

# Solutions to Selected Exercises in Bandit Algorithms

Tor Lattimore and Csaba Szepesvári

Draft of Friday 11<sup>th</sup> September, 2020

## Contents

<b>2</b>	<b>Foundations of Probability</b>	<b>5</b>
	2.1, 2.3, 2.5, 2.6, 2.7, 2.8, 2.9, 2.10, 2.11, 2.12, 2.14, 2.15, 2.16, 2.18, 2.19	
<b>3</b>	<b>Stochastic Processes and Markov Chains</b>	<b>11</b>
	3.1, 3.5, 3.8, 3.9	
<b>4</b>	<b>Stochastic Bandits</b>	<b>13</b>
	4.9, 4.10	
<b>5</b>	<b>Concentration of Measure</b>	<b>14</b>
	5.10, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, 5.18, 5.19	
<b>6</b>	<b>The Explore-then-Commit Algorithm</b>	<b>22</b>
	6.2, 6.3, 6.5, 6.6, 6.8	
<b>7</b>	<b>The Upper Confidence Bound Algorithm</b>	<b>26</b>
	7.1	
<b>8</b>	<b>The Upper Confidence Bound Algorithm: Asymptotic Optimality</b>	<b>27</b>
	8.1	
<b>9</b>	<b>The Upper Confidence Bound Algorithm: Minimax Optimality</b>	<b>28</b>
	9.1, 9.4	
<b>10</b>	<b>The Upper Confidence Bound Algorithm: Bernoulli Noise</b>	<b>29</b>
	10.1, 10.3, 10.4, 10.5	
<b>11</b>	<b>The Exp3 Algorithm</b>	<b>35</b>
	11.2, 11.5, 11.6, 11.7	
<b>12</b>	<b>The Exp3-IX Algorithm</b>	<b>37</b>
	12.1, 12.4	
<b>13</b>	<b>Lower Bounds: Basic Ideas</b>	<b>39</b>
	13.2	
<b>14</b>	<b>Foundations of Information Theory</b>	<b>39</b>
	14.4, 14.10, 14.11	
<b>15</b>	<b>Minimax Lower Bounds</b>	<b>41</b>
	15.1	
<b>16</b>	<b>Instance-Dependent Lower Bounds</b>	<b>42</b>
	16.2, 16.7	
<b>17</b>	<b>High-Probability Lower Bounds</b>	<b>43</b>
	17.1	

<b>18</b>	<b>Contextual Bandits</b>	<b>44</b>
	18.1, 18.6, 18.7, 18.8, 18.9	
<b>19</b>	<b>Stochastic Linear Bandits</b>	<b>47</b>
	19.3, 19.4, 19.5, 19.6, 19.7, 19.8	
<b>20</b>	<b>Confidence Bounds for Least Squares Estimators</b>	<b>54</b>
	20.1, 20.2, 20.3, 20.4, 20.5, 20.8, 20.9, 20.10, 20.11	
<b>21</b>	<b>Optimal Design for Least Squares Estimators</b>	<b>58</b>
	21.1, 21.2, 21.3, 21.5	
<b>22</b>	<b>Stochastic Linear Bandits for Finitely Many Arms</b>	<b>60</b>
<b>23</b>	<b>Stochastic Linear Bandits with Sparsity</b>	<b>60</b>
	23.2	
<b>24</b>	<b>Minimax Lower Bounds for Stochastic Linear Bandits</b>	<b>60</b>
	24.1	
<b>25</b>	<b>Asymptotic Lower Bounds for Stochastic Linear Bandits</b>	<b>61</b>
	25.3	
<b>26</b>	<b>Foundations of Convex Analysis</b>	<b>61</b>
	26.2, 26.3, 26.9, 26.13, 26.14, 26.15	
<b>27</b>	<b>Exp3 for Adversarial Linear Bandits</b>	<b>64</b>
	27.1, 27.4, 27.6, 27.8, 27.9, 27.11	
<b>28</b>	<b>Follow-the-Regularised-Leader and Mirror Descent</b>	<b>69</b>
	28.1, 28.5, 28.10, 28.11, 28.12, 28.13, 28.14, 28.15, 28.16, 28.17	
<b>29</b>	<b>The Relation between Adversarial and Stochastic Linear Bandits</b>	<b>77</b>
	29.2, 29.4	
<b>30</b>	<b>Combinatorial Bandits</b>	<b>78</b>
	30.4, 30.5, 30.6, 30.8	
<b>31</b>	<b>Non-stationary Bandits</b>	<b>81</b>
	31.1, 31.3	
<b>32</b>	<b>Ranking</b>	<b>82</b>
	32.2, 32.6	
<b>33</b>	<b>Pure Exploration</b>	<b>84</b>
	33.3, 33.4, 33.5, 33.6, 33.7, 33.9	
<b>34</b>	<b>Foundations of Bayesian Learning</b>	<b>90</b>
	34.4, 34.5, 34.13, 34.14, 34.15, 34.16	
<b>35</b>	<b>Bayesian Bandits</b>	<b>95</b>
	35.1, 35.2, 35.3, 35.6, 35.7	
<b>36</b>	<b>Thompson Sampling</b>	<b>100</b>
	36.3, 36.5, 36.6, 36.13	
<b>37</b>	<b>Partial Monitoring</b>	<b>103</b>
	37.3, 37.10, 37.12, 37.13, 37.14	
<b>38</b>	<b>Markov Decision Processes</b>	<b>106</b>
	38.2, 38.4, 38.5, 38.7, 38.8, 38.9, 38.10, 38.11, 38.12, 38.13, 38.14, 38.15, 38.16, 38.17, 38.19, 38.21, 38.22, 38.23, 38.24	





