminf_{n\to\infty} \frac{-\log\left(\mathbb{P}_{\nu\pi}(\Delta_{A_{n+1}} > 0)\right)}{\log(n)} > 1 + \varepsilon\,. \tag{33.1}$$

Notice that if Eq. (33.1) were not true then we would be done. Then let ν' be a Gaussian bandit in $\mathcal{E}_{\text{alt}}(\nu)$ with $\mu(\nu') = \mu(\nu)$ except that $\mu_i(\nu') = \mu_i(\nu) + \Delta_i(\nu)(1+\delta)$ where $i > 1$ and $\delta = \sqrt{1+\varepsilon} - 1$. By Theorem 14.2 and Lemma 15.1,

$$\begin{aligned}\mathbb{P}_{\nu\pi}(A_{n+1} \neq 1) + \mathbb{P}_{\nu'\pi}(A_{n+1} \neq i) &\geq \frac{1}{2}\exp\left(-\operatorname{D}(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi})\right) \\ &\geq \frac{1}{2}\exp\left(-\frac{(1+\delta)^2\Delta_i(\nu)^2\mathbb{E}_{\nu\pi}[T_i(n)]\log(n)}{2}\right).\end{aligned}$$

Because π is assumed to be asymptotically optimal $\lim_{n\to\infty} \mathbb{E}_{\nu\pi}[T_i(n)]/\log(n) = 2/\Delta_i(\nu)$ and hence

$$\mathbb{P}_{\nu\pi}(A_{n+1} \neq 1) + \mathbb{P}_{\nu'\pi}(A_{n+1} \neq i) = \Omega\left(\frac{1}{n^{1+\varepsilon}}\right).$$

Using Eq. (33.1) shows that

$$\limsup_{n\to\infty} n^{1+\varepsilon}\mathbb{P}_{\nu'\pi}(A_{n+1} \neq i) > 0\,,$$

which implies that

$$\liminf_{n\to\infty} \frac{-\log\left(\mathbb{P}_{\nu'\pi}(A_{n+1} \neq i)\right)}{\log(n)} \leq (1+\delta)^2 \leq 1 + \varepsilon\,.$$

(b) No. Consider the algorithm that chooses $A_n = 1 + (n \bmod k)$.

(c) The same argument as Part (a) shows there exists a $\nu \in \mathcal{E}$ with a unique optimal arm such that

$$\liminf_{n\to\infty} \frac{-\log\left(\mathbb{P}_{\nu\pi}(A_{n+1} \notin i^*(\nu))\right)}{\log(n)} = O(1)\,,$$

which means the probability of selecting a suboptimal arm decays only polynomially with n.

33.7

(a) Given a bandit $\tilde{\nu} \in \mathcal{E}$ define $\Phi : \mathcal{E} \times \mathcal{P}_{k-1} \to \mathbb{R}$ by

$$\Phi_{\tilde{\nu}}(\nu, x) = \sum_{i=1}^{k} x_i(\mu_i(\nu) - \mu_i(\tilde{\nu}))^2\,,$$

which is continuous for fixed $\tilde{\nu}$ and satisfies

$$\Phi(\nu, x) = \inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}(\nu)} \Phi_{\tilde{\nu}}(\nu, x)\,. \tag{33.2}$$

Let $\varepsilon > 0$ be sufficiently small so that $d(\nu, \xi) \leq \varepsilon$ implies that $i^*(\xi) = i^*(\nu)$, which exists since it was assumed that ν has a unique optimal arm. Define

$$\mathcal{E}_{\text{alt}} = \left\{ \tilde{\nu} \in \mathcal{E}_{\text{alt}}(\nu) : \min_i \mu_i(\tilde{\nu}) \geq \min_i \mu_i(\nu) - 2\varepsilon,\ \max_i \mu_i(\tilde{\nu}) \leq \max_i \mu_i(\nu) + 2\varepsilon \right\}$$

Then for any ξ and $\beta \in \mathcal{P}_{k-1}$ with $d(\nu, \xi) \leq \varepsilon$ it holds that

$$\Phi(\xi, \beta) = \inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}} \Phi_{\tilde{\nu}}(\xi, \beta)\,.$$

Hence on the ball $B = \{(\xi, \beta) : d(\nu, \xi) < \varepsilon\}$ the function Φ is an infimum over a compact set of continuous functions and hence Φ is continuous on B. That α^* is continuous now follows from the fact in the hint and because $\alpha^*(\nu)$ is unique.

(b) Define random variable $\Lambda \geq 1$ by

$$\Lambda = \min \left\{ \lambda \geq 1 : d(\hat{\nu}_t, \nu) \leq \sqrt{\frac{2 \log(\Lambda k t(t+1))}{\min_i T_i(t)}} \text{ for all } t \right\}.$$

Then $\mathbb{E}[\log(\Lambda)] \leq 1$. Hence $d(\hat{\nu}_t, \nu) \leq \delta$ for all $t \geq \tau$ given by

$$\tau_\nu = \min \left\{ t : \sqrt{\frac{2 \log(\Lambda k t(t+1))}{\min_i T_i(t)}} \leq \varepsilon(\delta) \right\}.$$

(c) Let $w(\varepsilon) = \inf\{x : d(\xi, \nu) \leq x \implies \|\alpha^*(\nu) - \alpha^*(\xi)\| \leq \varepsilon\}$, which by (a) satisfies $w(\varepsilon) > 0$ for all $\varepsilon > 0$. Hence $\mathbb{E}[\tau_\alpha(\varepsilon)] \leq \mathbb{E}[\tau_\nu(w(\varepsilon))] < \infty$.

(d) By definition of the algorithm $A_t = i$ implies that either $T_i(t-1) \leq \sqrt{t}$ or $A_t = \operatorname{argmax}_i \alpha_i^*(\hat{\nu}_{t-1}) - T_i(t-1)/(t-1)$. Now suppose that

$$t \geq \frac{2}{\varepsilon}(\tau_\alpha(\varepsilon/(2k)) + k(1 + \sqrt{t}))\,.$$

Then the definition of the algorithm implies that

$$T_i(t) \leq \max\left\{ T_i(\tau_\alpha(\varepsilon/(2k))), 1 + t(\alpha_i^*(\nu) + \varepsilon/(2k)), 1 + \sqrt{t} \right\} \leq t\alpha_i^*(\nu) + t\varepsilon\,.$$

Furthermore, since $\sum_{i=1}^k T_i(t) = t$ it follows that

$$\begin{aligned} T_i(t) &\geq t - \sum_{j \neq i} \max\left\{T_j(\tau_\alpha(\varepsilon/(2k))), 1 + t(\alpha_j^*(\nu) + \varepsilon/(2k)), 1 + \sqrt{t}\right\} \\ &\geq t\alpha_i^*(\nu) - \tau_\alpha(\varepsilon/(2k)) - k(1+\sqrt{t}) - t\varepsilon/2 \\ &\geq t\alpha_i^*(\nu) - t\varepsilon\,. \end{aligned}$$

And the result follows from the previous part, which ensures that

$$\mathbb{E}\left[\frac{2}{\varepsilon}(\tau_\alpha(\varepsilon/(2k)) + k(1+\sqrt{t}))\right] < \infty\,.$$

(e) Given $\varepsilon > 0$ let $\tau_\beta(\varepsilon) = \max\{t : t\Phi(\nu, \alpha^*(\nu)) \geq \beta_t(\delta) + \varepsilon t\}$ and

$$u(\varepsilon) = \sup_{(\xi,x)} \left\{\Phi(\xi, x) : d(\xi, \nu) \leq \varepsilon, \|x - \alpha^*(\nu)\|_\infty \leq \varepsilon\right\}\,.$$

Then for $t \geq \max\{\tau_\nu(\varepsilon), \tau_T(\varepsilon), \tau_\beta(u(\varepsilon))\}$ it holds that

$$t\Phi(\hat{\nu}_t, \alpha^*(\hat{\nu}_t)) \geq t(\Phi(\nu, \alpha^*(\nu)) - u(\varepsilon)) \geq \beta_t(\delta)\,,$$

which implies that

$$\tau \leq \max\{\tau_\nu(\varepsilon), \tau_T(\varepsilon), \tau_\beta(u(\varepsilon))\} \leq \tau_\nu(\varepsilon) + \tau_T(\varepsilon) + \tau_\beta(u(\varepsilon))\,.$$

Taking the expectation,

$$\mathbb{E}[\tau] \leq \mathbb{E}[\tau_\nu(\varepsilon)] + \mathbb{E}[\tau_T(\varepsilon)] + \mathbb{E}[\tau_\beta(u(\varepsilon))]\,.$$

Taking the limit as $\delta \to 0$ and using the previous parts shows that for any sufficiently small $\varepsilon > 0$,

$$\limsup_{\delta \to 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \leq \limsup_{\delta \to 0} \frac{\mathbb{E}[\tau_\beta(u(\varepsilon))]}{\log(1/\delta)} = \frac{1}{\Phi(\nu, \alpha^*(\nu)) - u(\varepsilon)}\,.$$

Continuity of Φ at $(\nu, \alpha^*(\nu))$ ensures that $\lim_{\varepsilon \to 0} u(\varepsilon) = 0$ and the result follows since $c^*(\nu) = 1/\Phi(\nu, \alpha^*(\nu))$.

33.8 For the first part we need to prove

$$H_2(\mu) \leq H_1(\mu) \leq (1 + \log(k))H_2(\mu)$$

where

$$H_1(\mu) = \sum_{i=1}^k \frac{1}{\Delta_i^2}\,, \qquad\qquad H_2(\mu) = \max_{i \in [k]} \frac{i}{\Delta_i^2}\,,$$

and where $\Delta_1 = \Delta_2 \leq \cdots \leq \Delta_k$. For the first inequality note that $1/\Delta_2^2 \geq \cdots \geq 1/\Delta_k^2$. Summing up the first i of these numbers, we see that $H_1(\mu) \geq \sum_{j=1}^{i} 1/\Delta_j^2 \geq i/\Delta_i^2$. Taking the max of both sides over i, we get $H_1(\mu) \geq H_2(\mu)$. To get the reverse inequality, note that

$$H_1(\mu) = \frac{1}{2}\frac{2}{\Delta_2^2} + \sum_{i=2}^{k} \frac{1}{i}\frac{i}{\Delta_i^2} \leq \left(\frac{1}{2} + \sum_{i=2}^{k} \frac{1}{i}\right) \max_{j \geq 2} \frac{j}{\Delta_j^2} \leq (1 + \log(k)) \max_{j \geq 1} \frac{j}{\Delta_j^2} \,.$$

For the second part note that whenever $\Delta_1 = \cdots = \Delta_k > 0$ it holds that $H_1(\mu) = H_2(\mu)$. Further, when $\Delta_i = \sqrt{ic}$ with $c > 0$ for $i \geq 2$, $i/\Delta_i^2 = 1/c$, hence $H_1(\mu) = (1/2 + \sum_{i=2}^{k} \frac{1}{i})\frac{1}{c} = L \max_i \frac{i}{\Delta_i^2}$.

33.10 We have $\mathbb{P}\left(\max_{i \in [n]} \mu(X_i) < \mu_\alpha^*\right) = \mathbb{P}\left(\mu(X_1) < \mu_\alpha^*\right)^n \leq (1-\alpha)^n \leq \delta$. Solving for n gives the required inequality.

Chapter 34 Foundations of Bayesian Learning

34.3

(a) By the 'sections' Lemma 1.26 in [Kallenberg, 2002], $d(x) = \int_\Theta p_\psi(x) q(\psi) d\nu(\psi)$ is $\mathcal{H}$-measurable. Therefore $N = d^{-1}(0) \in \mathcal{H}$ is measurable. Then

$$\begin{aligned}
0 &= \int_N \int_\Theta p_\psi(x) q(\psi) d\nu(\psi) d\mu(x) \\
&= \int_\Theta \int_N p_\psi(x) d\mu(x) q(\psi) d\nu(\psi) \\
&= \int_\Theta P_\psi(N) d\nu(\psi) \\
&= \mathbb{P}(X \in N) \\
&= \mathbb{P}_X(N) \,.
\end{aligned}$$

where the first equality follows from the definition of N. The second is Fubini's theorem, the third by the definition of the Radon-Nikodym derivative, the fourth by the definition of $\mathbb{P}$ and the last by the definition of $\mathbb{P}_X$.

(b) Note that $q(\theta \,|\, x) = p_\theta(x) q(\theta)/d(x)$, which is jointly measurable in θ and x. The fact that $Q(A \,|\, x)$ is a probability measure for all x is straightforward from the definition of expectation and because for $x \in N$,

$$Q(\Theta \,|\, x) = \frac{\int_\Theta q(\theta \,|\, x)}{\int_\Theta p_\psi(x) q(\psi) d\nu(\psi)} = 1 \,.$$

That $Q(A \,|\, \cdot)$ is $\mathcal{H}$-measurable follows from the sections lemma and the fact that $N \in \mathcal{H}$. Let

$A \in \mathcal{G}$ and $B \in \sigma(X) \subseteq \mathcal{F}$, which can be written as $B = \Theta \times C$ for some $C \in \mathcal{H}$. Then

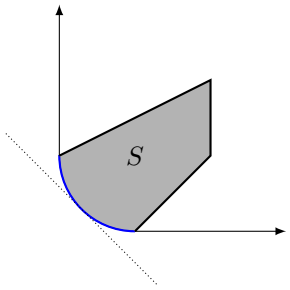
$$\begin{aligned}
\int_B Q(A \,|\, X(\omega)) d\mathbb{P}(\omega) &= \int_B \int_A q(\theta \,|\, X(\omega)) d\nu(\theta) d\mathbb{P}(\omega) \\
&= \int_\Theta \int_C \int_A q(\theta \,|\, x) d\nu(\theta) p_\theta(x) q(\theta) d\mu(x) d\nu(\theta) \\
&= \int_C d(x) \int_A q(\theta \,|\, x) d\nu(\theta) d\mu(x) \\
&= \int_C \int_A p_\theta(x) q(\theta) d\nu(\theta) d\mu(x) \\
&= \int_A p_\theta(C) q(\theta) d\nu(\theta) \\
&= \mathbb{P}(\theta \in A, X \in C) \\
&= \mathbb{P}(\theta \in A, X \in C) \\
&= \int_B \mathbb{I}_A(\theta) d\mathbb{P}\,,
\end{aligned}$$

which is the required averaging property.

34.11 Let π^* as in the problem definition. Let $\tilde{S}$ be the extension of S by adding rays in the positive direction: $\tilde{S} = \{x + u : x \in S, u \geq 0\}$. Clearly $\tilde{S}$ remains convex and $\lambda(S) \subseteq \partial\tilde{S}$ is on the boundary and is a subset of $\lambda(\tilde{S})$ (see figure) Let $x \in \lambda(S)$. By the supporting hyperplane theorem and the convexity of $\tilde{S}$ there exists a nonzero vector $a \in \mathbb{R}^N$ and $b \in \mathbb{R}$ such that $\langle a, \ell(\pi^*)\rangle = b$ and $\langle a, y\rangle \geq b$ for all $y \in \tilde{S}$. Furthermore $a \geq 0$ since $x + e_i \in \tilde{S}$ and so $\langle x + e_i, a\rangle = b + a_i \geq b$. Define $q(\nu_i) = a_i / \|a\|_1$. Then, for any policy π,

$$\sum_{\nu \in \mathcal{E}} q(\nu)\ell(\pi, \nu) = \frac{1}{\|a\|_1} \sum_{i=1}^N a_i \ell(\pi, \nu_i) = \frac{\langle a, \ell(\pi)\rangle}{\|a\|_1} \geq \frac{b}{\|a\|_1}$$

with equality for any policy π with $\ell(\pi) = \ell(\pi^*)$. Since a_i is nonnegative, $a \neq 0$, $q \in \mathcal{P}(\mathcal{E})$, finishing the proof.



34.12

1. Suppose that π is not admissible. Then there exists another policy π' with $\ell(\pi', \nu) \leq \ell(\pi, \nu)$ for all $\nu \in \mathcal{E}$. Clearly π' is also Bayesian optimal. But π was unique, which is a contradiction.

2. Suppose that $\Pi = \{\pi_1, \pi_2\}$ and $\mathcal{E} = \{\nu_1, \nu_2\}$ and $\ell(\pi, \nu_1) = 0$ for all π and $\ell(\pi_i, \nu_2) = \mathbb{I}\{i = 2\}$. Then any policy is Bayesian optimal for $Q = \delta_{\nu_1}$, but π_2 is dominated by π_1.

3. Suppose that π is not admisslbe. Then there exists another policy π' with $\ell(\pi', \nu) \leq \ell(\pi, \nu)$ for all $\nu \in \mathcal{E}$ and $\ell(\pi', \nu) < \ell(\pi, \nu)$ for at least one $\nu \in \mathcal{E}$. Then

$$\int_{\mathcal{E}} \ell(\pi, \nu) dQ(\nu) = \sum_{\nu \in \mathcal{E}} Q(\{\nu\}) \ell(\pi, \nu) > \sum_{\nu \in \mathcal{E}} Q(\{\nu\}) \ell(\pi, \nu) = \int_{\mathcal{E}} \ell(\pi', \nu) dQ(\nu) \,,$$

   which is a contradiction.

34.13 Let Π be the set of all policies and $\Pi_{\mathrm{D}} = \{e_1, \ldots, e_N\}$ the set of all deterministic policies, which is finite. Exercise 34.7 allows us to view probability measures on $(\Pi_{\mathrm{D}}, \mathfrak{B}(\Pi_{\mathrm{D}}))$, which because Π_{D} is finite are just probability vectors in $\mathcal{P}_{N-1}$. In this way Π inherits a metric and topology from $\mathcal{P}_{N-1}$. Even more straightforwardly, $\mathcal{E}$ is identified with $[0,1]^k$ and inherits a metric from that space. As metric spaces both Π and $\mathcal{E}$ are compact and the regret $R_n(\pi, \nu)$ is continuous in both arguments by Exercise 14.4. Let $(\nu_j)_{j=1}^{\infty}$ be a sequence of bandit environments that is dense in $\mathcal{E}$ and $\mathcal{E}_j = \{\nu_1, \ldots, \nu_j\}$. Using the notation of the previous exercise, let $R_{n,j}(\pi) = (R_n(\pi, \nu_1), \ldots, R_n(\pi, \nu_j))$, $S_j = R_{n,j}(\Pi) \subset \mathbb{R}^j$ and let $\lambda(S_j)$ be the Pareto frontier of S_j. Note that S_j is non-empty, closed and convex. Thus, $\lambda(S_j) \subset S_j$. Now let $\pi^{\mathrm{adm}} \in \Pi$ be an admissible policy. Then $R_{n,j}(\pi^{\mathrm{adm}}) \in \lambda(S_j)$ and by the result of the previous exercise there exists a distribution $Q_j \in \mathcal{P}(\mathcal{E})$ supported on $\mathcal{E}_j$ such that $\mathrm{BR}_n(\pi^{\mathrm{adm}}, Q_j) \leq \min_\pi \mathrm{BR}_n(\pi, Q_j)$. Let $\mathcal{Q}$ be the space of probability measures on $(\mathcal{E}, \mathfrak{B}(\mathcal{E}))$, which is compact with the weak* topology by Theorem 2.14. Hence $(Q_j)_{j=1}^{\infty}$ contains a convergent subsequence $(Q_i)_i$ converging to Q. Notice that $\nu \mapsto R_n(\pi, \nu)$ is a continuous function from $\mathcal{E}$ to $[0, n]$. Therefore by the definition of the weak* topology for any policy π,

$$\lim_{i \to \infty} \mathrm{BR}_n(\pi, Q_i) = \lim_{i \to \infty} \int_{\mathcal{E}} R_n(\pi, \nu) dQ(\nu) = \mathrm{BR}_n(\pi, Q) \,.$$

Hence $\mathrm{BR}_n(\pi^{\mathrm{adm}}, Q) \leq \min_\pi \mathrm{BR}(\pi, Q)$.

34.14 Clearly

$$\max_{Q \in \mathcal{Q}} \mathrm{BR}_n^*(Q) \leq R_n^*(\mathcal{E}) \,.$$

In the remainder we show the other direction. Since $[0,1]^k$ is compact with the usual topology, Theorem 2.14 shows that $\mathcal{Q}$ is compact with the weak* topology. Let Π_{D} be the space of all deterministic policies with the discrete topology and let

$$\mathcal{P} = \left\{ \sum_{\pi \in A} p_i \delta_\pi : p \in \mathcal{P}(A) \text{ and } A \subset \Pi_{\mathrm{D}} \text{ is finite} \right\} ,$$

which is a convex subspace of the topological vector space of all signed measures on $(\Pi_{\mathrm{D}}, 2^{\Pi_{\mathrm{D}}})$ with

the weak* topology. Let $g : \mathcal{P} \times \mathcal{Q} \to [0, n]$ be defined by

$$g(\mu, Q) = -\int_{\Pi_{\mathrm{D}}} \int_{\mathcal{E}} R_n(\pi, \nu) dQ(\nu) d\mu(\pi) \,.$$

The regret is bounded in $[0, n]$ and the discrete topology on Π_{D} means that all functions from Π_{D} to $\mathbb{R}$ are continuous, including $\pi \mapsto R_n(\pi, \nu)$. Then by the definition of the weak* topology, $g(\cdot, Q)$ is continuous in its first argument for all Q (see Note 16 in Chapter 2). The integral over Π_{D} with respect to $\mu \in \mathcal{P}$ is a finite sum and $\nu \mapsto R_n(\pi, \nu)$ is continuous for all π by the result in Exercise 14.4. Therefore g is continuous and linear in both arguments. By Sion's theorem (Theorem 28.12),

$$-\max_{Q \in \mathcal{Q}} \mathrm{BR}_n^*(Q) = \min_{Q \in \mathcal{Q}} \sup_{\mu \in \mathcal{P}} g(\mu, Q) = \sup_{\mu \in \mathcal{P}} \min_{Q \in \mathcal{Q}} g(\mu, Q) = -\inf_{\mu \in \mathcal{P}} R_n^*(\pi, \mathcal{E}) \,.$$

Therefore

$$\max_{Q \in \mathcal{Q}} \mathrm{BR}_n^*(Q) = \inf_{\mu \in \mathcal{P}} R_n^*(\pi, \mathcal{E}) \geq \inf_{\pi \in \Pi} R_n^*(\pi, \mathcal{E}) = R_n^*(\pi) \,.$$

Hence $R_n^*(\mathcal{E}) = \max_{Q \in \mathcal{Q}} \mathrm{BR}_n^*(Q)$.

Chapter 35 Bayesian Bandits

35.1 Let π be the policy of MOSS from Chapter 9, which for any 1-subgaussian bandit ν with rewards in $[0, 1]$ satisfies

$$R_n(\pi, \nu) \leq C \min\left\{\sqrt{kn}, \frac{k \log(n)}{\Delta_{\min}(\nu)}\right\} ,$$

where $\Delta_{\min}(\nu)$ is the smallest positive suboptimality gap. Let $\mathcal{E}_n$ be the set of bandits in $\mathcal{E}$ for which there exists an arm i with $\Delta_i \in (0, n^{-1/4})$. Then, for $C' = Ck$,

$$\begin{aligned}
\mathrm{BR}_n^*(Q) &\leq \mathrm{BR}_n(\pi, Q) \\
&= \int_{\mathcal{E}} R_n(\pi, \nu) dQ(\nu) \\
&= \int_{\mathcal{E}_n} R_n(\pi, \nu) dQ(\nu) + \int_{\mathcal{E}_n^c} R_n(\pi, \nu) dQ(\nu) \\
&\leq C' \sqrt{n} Q(\mathcal{E}_n) + C' \int_{\mathcal{E}_n^c} n^{1/4} \log(n) dQ(\nu) \\
&= C' \sqrt{n} Q(\mathcal{E}_n) + o(\sqrt{n}) \,.
\end{aligned}$$

The first part follows since $\cap_n \mathcal{E}_n = \emptyset$ and thus $\lim_{n \to \infty} Q(\mathcal{E}_n) = 0$ for any measure Q. For the second part we describe roughly what needs to be done. The idea is to make use of the minimax lower bound technique in Exercise 15.2, which shows that for a uniform prior concentrated on a finite set

of k bandits the regret is $\Omega(\sqrt{kn})$. The only problems are that *(a)* the rewards were assumed to be Gaussian and *(b)* the prior depends on n. The first issue is corrected by replacing the Gaussian distributions with Bernoulli distributions with means close to $1/2$. For the second issue you should compute this prior for $n \in \{1, 2, 4, 8, \ldots\}$ and denote them $Q_1, Q_2, \ldots$. Then let $Q = \sum_{j=1}^{\infty} p_j Q_j$ where $p_j \propto (j \log^2(j))^{-1}$. The result follows easily.

35.2 Recall that $E_t = U_t$ and for $t < n$,

$$E_t = \max\{U_t, \mathbb{E}[E_{t+1} \,|\, \mathcal{F}_t]\}\,.$$

Integrability of $(U_t)_{t=1}^n$ ensures that $(E_t)_{t=1}^n$ are integrable. By definition $E_t \geq \mathbb{E}[E_{t+1} \,|\, \mathcal{F}_t]$. Hence $(E_t)_{t=1}^n$ is a supermartingale adapted to $\mathbb{F}$. Hence for any stopping time $\kappa \in \mathfrak{R}_1^n$ the optional stopping theorem says that

$$\mathbb{E}[U_\kappa] \leq \mathbb{E}[E_\kappa] \leq E_1\,.$$

On the other hand, for τ satisfying the requirements of the lemma the process $M_t = E_{t \wedge \tau}$ is a martingale and hence $\mathbb{E}[U_\tau] = \mathbb{E}[M_\tau] = M_1 = E_1$.