## 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  $\Sigma_X$  is a  $\sigma$ -algebra we have  $V^c \in \Sigma$  and by definition of  $\Sigma_X$  we have  $U^c = X^{-1}(V^c) \in \Sigma_X$ . Therefore  $\Sigma_X$  is closed under complements. Finally let  $(U_i)_i$  be a countable sequence with  $U_i \in \Sigma_X$ . Then  $\bigcup_i U_i = X^{-1}(\bigcup_i X(U_i)) \in \Sigma_X$ , which means that  $\Sigma_X$  is closed under countable unions and the proof is completed.

### 2.5

(a) Let  $\mathcal{A}$  be the set of all  $\sigma$ -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  $\sigma$ -algebra containing  $\mathcal{G}$ . Clearly  $\mathcal{F}^*$  contains  $\mathcal{G}$  and is contained in all  $\sigma$ -algebras containing  $\mathcal{G}$ . Furthermore, by definition it contains exactly those  $A$  that are in every  $\sigma$ -algebra that contains  $\mathcal{G}$ . It remains to show that  $\mathcal{F}^*$  is a  $\sigma$ -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  $\sigma$ -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  $\sigma$ -algebra on  $\Omega$  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

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

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,  $\Omega$  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  $\mathcal{F} = 2^\Omega$ .

$\Omega$  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  $\Omega$  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  $i$ th 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  $\sigma$ -algebra is generated by all open sets and so  $f^{-1}(A) \in \mathfrak{B}(\mathbb{R})$  as required.
- (b) Since  $|\cdot| : \mathbb{R} \rightarrow \mathbb{R}$  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 and 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

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

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

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

**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

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

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 non-negative independent random variables. Then

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

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

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

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

**2.19** Using the fact that  $\int_0^a 1 dx = a$  for  $a \geq 0$  and the non-negativity of  $X$  we have

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

Then by Fubini's theorem,

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

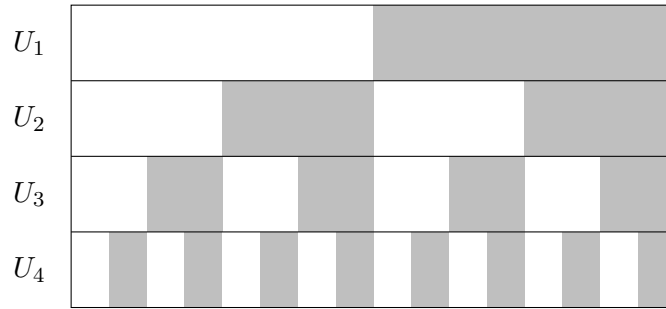
## Chapter 3 Stochastic Processes and Markov Chains

### 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

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

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

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

Let  $k = \max K$ . Then

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

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  $\mathbb{P}(X_{m,t} = 0) = \mathbb{P}(X_{m,t} = 1) = 1/2$  follows from Part (b).