35.3 Define $v^n(x) = \sup_{\tau \in \mathfrak{R}_1^n} \mathbb{E}_x[U_\tau]$. By assumption, $\mathbb{E}_x[|u(S_t)|] < \infty$ for all $x \in \mathcal{S}$ and $t \in [n]$. Therefore by Theorem 1.7 of Peskir and Shiryaev [2006],

$$v^n(x) = \max\{u(x), \int_{\mathcal{S}} v^{n-1}(y) P_x(dy)\}\,. \tag{35.1}$$

Recall that $v(x) = \sup_\tau \mathbb{E}_x[U_\tau]$. Clearly $v^n(x) \leq v(x)$ for all $x \in \mathcal{S}$. Let τ be an arbitrary stopping time. Then

$$v^n(x) \geq \mathbb{E}_x[U_{\tau \wedge n}] = \mathbb{E}_x[U_\tau] + \mathbb{E}_x[(U_n - U_\tau)\mathbb{I}\{\tau \geq n\}]\,.$$

Since $\sup_n |U_n|$ is $\mathbb{P}_x$-integrable for all x by assumption, the dominated convergence theorem shows that

$$\lim_{n\to\infty} \mathbb{E}_x[(U_n - U_\tau)\mathbb{I}\{\tau \geq n\}] = \mathbb{E}_x\left[\lim_{n\to\infty} (U_n - U_\tau)\mathbb{I}\{\tau \geq n\}\right] = 0\,,$$

where the second equality follows because $U_\infty = \lim_{n\to\infty} U_n$ exists $\mathbb{P}_x$-almost surely by assumption. Therefore $\lim_{n\to\infty} v^n(x) = v(x)$. Since convergence is monotone it follows that v is measurable. Taking limits in Eq. (35.1) shows that

$$\begin{aligned}
v(x) &= \lim_{n\to\infty} \max\{u(x), \int_{\mathcal{S}} v^{n-1}(y) P_x(dy)\} \\
&= \max\{u(x), \lim_{n\to\infty} \int_{\mathcal{S}} v^{n-1}(y) P_x(dy)\} \\
&= \max\{u(x), \int_{\mathcal{S}} v(y) P_x(dy)\}\,,
\end{aligned}$$

where the last equality follows from the monotone convergence theorem. Next, let $V_n = v(S_n)$. By

definition $V_n \geq U_n$. Hence

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}_x[|V_n - U_n|] &= \lim_{n\to\infty} \mathbb{E}_x[V_n - U_\infty] \\
&\leq \lim_{n\to\infty} \mathbb{E}_x \left[\mathbb{E}_x \left[\sup_{t\geq n} U_t \,\middle|\, S_n \right] - U_\infty \right] \\
&= \lim_{n\to\infty} \mathbb{E}_x \left[\sup_{t\geq n} U_t - U_\infty \right] \\
&= \mathbb{E}_x \left[\lim_{n\to\infty} \sup_{t\geq n} U_t - U_\infty \right] \\
&= 0\,,
\end{aligned}$$

where the exchange of limit and expectation is again justified by the dominated convergence theorem and the assumption that $\sup_n |U_n|$ is $\mathbb{P}_x$-integrable. Therefore $V_\infty = \lim_{n\to\infty} V_n = U_\infty$ $\mathbb{P}_x$-a.s.. Then

$$\mathbb{E}_x[V_{n+1} \,|\, S_n] = \int_{\mathcal{S}} v(y) P_{S_n}(dy) \leq V_n \text{ a.s.}\,.$$

Therefore $(V_n)_{n=1}^\infty$ is a supermartingale, which means that for any stopping time κ,

$$\mathbb{E}_x[U_\kappa] = \mathbb{E}_x \left[\lim_{n\to\infty} U_{\kappa\wedge n} \right] = \lim_{n\to\infty} \mathbb{E}_x[U_{\kappa\wedge n}] \leq \lim_{n\to\infty} \mathbb{E}_x[V_{\kappa\wedge n}] \leq v(x)\,,$$

where the exchange of limits and expectation is justified by the dominated convergence theorem and the fact that $U_{\kappa\wedge n} \leq \sup_n U_n$, which is $\mathbb{P}_x$-integrable by assumption. Consider a stopping time τ satisfying the conditions of Theorem 35.3. Then $(V_{n\wedge\tau})_{n=1}^\infty$ is a martingale and using the same argument as before we have

$$\mathbb{E}_x[U_\tau] = \mathbb{E}_x[V_\tau] = \mathbb{E}_x \left[\lim_{n\to\infty} V_{\tau\wedge n} \right] = \lim_{n\to\infty} \mathbb{E}_x \left[V_{\tau\wedge n} \right] = v(x)\,,$$

where the first inequality follows from the assumption on τ that on the event $\tau < \infty$, $U_\tau = V_\tau$ and the fact that $V_\infty = U_\infty$ $\mathbb{P}_x$-a.s..

35.4 Fix $x \in \mathcal{S}$ and let

$$g = \sup_{\tau\geq 2} \frac{\mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t r(S_t) \right]}{\mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t \right]}\,.$$

We will show that *(a)* $v_\gamma(x) > 0$ for all $\gamma < g$ and *(b)* $v_\gamma(x) = 0$ for all $\gamma \geq g$.

For *(a)*, assume $\gamma < g$. By the definition of g there exists a stopping time $\tau \geq 2$ such that

$$\mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t r(S_t) \right] > \gamma\, \mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t \right]\,,$$

which implies that

$$v_\gamma(x) \geq \mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t (r(S_t) - \gamma) \right] > 0 \,.$$

Moving now to *(b)*, first note that $v_\gamma(x) \geq 0$ for any $\gamma \in \mathbb{R}$ because when $\tau = 1$, $\mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t (r(S_t) - g) \right] = 0$. Hence, it suffices to show that $v_\gamma(x) \leq 0$ for all $\gamma \geq g$. Pick $\gamma \geq g$. By the definition of g, for any stopping time $\tau \geq 2$,

$$\mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t (r(S_t) - \gamma) \right] \leq 0 \,,$$

which implies that

$$\sup_{\tau \geq 2} \mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t (r(S_t) - \gamma) \right] \leq 0 \,.$$

If τ is a $\mathbb{F}$-stopping time then $\mathbb{P}_x(\tau = 1)$ is either zero or one (the stopping rule underlying τ either stops given $S_1 = x$, or does not stop – the stopping rule cannot inject any further randomness). From this it follows that

$$\begin{aligned} v_g(x) &= \sup_{\tau \geq 1} \mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t (r(S_t) - \gamma) \right] \\ &= \max \left\{ 0, \sup_{\tau \geq 2} \mathbb{E}_x \left[\sum_{t=1}^{\tau-1} \alpha^t (r(S_t) - \gamma) \right] \right\} \\ &\leq 0 \,, \end{aligned}$$

finishing the proof.

35.5 Define $S^n = (S_t)_{t=1}^n$. The sequence $(S^n)_{n=1}^\infty$ is a Markov chain on the space of finite sequences $\mathcal{S}^*$ with probability kernel characterized by

$$Q_{x_{1:n}}(B_1 \times \cdots \times B_{n+1}) = P_{x_n}(B_{n+1}) \prod_{t=1}^{n} \mathbb{I}\{x_t \in B_t\} \,,$$

where $B_t \subseteq \mathcal{S}$ is measurable. Note that for measurable $f : \mathcal{S}^* \to \mathbb{R}$ and $x_{1:n} \in \mathcal{S}^n$,

$$\int_{\mathcal{S}^*} f(y) Q_{x_{1:n}}(dy) = \int_{\mathcal{S}} f(x_{1:n} x_{n+1}) P_{x_n}(dx_{n+1}) \,.$$

Now define measurable function $u : \mathcal{S}^* \to \mathbb{R}$ by

$$u(x_{1:n}) = \sum_{t=1}^{n-1} \alpha^t (r(x_t) - \gamma) \,.$$

Let $U_t = u(S^t)$ and define

$$\bar{v}_\gamma(x_{1:n}) = \sup_{\tau \geq 1} \mathbb{E}_{x_{1:n}}[U_t]\,, \tag{35.2}$$

where $\mathbb{E}_{x_{1:n}}$ is the expectation on the Markov chain $(S^n)_{n=1}^\infty$ when the initial 'state' is $S^1 = x_{1:n} \in \mathcal{S}^*$. The definitions ensure that for any $x, y \in \mathcal{S}$ and $\gamma \in \mathbb{R}$,

$$v_\gamma(x) = \bar{v}_\gamma(x) \qquad \text{and} \qquad \bar{v}_\gamma(xy) = r(x) - \gamma + \alpha \bar{v}_\gamma(y)\,.$$

In order to apply Theorem 35.3 we need to check the existence and integrability conditions of $(U_n)_{n=1}^\infty$. By Assumption 35.6 and the dominated convergence theorem, $U = \lim_{n\to\infty} U_n$ exists $\mathbb{P}_{x_{1:n}}$-a.s. and $\sup_{n\geq 1} U_n$ is $\mathbb{P}_{x_{1:n}}$-integrable for all $x_{1:n} \in \mathcal{S}^*$. Then by Theorem 35.3 it follows that

$$\bar{v}_\gamma(x_{1:n}) = \max\{u(x_{1:n}), \int_{\mathcal{S}} \bar{v}_\gamma(x_{1:n}x_{n+1}) P_{x_n}(dx_{n+1})\}\,.$$

The proof of Part (a) is completed by noting that

$$\begin{aligned} v_\gamma(x) = \bar{v}_\gamma(x) &= \max\{0, \int_{\mathcal{S}} \bar{v}_\gamma(xy) P_x(dy)\} \\ &= \max\{0, r(x) - \gamma + \alpha \int_{\mathcal{S}} v_\gamma(y) P_x(dy)\}\,. \end{aligned}$$

For Part (b), when $\gamma < g(x)$ we have $v_\gamma(x) > 0$ by definition and hence using the previous part it follows that

$$v_\gamma(x) = r(x) - \gamma + \alpha \int_{\mathcal{S}} v_\gamma(y) P_x(dy)\,.$$

Note that $\sup_{x\in\mathcal{S}} |v_{\gamma+\delta}(x) - v_\gamma(x)| \leq |\delta|/(1-\alpha)$ and hence by continuity for $\gamma = g(x)$ we have

$$r(x) - \gamma + \alpha \int_{\mathcal{S}} v_\gamma(y) P_x(dy) = 0 = v_\gamma(x)\,.$$

For Part (c), applying Theorem 35.3 again shows that when $\bar{v}_\gamma(x) = 0$, then $\tau = \min\{t \geq 2 : \bar{v}_\gamma(S^t) = u(S^t)\}$ attains the stopping time in Eq. (35.2). Therefore

$$\mathbb{E}_x\left[\sum_{t=1}^{\tau-1} \alpha^t (r(S_t) - \gamma)\right] = \mathbb{E}_x[U_\tau] = \bar{v}_x(\gamma) = v_x(\gamma) = 0\,.$$

Notice finally that

$$\bar{v}_\gamma(x_{1:n}) - u(x_{1:n}) = \alpha^n v_\gamma(x_{1:n})\,,$$

which means that $\tau = \min\{t \geq 2 : \alpha^t v_\gamma(S_t) = 0\} = \min\{t \geq 2 : g(S_t) \leq \gamma\}$.

Chapter 36 Thompson Sampling

36.3 We need to show that

$$\mathbb{P}\left(A^* = \cdot \,|\, \mathcal{F}_{t-1}\right) = \mathbb{P}\left(A_t = \cdot \,|\, \mathcal{F}_{t-1}\right)$$

holds almost surely. For specificity, let $r : \mathcal{E} \to [k]$ be the (tie-breaking) rule that chooses the arm with the highest mean given a bandit environment so that $A_t = r(\nu_t)$ and $A^* = r(\nu)$. Recall that $\nu_t \sim Q_{t-1}(\cdot) = Q(\,\cdot \,|\, A_1, X_1, \ldots, A_{t-1}, X_{t-1})$. We have

$$\begin{aligned}
\mathbb{P}\left(A^* = i \,|\, \mathcal{F}_{t-1}\right) &= \mathbb{P}\left(r(\nu) = i \,|\, \mathcal{F}_{t-1}\right) && \text{(definition of } A^*\text{)} \\
&= Q_{t-1}(\{x \in \mathcal{E} \,:\, r(x) = i\}) && \text{(definition of } Q_{t-1}\text{)} \\
&= \mathbb{P}\left(r(\nu_t) = i \,|\, \mathcal{F}_{t-1}\right) && \text{(definition of } \nu_t\text{)} \\
&= \mathbb{P}\left(A_t = i \,|\, \mathcal{F}_{t-1}\right) . && \text{(definition of } A_t\text{)}
\end{aligned}$$

36.5 We have

$$\begin{aligned}
\sum_{t \in \mathcal{T}} \mathbb{I}\left\{A_t = i\right\} &\leq \sum_{t=1}^{n} \sum_{s=1}^{n} \mathbb{I}\left\{T_i(t) = s, T_i(t-1) = s-1, G_i(T_i(t-1)) > 1/n\right\} \\
&= \sum_{s=1}^{n} \mathbb{I}\left\{G_i(s-1) > 1/n\right\} \sum_{t=1}^{n} \mathbb{I}\left\{T_i(t) = s, T_i(t-1) = s-1\right\} \\
&= \sum_{s=1}^{n} \mathbb{I}\left\{G_i(s-1) > 1/n\right\} ,
\end{aligned}$$

where the first equality uses that when $A_t = i$, $T_i(t) = s$ and $T_i(t-1) = s-1$ for some $s \in [n]$ and that $t \in \mathcal{T}$ implies $G_i(T_i(t-1)) > 1/n$. The next equality is by algebra, and the last follows because for any $s \in [n]$, there is at most one time point $t \in [n]$ such that $T_i(t) = s$ and $T_i(t-1) = s-1$. For the next inequality, note that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t \notin \mathcal{T}} \mathbb{I}\left\{E_i^c(t)\right\}\right] &= \sum_t \mathbb{E}\big[\mathbb{E}[\mathbb{I}\left\{E_i^c(t), G_i(T_i(t-1)) \leq 1/n\right\} | \mathcal{F}_{t-1}]\big] \\
&= \sum_t \mathbb{E}\big[\mathbb{I}\left\{G_i(T_i(t-1)) \leq 1/n\right\} G_i(T_i(t-1))\big] \\
&\leq \sum_t \mathbb{E}\big[\mathbb{I}\left\{G_i(T_i(t-1)) \leq 1/n\right\} 1/n\big] \\
&= \mathbb{E}\left[\sum_{t \notin \mathcal{T}} 1/n\right] ,
\end{aligned}$$

where the second equality used that $\mathbb{I}\{G_i(T_i(t-1)) \leq 1/n\}$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[\mathbb{I}\{E_i^c(t)\}|\mathcal{F}_{t-1}] = 1 - \mathbb{P}(\theta_i(t) \leq \mu_1 - \varepsilon|\mathcal{F}_{t-1}) = G_i(T_i(t-1))$.

36.6

(a) Let $f(y) = \sqrt{s/(2\pi)}\exp(-sy^2/2)$ be the probability density function of a centered Gaussian with variance $1/s$ and $F(y) = \int_{-\infty}^{y} f(x)dx$ be its cumulative distribution function. Then

$$\begin{aligned} G_{1s} &= \int_{\mathbb{R}} f(y+\varepsilon)F(y)/(1-F(y))dy \\ &\leq \int_0^\infty f(y+\varepsilon)/(1-F(y)) + 2\int_{-\infty}^0 f(y+\varepsilon)F(y)dy\,. \end{aligned} \tag{36.1}$$

For the first term in Eq. (36.1), following the hint, we use the following bound on $1 - F(y)$ for $y \geq 0$:

$$1 - F(y) \geq \frac{\exp(-sy^2/2)}{y\sqrt{s} + \sqrt{sy^2+4}}\,.$$

Hence

$$\begin{aligned} \int_0^\infty \frac{f(y+\varepsilon)}{1-F(y)}dy &\leq \int_0^\infty f(y+\varepsilon)\exp(sy^2/2)(y\sqrt{s} + \sqrt{sy^2+4})dy \\ &\leq 2\exp(-s\varepsilon^2/2)\int_0^\infty \exp(-sy\varepsilon)(y\sqrt{s}+1)\sqrt{\frac{s}{2\pi}}dy \\ &= 2\frac{1+\varepsilon\sqrt{s}}{\varepsilon^2 s\sqrt{2\pi}}\exp(-s\varepsilon^2/2)\,. \end{aligned}$$

For the second term in Eq. (36.1),

$$2\int_{-\infty}^0 f(y+\varepsilon)F(y)dy \leq 2\int_{\mathbb{R}} f(y+\varepsilon)F(y)dy \leq 2\exp(-s\varepsilon^2)\,.$$

Summing from $s = 1$ to ∞ shows that

$$2\sum_{s=1}^\infty \left(\exp(-s\varepsilon^2) + \frac{1+\varepsilon\sqrt{s}}{\varepsilon^2 s\sqrt{2\pi}}\exp(-s\varepsilon^2/2)\right) \leq \frac{c}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)\,,$$

where the last line follows from a Mathematica slog.

(b) Let $\hat{\mu}_{is}$ be the empirical mean of arm i after s observations. Then $G_{is} \leq 1/n$ if

$$\hat{\mu}_{is} + \sqrt{\frac{2\log(n)}{s}} \leq \mu_1 - \varepsilon\,.$$

Hence for $s \geq u = \frac{2\log(n)}{(\Delta_i - \varepsilon)}$ we have

$$\begin{aligned}
\mathbb{P}\left(G_{is} > 1/n\right) &\leq \mathbb{P}\left(\hat{\mu}_{is} + \sqrt{\frac{2\log(n)}{s}} > \mu_1 - \varepsilon\right) \\
&= \mathbb{P}\left(\hat{\mu}_{is} - \mu_i > \Delta_i - \varepsilon - \sqrt{\frac{2\log(n)}{s}}\right) \\
&\leq \exp\left(-\frac{s\left(\Delta_i - \varepsilon - \sqrt{\frac{2\log(n)}{s}}\right)^2}{2}\right).
\end{aligned}$$

Summing,

$$\begin{aligned}
\sum_{s=1}^{n} \mathbb{P}\left(G_{is} > 1/n\right) &\leq u + \sum_{s=\lceil u \rceil}^{n} \exp\left(-\frac{s\left(\Delta_i - \varepsilon - \sqrt{\frac{2\log(n)}{s}}\right)^2}{2}\right) \\
&\leq 1 + \frac{2}{(\Delta_i - \varepsilon)^2}(\log(n) + \sqrt{\pi \log(n)} + 1)\,,
\end{aligned}$$

where the last inequality follows by bounding the sum by an integral as in the proof of Lemma 8.2.

36.7 Abbreviate $\mathbb{P}_{\mathcal{G}}(\cdot) = \mathbb{P}(\,\cdot \,|\, \mathcal{G})$. For Lemma 36.5, recall that X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\mathcal{F}_t)_{t=0}^n$ a filtration of $\mathcal{F}$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathbb{P}_t(\cdot) = \mathbb{P}\left(\,\cdot \,|\, \mathcal{F}_t\right)$. The problem is to prove

$$\mathbb{E}\left[\sum_{t=1}^{n} I_{\mathbb{P}_{t-1}}(X; \mathcal{F}_t)\right] = I_{\mathbb{P}}(X; \mathcal{F}_n)\,.$$

Recall that $H_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[\log 1/\mathbb{P}_X(X)]$, where we use $\mathbb{E}_{\mathbb{P}}$ to denote the expectation operator with respect to measure $\mathbb{P}$. For $\mathcal{G} \subset \mathcal{F}$ sub-σ-algebra, define $\mathbb{P}_{\mathcal{G}}$ via $\mathbb{P}_{\mathcal{G}}(A) = \mathbb{P}(A|\mathcal{G})$, $A \in \mathcal{F}$. Then, $H_{\mathbb{P}}(X|\mathcal{G}) = \mathbb{E}_{\mathbb{P}}[H_{\mathbb{P}_{\mathcal{G}}}(X)]$. With the help of this notation, we have

$$I_{\mathbb{P}_{t-1}}(X; \mathcal{F}_t) = H_{\mathbb{P}_{t-1}}(X) - H_{\mathbb{P}_{t-1}}(X|\mathcal{F}_t) = H_{\mathbb{P}_{t-1}}(X) - \mathbb{E}_{\mathbb{P}_{t-1}}[H_{\mathbb{P}_t}(X)]$$

and thus

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n} I_{\mathbb{P}_{t-1}}(X; \mathcal{F}_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{n} H_{\mathbb{P}_{t-1}}(X) - \mathbb{E}_{\mathbb{P}_{t-1}}[H_{\mathbb{P}_t}(X)]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} H_{\mathbb{P}_{t-1}}(X) - H_{\mathbb{P}_t}(X)\right] \\
&= \mathbb{E}\left[H_{\mathbb{P}_0}(X) - H_{\mathbb{P}_n}(X)\right].
\end{aligned}$$

The result follows by noting that $\mathbb{P}_0 = \mathbb{P}$.

For Lemma 36.6, recall that X and Y are random variables on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where X is discrete and $I(X;Y)$ exists. The problem is to show that $I(X;Y) = \mathbb{E}\left[\mathrm{D}(\mathbb{P}(Y \in \cdot \,|\, X), \mathbb{P}(Y \in \cdot))\right]$.

Let $o_{X|Y}(x|Y) = \mathbb{P}(X = x|Y)$. This is almost-surely well-defined. Further, by the note that comes with Fig. 2.4, there exist some Borel function f_x such that $p_{X|Y}(x|Y) = f_x(Y)$ holds almost surely. This, allows us to define $p_{X|Y}(x|y)$ as $f_x(y)$ for $y \in \mathbb{R}$. Then,

$$I(X;Y) = \mathbb{E}\left[\log \frac{1}{\mathbb{P}_X(X)} - \log \frac{1}{p_{X|Y}(X|Y)}\right].$$

Let μ be a measure that dominates the measures $P_Y(\cdot) = \mathbb{P}(Y \in \cdot)$ and $P_{Y|X}(\,\cdot\,|\,x) = \mathbb{P}(Y \in \cdot|X = x)$ where $x \in \mathrm{range}(X)$. This measure exists because X is discrete (in fact, we can take $\mu = P_Y$). Let $p_Y = dP_Y/d\mu$ and $p_{Y|X}(\,\cdot\,|\,x) = dP_{Y|X}(\,\cdot\,|\,x)/d\mu$. Then, for any $U \subset \mathbb{R}$ Borel,

$$\begin{aligned}\int_U P_{X|Y}(x|y)p_Y(y)d\mu(y) &= \mathbb{E}\left[P_{X|Y}(x|Y)\mathbb{I}\{Y \in U\}\right] \\ &= \mathbb{E}\left[\mathbb{P}(X = x|Y)\,\mathbb{I}\{Y \in U\}\right] = \mathbb{P}(X = x, Y \in U) \\ &= \mathbb{P}(Y \in U|X = x)\,\mathbb{P}(X = x) = \int_U p_{Y|X}(y|x)\mathbb{P}_X(x)d\mu(y).\end{aligned}$$

Thus, $p_{X|Y}(x|y)P_Y(y) = p_{Y|X}(y|x)\mathbb{P}_X(x)$ holds for μ-almost all y and it follows that for μ-almost all y,

$$\frac{p_{X|Y}(x|y)}{\mathbb{P}_X(x)} = \frac{p_{Y|X}(y|x)}{p_Y(y)}.$$

By the definition of μ, almost surely,

$$\frac{p_{X|Y}(x|Y)}{\mathbb{P}_X(x)} = \frac{p_{Y|X}(Y|x)}{p_Y(Y)}.$$

Plugging X in the place of x, taking the logarithm of both sides and then expectation, we see that the left-hand side is $I(X;Y)$, while the right-hand side is the expected relative entropy between $\mathbb{P}(Y \in \cdot\,|\,X)$ and $\mathbb{P}(Y \in \cdot)$, as required.

36.9 Let $J_t = I_{t-1}(A^*; \mathcal{F}_t)$. Then

$$\mathrm{BR}_n = \mathbb{E}\left[\sum_{t=1}^n \sqrt{I_{t-1}(A^*;\mathcal{F}_t)\Gamma_t}\right] = \mathbb{E}\left[\langle J^{1/2}, \Gamma^{1/2}\rangle\right],$$

By Cauchy-Schwarz, $\langle J^{1/2}, \Gamma^{1/2}\rangle \le \|J^{1/2}\|_2 \|\Gamma^{1/2}\|_2$. By the assumption that $\frac{1}{n}\sum_{t=1}^n \Gamma_t \le \bar{\Gamma}$ it follows that $\|\Gamma^{1/2}\|_2 \le \sqrt{n\bar{\Gamma}}$ and hence

$$\mathrm{BR}_n \le \sqrt{n\bar{\Gamma}}\,\mathbb{E}\left[\|J^{1/2}\|_2\right] \le \sqrt{n\bar{\Gamma}\,\mathbb{E}\left[\|J^{1/2}\|_2^2\right]} = \sqrt{n\bar{\Gamma}\,H(A^*)},$$

where the second equality follows from Jensen's inequality, while last follows from Lemma 36.5.