- (e) By the previous parts  $X_t = \sum_{m=1}^{\infty} X_{m,t} 2^{-t}$  is a weighted sum of an independent sequence of uniform Bernoulli 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  $\sigma$ -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  $\lim_{n \rightarrow \infty} U_n(\omega) = \int_{\mathcal{X}} \lim_{n \rightarrow \infty} X_n(x) K(\omega, dx) = \int_{\mathcal{X}} X(x) K(\omega, dx) = U(\omega)$ . Hence  $\lim_{n \rightarrow \infty} U_n(\omega) = U(\omega)$  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  $\mathbb{E}[X_t \mathbb{I}\{\tau = t\}] \leq \mathbb{E}[\mathbb{I}\{\tau = t\} X_n]$ .

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

**3.9** Let  $\Sigma_{\mathcal{X}}$  ( $\Sigma_{\mathcal{Y}}$ ) be the  $\sigma$ -algebra underlying  $\mathcal{X}$  (respectively,  $\mathcal{Y}$ ). It suffices to verify that for

$A \in \Sigma_{\mathcal{X}}, B \in \Sigma_{\mathcal{Y}}, \mathbb{P}_{(X,Y)}(A \times B) = (\mathbb{P}_Y \otimes \mathbb{P}_{X|Y})(A \times B)$ . We have

$$\begin{aligned}
\mathbb{P}_{(X,Y)}(A \times B) &= \mathbb{P}(X \in A, Y \in B) \\
&= \mathbb{E}[\mathbb{E}[\mathbb{I}\{X \in A\} \mathbb{I}\{Y \in B\} \mid Y]] && \text{(tower rule)} \\
&= \mathbb{E}[\mathbb{I}\{Y \in B\} \mathbb{E}[\mathbb{I}\{X \in A\} \mid Y]] && (\mathbb{I}\{Y \in B\} \text{ is } \sigma(Y)\text{-measurable}) \\
&= \mathbb{E}[\mathbb{I}\{Y \in B\} \mathbb{P}(X \in A \mid Y)] && \text{(relation of expectation and probability)} \\
&= \mathbb{E}[\mathbb{I}\{Y \in B\} \mathbb{P}_{X|Y}(X \in A \mid Y)] && \text{(definition of } \mathbb{P}_{X|Y}) \\
&= \int_B \mathbb{P}_Y(dy) \mathbb{P}_{X|Y}(X \in A \mid y) && \text{(pushforward property)} \\
&= (\mathbb{P}_Y \otimes \mathbb{P}_{X|Y})(B \times A). && \text{(definition of } \otimes)
\end{aligned}$$

## Chapter 4 Stochastic Bandits

### 4.9

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

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

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

(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  $t \in \{1, 2, 4, 8, 16, \dots\}$ .

### 4.10

(a) (Sketch) Fix policy  $\pi$  and  $n$  and assume without loss of generality the reward-stack model. We turn policy  $\pi$  into a retirement policy  $\pi'$  by reshuffling the order in which  $\pi$  uses the two arms during the  $n$  rounds so that if  $\pi$  uses action 1, say  $m$  times out of the  $n$  rounds,  $\pi'$  will use action 1 in the first  $m$  rounds and then switches to arm 2. By using the regret decomposition, we see that this suffices to show that  $\pi'$  achieves no more regret than  $\pi$  (actually, achieves the same regret).

So what is policy  $\pi'$ ? Policy  $\pi'$  will keep querying policy  $\pi$  for at most  $n$  times. If  $\pi$  returns by proposing to use action 1,  $\pi'$  will play this action, get the reward from the environment and feeds the obtained reward to policy  $\pi$ . If  $\pi$  returns by proposing to play action 2,  $\pi'$  does not play this action for now, just feeds  $\pi$  with zero. After  $\pi$  was queried  $n$  times, action 2 is played up in the remaining rounds out of the total  $n$  rounds.