36.10 Let $\Delta_t = X_{tA^*} - X_{tA_t}$ and $I_t = I_{\mathbb{P}_{t-1}}(A^*; A_t, X_t)$. Then information directed sampling chooses $A_t \sim \pi_t$ where π_t minimizes $\Gamma_t = \mathbb{E}_{t-1}[\Delta_t]^2 / I_t$. Then,

$$\begin{aligned}\mathrm{BR}_n &= \mathbb{E}\left[\sum_{t=1}^n (X_{tA^*} - X_{tA_t})\right] = \mathbb{E}\left[\sum_{t=1}^n \Delta_t\right] = \mathbb{E}\left[\sum_{t=1}^n \sqrt{\Gamma_t I_t}\right] \\ &\leq \mathbb{E}\left[\sqrt{n \sum_{t=1}^n \Gamma_t I_t}\right] \leq \sqrt{n \mathbb{E}\left[\sum_{t=1}^n \Gamma_t I_t\right]}\,.\end{aligned}$$

Since π_t is chosen to minimize Γ_t it follows by Lemma 36.9 that $\Gamma_t \leq k/2$ almost surely. Hence

$$\mathrm{BR}_n \leq \sqrt{\frac{nk}{2} \mathbb{E}\left[\sum_{t=1}^n I_t\right]} \leq \sqrt{\frac{nk \log(k)}{2}}\,.$$

36.13 The plan is to use Sion's theorem (Theorem 28.12). First notice that $\mathcal{Q}$ is a convex subset of the linear topological space of all signed measures on $(\mathcal{E}, 2^{\mathcal{E}})$ with a topology induced by the total variation distance. Let Π_{D} be the space of deterministic policies, which is finite. Therefore $\mathcal{P}(\Pi_{\mathrm{D}})$ is convex and compact with respect to the total variation distance. For distribution $\mu \in \mathcal{P}(\Pi_{\mathrm{D}})$ with associated policy π_μ given by Exercise 34.7 and $Q \in \mathcal{P}(\mathcal{E})$ define

$$g(\mu, Q) = \mathrm{BR}_n(\pi_\mu, Q) = \int_{\mathcal{E}} \int_{\Pi_{\mathrm{D}}} R_n(\pi, \nu) d\mu(\pi) dQ(\nu)\,.$$

Since $R_n(\pi, \nu)$ is bounded in $[-n, n]$ it follows from Fubini's theorem and Exercise 14.4 that g is linear and continuous in both arguments. Therefore

$$\begin{aligned}R_n^*(\mathcal{E}) &= \min_{\pi \in \Pi} \max_{x \in \mathcal{E}} R_n(\pi, x) \\ &= \min_{\pi \in \Pi} \max_{Q \in \mathcal{Q}} \mathrm{BR}_n(\pi, Q) \\ &= \max_{Q \in \mathcal{Q}} \min_{\pi \in \Pi} \mathrm{BR}_n(\pi, Q) \\ &= \max_{Q \in \mathcal{Q}} \mathrm{BR}_n^*(Q)\,.\end{aligned}$$

where the exchange of minimum and maximum is by Sion's theorem.

36.15 Let π be a minimax optimal policy for $\{0,1\}^{n \times k}$. Given an arbitrary adversarial bandit $x \in [0,1]^{n \times k}$. Choose $\tilde{\pi}$ to be the policy obtained by observing $X_t = x_{tA_t}$ and then sampling $\tilde{X}_t \sim \mathcal{B}(X_t)$ and passing X_t to π. Then

$$R_n(\tilde{\pi}, x) \leq \sum_{\tilde{x} \in \{0,1\}^{n \times k}} \prod_{t=1}^n \prod_{i=1}^k x_{ti}^{\tilde{x}_{ti}} (1 - x_{ti})^{1 - \tilde{x}_{ti}} R_n(\pi, \tilde{x}) \leq R_n^*(\{0,1\}^{n \times k})\,.$$

Therefore $R_n^*([0,1]^{n\times k}) \le R_n^*(\{0,1\}^{n\times k})$. The other direction is obvious.

Chapter 37 Partial Monitoring

37.8 Abbreviate $P = Q_t$ and let $P^* = \lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} P^t$. By Parts (c) and (d) of Exercise 38.7, P^* is right stochastic and $P^*P = P^*$. Let $\nu^\top = e_1^\top P^*$. Then $\nu^\top P = e_1^\top P^* P = e_1^\top P^* = \nu_1^\top$, which proves the existence of a stationary distribution. For uniqueness, notice that for each $a \in \mathcal{A}$ and $b \in [k]$ there exists an $i \in \mathbb{N}$ such that $P^i_{ba} > 0$. Similarly, for any $c \in [k]$ and $b \notin \mathcal{A}$ we have $P^*_{cb} = 0$. Then since $P^* = P^* P^i$ it follows that for $a \in \mathcal{A}$ we have $P^*_{ba} > 0$ for all $b \in [k]$. Suppose ν is a stationary distribution and $b \notin \mathcal{A}$. Then $\nu_b = \sum_a P^*_{ab}\nu_a = 0$ and hence the stationary distribution is unique and vanishes outside of $\mathcal{A}$, which establishes the final claim in the lemma. Now suppose that μ is another stationary distribution and $\mu \neq \nu$. Let $a = \operatorname{argmax}_a(\nu_a - \mu_a)$, which means that $\nu_a > \mu_a$ and so $a \in \mathcal{A}$. Since both are stationary we have $0 < \varepsilon = \nu_a - \mu_a = \sum_b P^*_{ba}(\nu_b - \mu_b)$ and hence $\nu_b = \mu_b + \varepsilon$ whenever $P^*_{ba} > 0$. But this means that $1 = \|\nu\|_1 > \|\mu\|_1 = 1$, which is a contradiction.

37.9 First we show that P_t is indeed a probability vector. By assumption $\tilde{P}_t$ is the stationary distribution, which is a probability distribution. Let $\bar{P}_t = \textsc{Redistribute}(\tilde{P}_t)$ so that

$$P_t = (1-\gamma)\bar{P}_t + \frac{\gamma}{k}\mathbf{1}\,,$$

which means we need to show that $\bar{P}_t$ is a probability distribution. Since $\bar{P}_t$ is obtained by the iterative procedure given in the Redistribute function it is sufficient to show that the vector q tracked by this algorithm is indeed a distribution. The claim is that each loop of the Redistribute function does not break this property. The first observation is that the algorithm always moves mass from actions in $\mathcal{A}$ to actions in $\mathcal{D}$. All that must be shown is that $\bar{P}_{ta} \ge 0$ for all $a \in \mathcal{A}$. To see this note first that if $a \in \mathcal{A}$ is one of the choices of the algorithm in Line 11, then $\rho c_a q_a \le p_a/(2k)$ and so

$$\bar{P}_{ta} \ge \tilde{P}_{ta}/2 \qquad \text{for all } a \in \mathcal{A} \ge 0\,. \tag{37.1}$$

Part (a): Since $\gamma \le 1/2$ this follows from Eq. (37.1).

Part (b): First we show that $\sum_{a\in[k]}(\bar{P}_{ta} - \tilde{P}_{ta})\ell_a = 0$. It suffices to show that the redistribution in each inner loop of the algorithm does not change this value, which is true because

$$\begin{aligned}
(c_a q_a + c_b q_b)\ell_e &= (c_a q_a + c_b q_b)(\alpha\ell_a + (1-\alpha)\ell_b) \\
&= \frac{q_a q_b}{\alpha q_b + (1-\alpha)q_a}(\alpha\ell_a + (1-\alpha)\ell_b) \\
&= \rho c_a q_a \ell_a + \rho c_b q_b \ell_b\,.
\end{aligned}$$

Then using the definition of P_t we have

$$\begin{aligned}\left|\sum_{a\in[k]}(P_{ta}-\tilde{P}_{ta})\langle\ell_a,u\rangle\right| &= \left|\sum_{a\in[k]}(P_{ta}-\bar{P}_{ta})\langle\ell_a,u\rangle\right| \\ &= \gamma\left|\sum_{a\in[k]}\left(\frac{1}{k}-\bar{P}_{ta}\right)\langle\ell_a,u\rangle\right| \\ &\le \gamma\,,\end{aligned}$$

where we used the assumption that $\ell_a \in [0,1]^d$ for all actions and $u \in \mathcal{P}_{d-1}$ so that $\langle \ell_a, u\rangle \in [0,1]$.

(c): There are three cases: Either $b = j$ or $b = a$ or b is degenerate. If $b = j$, then the result is immediate from Part (a). If $b = a$, then, Part (a) combined with (37.11) implies that $P_{tb} = P_{ta} \ge \tilde{P}_{ta}/4 \ge \tilde{P}_{tj}Q_{tja}/4 \ge \tilde{P}_{tj}Q_{tja}/(4k)$. Finally, if b is degenerate, then by the definition of the rebalancing algorithm we have

$$\bar{P}_{tb} \ge \frac{\min(\tilde{P}_{tj},\tilde{P}_{ta})}{2k} \ge \frac{\min(\tilde{P}_{tj},\tilde{P}_{tj}Q_{tja})}{2k} = \frac{\tilde{P}_{tj}Q_{tja}}{2k}$$

and the result follows from Eq. (37.1).

(d): This is trivial from the definition of P_t.

(e): Let $a, b \in \mathcal{A}$ be the Pareto optimal actions that share their probability with e in REDISTRIBUTE. Thus, $e \in \mathcal{N}_{ab}$. Since e is a duplicate of $j \in \mathcal{A}$, and $\mathcal{A}$ is duplicate-free, it must be that j is either equal to a or b. With no loss of generality assume that $j = a$. Then, $\ell_e = \ell_a$ and it follows that $\alpha = 1$ and so $c_a = 1$ and $c_b = 0$. This means that $\bar{P}_{te} = \tilde{P}_{ta}/(2k)$ and using Eq. (37.1) again yields the result.

37.10

(a) This follows immediately from Part (e) of Lemma 37.15.

(b) There are two cases: Either b is a duplicate of a, or it is not a duplicate action of either j or a. If b is a duplicate of a, Part (e) of Lemma 37.15 gives that $P_{tb} \ge \tilde{P}_{ta}/(4k)$. Hence,

$$\frac{Q_{tja}\tilde{P}_{tj}}{P_{tb}} \le 4k\frac{Q_{tja}\tilde{P}_{tj}}{\tilde{P}_{ta}} \le 4k\frac{\sum_j Q_{tja}\tilde{P}_{tj}}{\tilde{P}_{ta}} = 4k\frac{\tilde{P}_{ta}}{\tilde{P}_{ta}} = 4k\,,$$

where the second to last equality follows from the definition of $\tilde{P}_t$. In the remaining case, Part (c) of Lemma 37.15 gives the result immediately.

(c) That S and S_a are disjoint is clear. Next let a and a' be distinct elements of $\mathcal{N}_j \cap \mathcal{A}$ and suppose that $b \in \mathcal{N}_{aj} \cap \mathcal{N}_{a'j}$. By Lemma 37.7(b), either $b = j$ or $C_b = C_a \cap C_j = C_{a'} \cap C_j$. This is impossible, however, since j, a and a' are all distinct Pareto optimal actions.

37.12

(a) We only sketch the argument. Let $\mathcal{V} = \{a, b\} \times [F]$ be the vertices of a graph and let $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ be the set of edges where $((a, f), (b, f')) \in \mathcal{E}$ if there exists an $i \in [d]$ such that $\Phi_{ai} = f$ and $\Phi_{bi} = f'$, which makes $(\mathcal{V}, \mathcal{E})$ a bipartite graph. We assume without loss of generality this graph is fully connected, since if not, then the following argument can be applied to each connected component. Let $f, f' \in [F]$ and use the fact there exists a path between (a, f) and (a, f') to show that $v(a, f) - v(a, f') \leq 2F$. By translating $v(a, f)$ in one direction and $v(b, f)$ in the other we may choose v so that $\max_f |v(a, f)| \leq F$. Conclue that $\max_f |v(b, f)| \leq F + 1$ using the fact that $v(b, f) + v(a, f') = \ell_{ai} - \ell_{bi}$ for some $f' \in [F]$ and $i \in [d]$.

(b) Let $\ell_1 = (1, 0, 0, 0 \ldots, 0, 0)$ and $\ell_2 = (0, 1, 0, 0, \ldots, 0, 0)$. For all other actions $a > 2$ let $\ell_a = \mathbf{1}$. Hence $\ell_1 - \ell_2 = (1, -1, 0, 0, \ldots, 0, 0)$. Next we define the feedback matrix by

$$\Phi = \begin{pmatrix} 1 & 2 & 2 & 1 & 2 & 1 & 2 & \\ 1 & 1 & 2 & 2 & 1 & 2 & 1 & \ddots \\ 1 & 1 & 1 & 2 & 2 & 1 & 2 & \ddots \\ 1 & 1 & 1 & 1 & 2 & 2 & 1 & \ddots \\ 1 & 1 & 1 & 1 & 1 & 2 & 2 & \ddots \\ 1 & 1 & 1 & 1 & 1 & 1 & 2 & \ddots \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

For this choice we have $v(1, 2) = v(1, 1) - 2$ and $v(2, 2) = v(2, 1) + 2$ and $v(3, 2) = v(3, 1) - 2$. For $j > 3$ we have $|v(j, 2) - v(j, 1)| = |v(j - 2, 2) - v(j - 2, 1)| + |v(j - 1, 2) - v(j - 1, 1)|$. Hence $|v(j, 2) - v(j, 1)|$ grows exponentially by the usual techniques for evaluating Fibonacci sequences.

Chapter 38 Markov Decision Processes

38.2 The solution to Part (b) is immediate from Part (a), so we only show the solution to Part (a). Abbreviate $\mathbb{P}_\mu^\pi$, to $\mathbb{P}$ and $\mathbb{P}_\mu^{\pi'}$ to $\mathbb{P}'$. Let $\pi' = (\pi'_1, \pi'_2, \ldots)$ be the Markov policy to be constructed: For each $t \geq 1$, $s \in \mathcal{S}$, $\pi'_t(\cdot|s)$ is a distribution over $\mathcal{A}$.

Fix $(s, a) \in \mathcal{S} \times \mathcal{A}$ and consider first $t = 1$. We want $\mathbb{P}'(S_1 = s, A_1 = a) = \mathbb{P}(S_1 = s, A_1 = a)$. By the definition of $\mathbb{P}'$, $\mathbb{P}$ and by the definition of conditional probabilities, we have $\mathbb{P}'(S_1 = s, A_1 = a) = \mathbb{P}'(A_1 = a|S_1 = s)\mathbb{P}'(S_1 = s) = \pi'_1(a|s)\mu(s) = \mathbb{P}(S_1 = a, A_1 = a) = \mathbb{P}(A_1 = a|S_1 = s)\mu(s)$. Thus, defining $\pi'_1(a|s) = \mathbb{P}(A_1 = a|S_1 = s)$ we see that the desired equality holds for $t = 1$. Now, for $t > 1$ first notice that from $\mathbb{P}'(A_{t-1} = a, S_{t-1} = s) = \mathbb{P}'(A_{t-1} = a, S_{t-1} = s)$, $(s, a) \in \mathcal{S} \times \mathcal{A}$ it follows by summing these equations over a that $\mathbb{P}'(S_{t-1} = s) = \mathbb{P}'(S_{t-1} = s)$. Hence, the same calculation as for $t = 1$ applies, showing that $\pi'_t(a|s) = \mathbb{P}(A_t = a|S_t = s)$ will work.

38.4 We show that $D(M) < \infty$ implies that M is strongly connected and that $D(M) = \infty$ implies that M is not strongly connected. Assume first that $D(M) < \infty$. Take any $s, s' \in \mathcal{S}$. Assume first that $s \neq s'$. By definition, there is a policy whose expected travel time from state s to s' is finite.

Take this policy. It follows that this policy reaches state s' from state s with positive probability, because otherwise the expected travel time would be infinite. Formally, if T is the random travel time of a policy whose expected travel time between s and s' is finite, $\{T = \infty\}$ is the event that the policy does not reach state s'. Now, for any $n \in \mathbb{N}$, $T > n\mathbb{I}\{T = \infty\}$. Taking expectations and reordering gives $\mathbb{E}[T]/n > \mathbb{P}(T = \infty)$. Letting $n \to \infty$, we see that $\mathbb{P}(T = \infty) = 0$ (thus we see that the policy reaches state s' in fact with probability one). It remains to consider the case when $s = s'$. If the MDP has a single state, it is strongly connected by definition. Otherwise, there exist a state $s'' \in \mathcal{S}$ that is distinct from $s = s'$. Since $D(M)$ is finite, there is a policy that reaches s' from s with positive probability and another one that reaches s again from s' with positive probability. Compose these two policies the obvious way to find the policy that travels from s to s with positive probability.

Assume now that $D(M) = \infty$, while M is strongly connected (proof by contradiction). Since M is strongly connected, for any s, s', there is a policy that has a positive probability of reaching s' from s. But this means that the uniformly random policy (the policy which chooses uniformly at random between the actions at any state) has also a positive probability of reaching any state from any other state. We claim that the expected travel time of this policy is finite between any pairs of states. Indeed, this follows by noticing that the states under this policy form a time-homogenous Markov chain whose transition probability matrix is irreducible and the hitting times in some a Markov chain, which coincide with the expected travel times in the MDP for the said policy, are finite. link

38.5 We follow the advice of the hint. For the second part, note that the minimum in the definition of $d^*(\mu_0, U)$ is attained when n_k is maximized for small indices until $|U|$ is exhausted. In particular, if $(n_k)_{0 \le k \le m}$ denotes the optimal solution ($n_k = 0$ for $k > m$) then $n_0 = \mathrm{A}^0, \dots, n_{m-1} = \mathrm{A}^k$, $0 \le n_m = |U| - \sum_{k=0}^{n-1} \mathrm{A}^k (= |U| - \frac{\mathrm{A}^m - 1}{\mathrm{A} - 1}) < \mathrm{A}^m$. Hence, $|U| < \mathrm{A}^m + \frac{\mathrm{A}^m - 1}{\mathrm{A} - 1} \le 2\mathrm{A}^m$, implying that $m \ge \log_{\mathrm{A}}(|U|/2)$. Thus,

$$
\begin{aligned}
d^*(\mu_0, U) &= \sum_{k=0}^{m-1} k\,\mathrm{A}^k + m\,n_m \\
&= m|U| + \sum_{k=0}^{m-1} (k - m)\mathrm{A}^k \\
&\overset{\text{(a)}}{=} m|U| + \frac{m}{\mathrm{A} - 1} - \frac{\mathrm{A}^{m+1} - \mathrm{A}}{(\mathrm{A} - 1)^2} \\
&\ge |U| \left(m - 1 - \frac{1}{\mathrm{A} - 1} \right) \\
&\ge |U| (\log_{\mathrm{A}}(|U|) - 3)\,,
\end{aligned}
$$

where step (a) follows since $|U| < \frac{\mathrm{A}^m - 1}{\mathrm{A} - 1}$. Choosing $U = \mathcal{S}$, we see that the expected minimum time to reach a random state in $\mathcal{S}$ is lower bounded by $\log_{\mathrm{A}}(\mathrm{S}) - 3$. The expected minimum time to reach an arbitrary state in $\mathcal{S}$ must also be above this quantity, proving the desired result.

38.7

(a) $A_n\mathbf{1} = \frac{1}{n}\sum_{t=0}^{n-1} P^t\mathbf{1} = \mathbf{1}$, which means that A_n is right stochastic.

(b) This follows immediately from the definitions.

(c) Let (B_n) and (C_n) be convergent subsequences of (A_n) with $\lim_{n\to\infty} B_n = B$ and $\lim_{n\to\infty} C_n = C$. It suffices to show that $B = C$. From Part (b), $B_m + \frac{1}{n_m}(P^{n_m} - I) = B_m P = PB_m$. Taking the limit as m tends to infinity and using the fact that P^n is $[0,1]$-valued we see that $B = BP = PB$. Similarly, $C = CP = PC$ and it follows that $B = BP^i = P^i B$ and $C = CP^i = PC^i$ hold for any $i \geq 0$. Hence $B = BC_m = C_m B$ and $C = CB_m = B_m C$ for any $m \geq 1$. Taking limit as m tends to infinity shows that $B = BC = CB$ and $C = CB = BC$, which together imply that $B = C$.

(d) We have already seen in the proof of Part (c) that $P^* = P^*P = PP^*$. From this, it follows that $P^* = P^*P^i$ for any $i \geq 0$, which implies that $P^* = P^*A_n$ holds for any $n \geq 1$. Taking limit shows that $P^* = P^*P^*$.

(e) Let $B = P - P^*$. By algebra $I - B^i = (I - B)(I + B + \cdots + B^{i-1})$. Summing over $i = 1, \ldots, n$ and dividing by n and using the fact that $B^i = P^i - P^*$ for all $i \geq 1$,

$$I - \frac{1}{n}\sum_{i=1}^{n} P^i + P^* = (I - B)H_n\,, \tag{38.1}$$

where $H_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P - P^*)^k$. The limit of the left-hand side of (38.1) exists and is equal to the identity matrix I. Hence the limit of the right-hand side also exists and in particular the limit of H_n must exist. Denoting this by H_∞ we find that $I = (I - B)H_\infty$ and thus $I - B$ is invertible and its inverse H is equal to H_∞.

(f) Let $U_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P^k - P^*)$. Then

$$B^k = \begin{cases} I, & \text{if } k = 0\,; \\ P^k - P^*, & \text{otherwise}\,. \end{cases}$$

Using this we calculate $H_n - U_n = \frac{1}{n}\sum_{i=1}^{n}(P - P^*)^0 - \frac{1}{n}\sum_{i=1}^{n}(P^0 - P^*) = I - I + P^* = P^*$. Hence $H - \lim_{n\to\infty} U_n = P^*$. From the definition of U we have $U = \lim_{n\to\infty} U_n$.

(g) This follows immediately because

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1} P^k(r - \rho) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P^k - P^*)r = Ur\,.$$

(h) One way to prove this is to note that by the previous part $v = Ur = (H - P^*)r$, hence $r = H^{-1}(v+\rho)$. Now, $H^{-1}(v+\rho) = (I - P + P^*)(v+\rho) = v - Pv + P^*v + \rho - P\rho + P^*\rho = v - Pv + \rho$, where we used that $P^*v = P^*Ur$ and $P^*U = P^*(H - P^*) = P^*H - P^* = (P^* - P^*H^{-1})H = 0$ and that $P\rho = PP^*r = P^*r = P^*\rho = P^*P^*r$.

Alternatively, the following direct argument also works. In this argument we only use that v is well defined. Let $v_n = \sum_{t=0}^{n-1} P^t(r - \rho)$, $\bar{v}_n = \frac{1}{n}\sum_{i=1}^{n} v_i$. Note that $\lim_{n\to\infty}\bar{v}_n = v$. Then,

$v_{k+1} = Pv_k + (r - \rho)$. Taking the average of these over $k = 1, \dots, n$ we get

$$\frac{1}{n}\left((n+1)\bar{v}_{n+1} - v_1\right) = P\bar{v}_n + (r - \rho)\,.$$

Taking the limit of both sides proves that $v = Pv + r - \rho$, which, after reordering gives $v + \rho = r + Pv$.

38.8

(a) First note that $|\max_x f(x) - \max_y g(y)| \le \max_x |f(x) - g(x)|$. Then for $v, w \in \mathbb{R}^{\mathrm{S}}$,

$$\begin{aligned}\|T_\gamma v - T_\gamma w\|_\infty &\le \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \gamma\,|\langle P_a(s), v - w\rangle| \\ &\le \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \gamma \|P_a(s)\|_1 \|v - w\|_\infty \\ &= \gamma\|v - w\|_\infty\,.\end{aligned}$$

Hence T is a contraction with respect to the supremum norm as required.

(b) This follows immediately from the Banach fixed point theorem, which also guarantees the uniqueness of a value function v satisfying $v = T_\gamma v$.

(c) Recall that the greedy policy is $\pi(s) = \operatorname{argmax}_a r_a(s) + \gamma\langle P_a(s), v\rangle$. Then

$$v(s) = \max_{a\in\mathcal{A}} r_a(s) + \gamma\langle P_a(s), v\rangle = r_\pi(s) + \gamma\langle P_\pi(s), v\rangle\,.$$

(d) We have $v = r + \gamma P_\pi v$. Solving for v completes the result.

(e) If π is a memoryless policy, it is trivial to see that $v_\gamma^\pi = r_\pi + \gamma P_\pi v_\gamma^\pi$. Let π^* be the greedy policy with respect to v, the unique solution of $v = T_\gamma v$. By the previous part of this exercise, it follows that $v_\gamma^{\pi^*} = v$. By Exercise 38.2, it suffices to show that for any Markov policy π, $v_\gamma^\pi \le v$. If π_t is the memoryless policy used in time step t when following π, $v_\gamma^\pi = \sum_{t=1}^\infty \gamma^{t-1} P_\pi^{(t-1)} r_{\pi_t}$, where $P^{(0)} = I$ and for $t \ge 1$, $P^{(t)} = P_{\pi_1} \dots P_{\pi_t}$. For $n \ge 1$, let $v_{\gamma,n}^\pi = \sum_{t=1}^n \gamma^{t-1} P_\pi^{(t-1)} r_{\pi_t}$. It is easy to see that $v_{\gamma,1}^\pi = r_{\pi_1} \le T\mathbf{0}$. Assume that for some $n \ge 1$,

$$v_{\gamma,n}^\pi \le T^n \mathbf{0}\,. \tag{38.2}$$

Notice that $f \le g$ implies $Tf \le Tg$. Hence, $Tv_{\gamma,n}^\pi \le T^{n+1}\mathbf{0}$. Further, $v_{\gamma,n+1}^\pi = r_{\pi_{n+1}} + \gamma P_{\pi_{n+1}} v_{\gamma,n}^\pi \le Tv_{\gamma,n}^\pi$. This shows that (38.2) holds for all $n \ge 1$. Letting $n \to \infty$, the right-hand side converges to v, while the left-hand side converges to v_γ^π. Hence, $v_\gamma^\pi \le v$.

38.9

(a) Let $0 \le \gamma < 1$. Algebra gives

$$P_\gamma^* P = (1-\gamma)\sum_{t=0}^\infty \gamma^t P^t P = \frac{P_\gamma^* - (1-\gamma)I}{\gamma}\,.$$

Hence $\gamma P_\gamma^* P = P_\gamma^* - (1-\gamma)I$. It is easy to check that P_γ^* is right stochastic. By the compactness of the space of right stochastic matrices, $(P_\gamma^*)_\gamma$ has at least one cluster point A as $\gamma \to 1-$. It follows that $AP = A$, which implies that $AP^* = A$. Now, $(P_\gamma^*)^{-1}P^* = \frac{(I-\gamma P)P^*}{1-\gamma} = P^*$, which implies that $P^* = AP^* = A$. Since this holds for any cluster point we conclude that $\lim_{\gamma\to 1-} P_\gamma^* = P^*$.

(b) Since $I - P + P^*$ is invertible and $PP^* = P^* = P^*P^*$, the required claim is equivalent to

$$\lim_{\gamma\to 1-} (I - P + P^*)\left(\frac{P_\gamma^* - P^*}{1-\gamma}\right) = I - P^* .$$

Rewriting $P_\gamma^* = (1-\gamma)\sum_{t=0}^\infty \gamma^t P^t$ shows that

$$\begin{aligned}(I - P + P^*)\left(\frac{P_\gamma^* - P^*}{1-\gamma}\right) &= (I - P + P^*)\left(\sum_{t=0}^\infty \gamma^t P^t - \frac{P^*}{1-\gamma}\right) \\ &= \sum_{t=0}^\infty (I-P)\gamma^t P^t = \frac{1}{\gamma}(I - P_\gamma^*) .\end{aligned}$$

The result is completed by taking the limit as γ tends to one from below and using Part (a).

38.10

(a) Let (γ_n) be an arbitrary increasing sequence with $\gamma_n < 1$ for all n and $\lim_{n\to\infty}\gamma_n = 1$. Let v_n be the fixed point of T_{γ_n} and π_n be the greedy policy with respect to v_n. Since greedy policies are always deterministic and there are only finitely many deterministic policies it follows there exists a subsequence $n_1 < n_2 < \cdots$ and policy π such that $\pi_{n_k} = \pi$ for all k.

(b) For arbitrary π and $0 \le \gamma < 1$, let v_γ^π be the value function of policy π in the γ-discounted MDP, v^π the value function of π and ρ^π its gain. Let U_π be the deviation matrix underlying P_π. Define f_γ^π by

$$v_\gamma^\pi = \frac{\rho^\pi}{1-\gamma} + v_\pi + f_\gamma^\pi . \tag{38.3}$$

By Part (b) of Exercise 38.9, and because $\rho^\pi = P_\pi^* r_\pi$ and $v_\pi = U_\pi r_\pi$, it holds that $\|f_\gamma^\pi\|_\infty \to 0$ as $\gamma \to 1$.

Fix now π to be the policy whose existence is guaranteed by the previous part. By Part (e) of Exercise 38.8, π is γ_n-discount optimal for all $n \ge 1$. Suppose that ρ^π is not a constant. In any case, ρ^π is piecewise constant on the recurrent classes of the Markov chain with transition probabilities P^π. Let $\rho_*^\pi = \max_{s\in\mathcal{S}} \rho^\pi(s)$. Let $R \subset \mathcal{S}$ be the recurrent class in this Markov chain where ρ^π is the largest and take a policy π' that is identical to π over R, while π' is set up such that it gets to R with probability one. Such a π' exist because the MDP is strongly connected. Fix any $s \in \mathcal{S} \setminus R$. We claim that there exists some $\gamma^* \in (0,1)$ such that for all $\gamma \ge \gamma^*$,

$$v_\gamma^{\pi'}(s) > v_\gamma^\pi(s) . \tag{38.4}$$

If this was true and n is large enough so that $\gamma_n \geq \gamma^*$ then, since π is γ_n-discount optimal, $v^{\pi}_{\gamma_n}(s) \geq v^{\pi'}_{\gamma_n}(s) > v^{\pi'}_{\gamma_n}(s)$, which is a contradiction.

Hence, it remains to show (38.4). By the construction of π', $\rho^{\pi'}(s) = \rho^{\pi}_* > \rho^{\pi}(s)$. From (38.3),

$$v^{\pi'}_{\gamma}(s) = \frac{\rho^{\pi'}(s)}{1-\gamma} + v_{\pi'}(s) + f^{\pi'}_{\gamma}(s) > \frac{\rho^{\pi}(s)}{1-\gamma} + v_{\pi}(s) + f^{\pi}_{\gamma}(s) = v^{\pi}_{\gamma}(s)\,,$$

where the inequality follows by taking $\gamma \geq \gamma^*$ for some $\gamma^* \in (0,1)$. The existence of any appropriate γ^* follows because $f^{\pi'}_{\gamma}(s), f^{\pi}_{\gamma}(s) \to 0$, while $1/(1-\gamma) \to \infty$ as $\gamma \to 1$.

(c) Since v is the value function of π and ρ is its gain, by (38.4) we have $\rho\mathbf{1} + v = r_\pi + P_\pi v$. Let π' be an arbitrary stationary policy and v_n be as before. Let $f_n = f_{\gamma_n}$ where f_γ is defined by (38.3). Note that $\|f_n\|_\infty \to 0$ as $n \to \infty$. Then,

$$\begin{aligned}
0 &\geq r_{\pi'} + (\gamma_n P_{\pi'} - I)v_n \\
&= r_{\pi'} + (\gamma_n P_{\pi'} - I)\left(\frac{\rho\mathbf{1}}{1-\gamma_n} + v + f_n\right) \\
&= r_{\pi'} - \rho\mathbf{1} + \gamma_n P_{\pi'} v - v + (\gamma_n P_{\pi'} - I) f_n\,.
\end{aligned}$$

Note that when $\pi' = \pi$, the first inequality becomes an equality. Taking the limit as n tends to infinity and rearranging shows that

$$\rho\mathbf{1} + v \geq r_{\pi'} + P_{\pi'} v$$

and

$$\rho\mathbf{1} + v = r_\pi + P_\pi v\,. \tag{38.5}$$

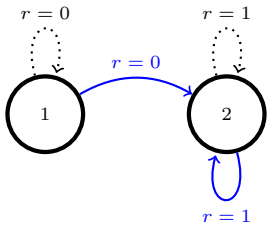
Since π' was arbitrary, $\rho + v(s) \geq \max_a r_a(s) + \langle P_a(s), v\rangle$ holds for all $s \in \mathcal{S}$. This combined with (38.5) shows that the pair (ρ, v) satisfies the Bellman optimality equation as required.

38.11 Clearly the optimal policy is to take action STAY in any state and this policy has gain $\rho^* = 0$. Pick any solution (ρ, v) to the Bellman optimality equations. Therefore $\rho = \rho^* = 0$ by Theorem 38.2. The Bellman optimality equation for state 1 is $v(1) = \max(v(1), -1 + v(2))$, which is equivalent to $v(1) \geq -1 + v(2)$. Similarly, the Bellman optimality equation for state 2 is equivalent to $v(2) \geq -1 + v(1)$. Thus the set of solutions is a subset of

$$\{(\rho, v) \in \mathbb{R} \times \mathbb{R}^2 \,:\, \rho = 0,\, v(1) - 1 \leq v(2) \leq v(1) + 1\}\,.$$

The same argument shows that any element of this set is a solution to the optimality equations.

38.12 Consider the deterministic MDP below with two states and two actions, $\mathcal{A} = \{\text{SOLID}, \text{DASHED}\}$.



Clearly the optimal policy is to choose $\pi(1) = \text{SOLID}$ and $\pi(2)$ arbitrarily which leads to a gain of 1. On the other hand, choosing $\rho = 1$ and $v = (2, 1)$ satisfies the linear program in Eq. (38.6) and the greedy policy with respect to this value function chooses $\pi(1) = \text{DASHED}$ and $\pi(2)$ arbitrary.

38.13 Let $T : \mathbb{R}^{\mathrm{S}} \to \mathbb{R}^{\mathrm{S}}$ be defined by $(Tv)(s) = \max_a r_a(s) - \rho^* + \langle P_a(s), v\rangle$ so the Bellman optimality equation Eq. (38.5) can be written in the compact form $v = Tv$. Let $v \in \mathbb{R}^{\mathrm{S}}$ be a solution to Eq. (38.5). The proof follows from the definition of the diameter and by showing that for any states $s_1, s_2 \in \mathcal{S}$ and memoryless policy π it holds that

$$v(s_2) \le v(s_1) + (\rho^* - \min_{s,a} r_a(s))\mathbb{E}^{\pi}[\tau_{s_2} \,|\, S_1 = s_1]\,.$$

The remainder of the proof is devoted to proving this result for fixed $s_1, s_2 \in \mathcal{S}$ and memoryless policy π. Abbreviate $\tau = \tau_{s_2}$ and let $\mathbb{E}[\cdot]$ denote the expectation with respect to the measure induced by the interaction of π and the MDP conditioned on $S_1 = s_1$. Since the result is trivial when $\mathbb{E}[\tau] = \infty$, for the remainder we assume that $\mathbb{E}[\tau] < \infty$. Define operator $\bar{T} : \mathbb{R}^{\mathrm{S}} \to \mathbb{R}^{\mathrm{S}}$ by

$$(\bar{T}u)(s) = \begin{cases} \min_{s,a} r_a(s) - \rho^* + \langle P_\pi(s), u\rangle, & \text{if } s \neq s_2\,; \\ v(s_2), & \text{otherwise}\,. \end{cases}$$

Since $r_\pi(s) - \rho^* \ge \min_{s,a} r_a(s) - \rho^*$ and $Tv = v$ it follows that $(\bar{T}v)(s) \le (Tv)(s) = v(s)$. Notice that for $u \le w$ it holds that $\bar{T}u \le \bar{T}w$. Then by induction we have $\bar{T}^n v \le v$ for all $n \in \mathbb{N}^+$. By unrolling the recurrence we have

$$v(s_1) \ge (\bar{T}^n v)(s_1) = \mathbb{E}\left[-(\rho^* - \min_{s,a} r_a(s))(n \wedge \tau) + v(S_{\tau \wedge n})\right]\,.$$

Taking the limit as n tends to infinity shows that $v(s_1) \ge v(s_2) - (\rho^* - \min_{s,a} r_a(s))\mathbb{E}[\tau]$, which completes the result.

38.14

(a) It is clear that Algorithm 26 returns TRUE if and only if (ρ, v) is feasible for Eq. (38.6). Note that feasibility can be written in the compact form $\rho\mathbf{1}v \ge Tv$. It remains to show that when (ρ, v) is not feasible then $u = (1, e_s - P_{a_s^*}(s))$ is such that for any (ρ', v') feasible, $\langle(\rho', v'), u\rangle > \langle(\rho, v), u\rangle$. For this, we have $\langle(\rho', v'), u\rangle = \rho' + v'(s) - \langle P_{a_s^*}(s), v'\rangle \ge r_{a_s^*}(s)$, where the inequality used that (ρ', v') is feasible. Further, $\langle(\rho, v), u\rangle = \rho + v(s) - \langle P_{a_s^*}(s), v\rangle < r_{a_s^*}(s)$, by the construction of u. Putting these together gives the result.

(b) Relax the constraint that $v(\tilde{s}) = 0$ to $-\varepsilon \le v(\tilde{s}) \le \varepsilon$. Then add the ε of slack to the first constraint of Eq. (38.7) and add the additional constraints used in Eq. (38.9). Now the ellipsoid method can be applied as for Eq. (38.9).

38.15 Let $\phi_k(x) = \text{TRUE}$ if $\langle a_k, x\rangle \ge b_k$ and $\phi_k(x) = a_k$ otherwise. Then the new separation oracle returns true if $\phi(x)$ and $\phi_k(x)$ are true for all k. Otherwise return the separating hyperplane provided by some ϕ or ϕ_k that did not return true.

38.16

(a) Let $\pi = (\pi_1, \pi_2, \ldots)$ be an arbitrary Markov policy where π_t is the policy followed in time step t. Using the notation and techniques from the proof of Theorem 38.2,

$$\begin{aligned} P_\pi^{(t-1)} r_{\pi_t} &= P_\pi^{(t-1)}(r_{\pi_t} + P_{\pi_t} v - P_{\pi_t} v) \le P_\pi^{(t-1)}(r_{\tilde{\pi}} + P_{\tilde{\pi}} v - P_{\pi_t} v) \\ &\le P_\pi^{(t-1)}((\rho+\varepsilon)\mathbf{1} + v - P_{\pi_t} v) = (\rho+\varepsilon)\mathbf{1} + P_\pi^{(t-1)} v - P_\pi^{(t)} v\,. \end{aligned}$$

Taking the average and then the limit shows that $\bar{\rho}^\pi(s) \le \rho + \varepsilon$ for all $s \in \mathcal{S}$. By the claim in Exercise 38.2, $\rho^* \le \rho + \varepsilon$.

(b) We have $r_{\tilde{\pi}}(s) + \langle P_{\tilde{\pi}(s)}(s), v\rangle \ge \max_a r_a(s) + \langle P_a(s), v\rangle - \varepsilon' \ge \rho + v(s) - (\varepsilon + \varepsilon')$. Therefore,

$$P_{\tilde{\pi}}^{t-1} r_{\tilde{\pi}} = P_{\tilde{\pi}}^{t-1}(r_{\tilde{\pi}} + P_{\tilde{\pi}} v - P_{\tilde{\pi}} v) \ge P_{\tilde{\pi}}^{t-1}((\rho - (\varepsilon+\varepsilon'))\mathbf{1} + v - P_{\tilde{\pi}} v)\,.$$

Taking the average and the limit again shows that $\rho^{\tilde{\pi}}(s) \ge \rho - (\varepsilon + \varepsilon')$. The claim follows from combining this with the previous result.

(c) Let $\varepsilon = v + \rho\mathbf{1} - Tv$, which by the first constraint satisfies $\varepsilon \ge 0$. Let π^* be an optimal policy satisfying the requirements of the theorem statement and π be the greedy policy with respect to v. Then

$$P_{\pi^*}^t r_{\pi^*} \le P_{\pi^*}^t(r_\pi + P_\pi v - P_{\pi^*} v) = P_{\pi^*}^t(\rho\mathbf{1} + v - \varepsilon - P_{\pi^*} v)\,.$$

Hence $\rho^*\mathbf{1} = \rho^{\pi^*}\mathbf{1} \le \rho\mathbf{1} - P_{\pi^*}^* \varepsilon$, which means that $P_{\pi^*}^* \varepsilon \le \delta\mathbf{1}$. By the definition of $\tilde{s}$ there exists a state s with Therefore

$$\delta \ge (P_{\pi^*}\varepsilon)(s) = \sum_{s' \in \mathcal{S}} P_{\pi^*}(s, s')\varepsilon(s') \ge P_{\pi^*}(s, \tilde{s})\varepsilon(\tilde{s})$$

and so $\varepsilon(\tilde{s}) \le \delta|\mathcal{S}|$. Notice that $\tilde{v} = v - \varepsilon + \varepsilon(\tilde{s})\mathbf{1}$ also satisfies the constraints in Eq. (38.7) and hence

$$\langle v - \varepsilon + \varepsilon(\tilde{s})\mathbf{1}, \mathbf{1}\rangle \ge \langle v, \mathbf{1}\rangle\,,$$

which implies that $\langle \varepsilon, \mathbf{1}\rangle \le |\mathcal{S}|\varepsilon(\tilde{s}) \le |\mathcal{S}|^2\delta$. Hence (ρ, v) approximately satisfies the Bellman optimality equation.

38.17 Define operator $T : \mathbb{R}^{\mathrm{S}} \to \mathbb{R}^{\mathrm{S}}$ by $(Tu)(s) = \max_{a \in \mathcal{A}} r_a(s) + \langle P_a(s), u \rangle$, which is chosen so that $v_n^* = T^n \mathbf{0}$. Let v be a solution to the Bellman optimality equation with $\min_s v(s) = 0$. Then $Tv = \rho^* \mathbf{1} + v$ and

$$v_n^* = T^n \mathbf{0} \leq T^n v = n\rho^* \mathbf{1} + v \leq n\rho^* \mathbf{1} + D\mathbf{1}\,,$$

where the last inequality follows from the previous exercise and the assumption that $\min_s v(s) = 0$.

38.19

(a) There are four memoryless policies in this MDP. All are optimal except the policy π that always chooses the dashed action.

(b) The optimistic rewards are given by

$$\tilde{r}_{k,\mathrm{STAY}}(s) = \frac{1}{2} + \sqrt{\frac{L}{2(1 \vee T_{k-1}(s, \mathrm{STAY}))}}$$
$$\tilde{r}_{k,\mathrm{GO}}(s) = \sqrt{\frac{L}{2(1 \vee T_{k-1}(s, \mathrm{GO}))}}\,.$$

Whenever $T_{k-1}(s,a) \geq 1$ the transition estimates $\hat{P}_{k-1,a}(s) = P_a(s)$. Let S'_t be the state with $S_t \neq S'_t$. Suppose that $T_{k-1}(S_t, \mathrm{STAY}) > T_{k-1}(S'_t, \mathrm{STAY})$ and $\tilde{r}_{k,\mathrm{STAY}}(S_t) < 1$. Then $\tilde{r}_{k,\mathrm{STAY}}(S'_t) > \tilde{r}_{k,\mathrm{STAY}}(S_t)$. Once this occurs the optimal policy in the optimistic MDP is to choose action GO. It follows easily that once t is sufficiently large the algorithm will alternate between choosing actions GO and STAY and subsequently suffer linear regret. Note that the uncertainty in the transitions does not play a big role here. In the optimistic MDP they will always be chosen to maximize the probability of transitioning to the state with the largest optimistic reward.

38.21 Here we abuse notation by letting $\hat{P}_{u,a}(s)$ be the empirical next-state transitions after u visits to state-action pair (s,a). By a union bound and the result in Exercise 5.17,

$$\begin{aligned}
\mathbb{P}(F) &\leq \mathbb{P}\left(\exists u \in \mathbb{N}, s, a \in \mathcal{S} \times \mathcal{A} : \|P - \hat{P}_{u,a}(s)\|_1 \geq \sqrt{\frac{2\mathrm{S}\log(4\mathrm{SA}u(u+1)/\delta)}{u}}\right) \\
&\leq \sum_{s,a \in \mathcal{S} \times \mathcal{A}} \sum_{u=1}^{\infty} \mathbb{P}\left(\|P - \hat{P}_{u,a}(s)\|_1 \geq \sqrt{\frac{2\mathrm{S}\log(4\mathrm{SA}u(u+1)/\delta)}{u}}\right) \\
&\leq \sum_{s,a \in \mathcal{S} \times \mathcal{A}} \sum_{u=1}^{\infty} \frac{\delta}{2\mathrm{SA}u(u+1)} = \frac{\delta}{2}\,.
\end{aligned}$$

The statement in Exercise 5.17 makes an independence assumption that is not exactly satisfied here. We are saved by the Markov property, which provides the conditional independence required.

38.22 If $\sum_{k=1}^{m-1} a_k \leq 1$, then $A_0 = A_1 = \cdots = A_{m-1} = 1$. Hence $a_m \leq 1$ also holds and using

$A_m \geq 1$,

$$\sum_{k=1}^{m} \frac{a_k}{\sqrt{A_{k-1}}} = \sum_{k=1}^{m-1} a_k \, + \, a_m \leq 1 + 1 \leq (\sqrt{2}+1) A_m \, .$$

Let us now use induction on m. As long as m is so that $\sum_{k=1}^{m-1} a_k \leq 1$, the previous argument covers us. Thus, consider any $m > 1$ such that $\sum_{k=1}^{m-1} a_k > 1$ and assume that the statement holds for $m-1$ (note that $m = 1$ implies $\sum_{k=1}^{m-1} a_k \leq 1$). Let $c = \sqrt{2}+1$. Then,

$$\begin{aligned}
\sum_{k=1}^{m} \frac{a_k}{\sqrt{A_{k-1}}} &\leq c\sqrt{A_{m-1}} + \frac{a_m}{\sqrt{A_{m-1}}} && \text{(split sum, induction hypothesis)} \\
&= \sqrt{c^2 A_{m-1} + 2ca_m + \frac{a_m^2}{A_{m-1}}} \\
&\leq \sqrt{c^2 A_{m-1} + (2c+1)a_m} && (a_m \leq A_{m-1}) \\
&= c\sqrt{A_{m-1} + a_m} && \text{(choice of } c) \\
&= c\sqrt{A_m} \, . && (A_{m-1} \geq 1 \text{ and definition of } A_m)
\end{aligned}$$

38.23 We only outline the necessary changes to the proof of Theorem 38.6. The first step is to augment the failure event to include the event that there exists a phase k and state-action pair s, a such that

$$|\tilde{r}_{k,a}(s) - r_a(s)| \geq \sqrt{\frac{2L}{T_t(s,a)}} \, .$$

The likelihood of this event is at most $\delta/2$ by Hoeffding's bound combined with a union bound. Like in the proof of Theorem 38.6 we now restrict our attention to the regret on the event that the failure does not occur. The first change is that $r_{\pi_k}(s)$ in Eq. (38.18) must be replaced with $\tilde{r}_{k,\pi_k}(s)$. Then the reward terms no longer cancel in Eq. (38.20), which means that now

$$\begin{aligned}
\tilde{R}_k &= \sum_{t \in E_k} \left(-v_k(S_t) + \langle P_{k,A_t}(S_t), v_k \rangle + \tilde{r}_{k,A_t}(S_t) - r_{A_t}(S_t) \right) \\
&\leq \sum_{t \in E_k} \left(\langle P_{A_t}, v_k \rangle - v_k(S_t) \right) + \frac{D}{2} \sum_{t \in E_k} \| P_{k,A_t}(S_t) - P_{A_t}(S_t) \|_1 \\
&\qquad + \sum_{t \in E_k} \sqrt{\frac{2L}{1 \vee T_{\tau_k - 1}(S_t, A_t)}} \, .
\end{aligned}$$

The first two terms are the same as the proof of Theorem 38.6 and are bounded in the same way, which result in the same contribution to the regret. Only the last term is new. Summing over all

phases and applying the result from Exercise 38.22 and Cauchy-Schwarz,

$$\sum_{k=1}^{K} \sum_{t \in E_k} \sqrt{\frac{2L}{1 \vee T_{\tau_k - 1}(S_t, A_t)}} = \sqrt{2L} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \frac{T_{(k)}(s,a)}{\sqrt{1 \vee T_{\tau_k - 1}(s,a)}}$$
$$\leq (\sqrt{2}+1)\sqrt{2L\mathrm{SA}n}\,.$$

This term is small relative to the contribution due to the uncertainty in the transitions. Hence there exists a universal constant C such that with probability at least $1 - 3\delta/2$ the regret of this modified algorithm is at most

$$\hat{R}_n \leq CD(M)\mathrm{S}\sqrt{n\mathrm{A}\log\left(\frac{n\mathrm{SA}}{\delta}\right)}\,.$$

38.24

(a) An easy calculation shows that the depth d of the tree is bounded by $2 + \log_{\mathrm{A}} \mathrm{S}$, which by the conditions in the statement of the lower bound implies that $d + 1 \leq D/2$. The diameter is the maximum over all distinct pairs of states of the expected travel time between those two states. It is not hard to see that this is maximized by the pair s_g and s_b, so we restrict our attention to bounding the expected travel time between these two states under some policy. Let $\tau = \min\{t : S_t = s_b\}$ and let π be a policy that traverses the tree to a decision state with $\varepsilon(s,a) = 0$. We will show that for this policy

$$\mathbb{E}[\tau \,|\, S_1 = s_g] \leq D\,.$$

Let $X_1, X_2, \ldots$ be a sequence of random variables where $X_i \in \mathbb{N}^+$ is the number of rounds until the policy leaves state s_g on the ith series of visits to s_g. Then let M be the number of visits to state s_g before s_b is reached. All of these random variables are independent and geometrically distributed. An easy calculation shows that

$$\mathbb{E}[X_i] = 1/\delta \qquad\qquad \mathbb{E}[M] = 2\,.$$

Then $\tau = \sum_{i=1}^{M}(X_i + d + 1)$, which has expectation $\mathbb{E}[\tau] = 2(1/\delta + d + 1) \leq 2/\delta + D/2 \leq D$.

(b) The definition of stopping time τ ensures that $T_\sigma \leq n/D + 1 \leq 2n/D$ almost surely and hence $D\mathbb{E}[T_\sigma]/n \leq 2$ is immediate. For the second part note that

$$\mathbb{P}\left(\sum_{t=1}^{n/(2D)} D_t \geq \frac{n}{D}\right)$$

(c)

Bibliography

O. Kallenberg. *Foundations of modern probability*. Springer-Verlag, 2002. [10, 38, 66]

T. Needham. A visual explanation of jensen's inequality. *American Mathematical Monthly*, 100(8): 768–771, 1993. [45]

G. Peskir and A. Shiryaev. *Optimal stopping and free-boundary problems*. Springer, 2006. [70]