(b) Assume that arm 1 has a Bernoulli payoff with parameter  $p \in [0, 1]$  and arm 2 has a fixed payoff of 0.5 (so  $\mu_1 = p$  and  $\mu_2 = 0.5$ ). Note that whether  $\pi$  ever retires on these Bernoulli

environments depends on whether there exists some  $t > 0$  and  $x_1, \dots, x_{t-1} \in \{0, 1\}$  such that  $\pi_t(2|1, x_1, \dots, 1, x_{t-1}) > 0$ , or

$$\sup_{t>0} \sup_{x_1, \dots, x_{t-1} \in \{0, 1\}} \pi_t(2|1, x_1, \dots, 1, x_{t-1}) > 0. \quad (4.1)$$

We have the two cases. When (4.1) does not hold then  $\pi$  will have linear regret when  $p < 0.5$ . When (4.1) does hold then take the  $t > 0$  and  $x_1, \dots, x_{t-1} \in \{0, 1\}$  such that  $\rho = \pi_t(2|1, x_1, \dots, 1, x_{t-1}) > 0$  (these must exist). Assume that  $t > 0$  is smallest possible: Hence,  $\pi_s(1|1, x'_1, \dots, 1, x'_{s-1}) = 1$  for any  $s < t$  and  $x'_1, \dots, x'_{s-1} \in \{0, 1\}$ . Now, take an environment when  $p > 0.5$  (so arm 1 is the optimal arm) and let  $R_n$  denote the regret of  $\pi$  in this environment. Then letting  $\Delta = p - 0.5 > 0$ , we have

$$\begin{aligned} R_n &= \Delta \mathbb{E}[T_2(n)] \\ &\geq \Delta \mathbb{E}[\mathbb{I}\{A_1 = 1, X_1 = x_1, \dots, A_{t-1} = 1, X_{t-1} = x_{t-1}, A_t = 2\} T_2(n)] \\ &= \Delta \mathbb{E}[\mathbb{I}\{A_1 = 1, X_1 = x_1, \dots, A_{t-1} = 1, X_{t-1} = x_{t-1}, A_t = 2\} (n - t + 1)] \\ &= \Delta \mathbb{P}(A_1 = 1, X_1 = x_1, \dots, A_{t-1} = 1, X_{t-1} = x_{t-1}, A_t = 2) (n - t + 1) \\ &= \Delta (n - t + 1) \rho \prod_{s=1}^t p^{x_s} (1 - p)^{1-x_s} \\ &\geq c(n - t + 1), \end{aligned}$$

where  $c = \Delta \rho \prod_{s=1}^{t-1} p^{x_s} (1 - p)^{1-x_s} > 0$ . It follows that  $\liminf_{n \rightarrow \infty} R_n/n \geq c > 0$ .

## Chapter 5 Concentration of Measure

### 5.10

(a) The Cramér-Chernoff method gives that for any  $\lambda \geq 0$ ,

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

Taking logarithm of both sides and reordering gives

$$\frac{1}{n} \log \mathbb{P}(\hat{\mu}_n \geq \varepsilon) \leq -(\lambda\varepsilon - M_X(\lambda)).$$

Since this holds for any  $\lambda \geq 0$ , taking the supremum over  $\lambda \in \mathbb{R}$  gives the desired inequality (allowing  $\lambda < 0$  potentially makes the resulting inequality loser).

(b) Let  $X$  be a Rademacher variable. We have  $\psi_X(\lambda) = \frac{1}{2}(\exp(-\lambda) + \exp(\lambda)) = \cosh(\lambda)$ . To get the Fenchel dual of  $\log \psi_X$ , we find the maximum value of  $f(\lambda) = \lambda\varepsilon - \log \psi_X(\lambda) =$

$\lambda\varepsilon - \log \cosh(\lambda)$ . We have  $\frac{d}{d\lambda} \log \cosh(\lambda) = (e^\lambda - e^{-\lambda})/(e^\lambda + e^{-\lambda}) = \tanh(\lambda) \in [-1, 1]$ . Hence,  $\sup_\lambda f(\lambda) = +\infty$  when  $|\varepsilon| > 1$ . In the other case, we get that the maximum is  $\psi_X^*(\varepsilon) = f(\tanh^{-1}(\varepsilon)) = \tanh^{-1}(\varepsilon)\varepsilon - \log \cosh(\tanh^{-1}(\varepsilon))$ . Using  $\tanh^{-1}(\varepsilon) = \frac{1}{2} \log(\frac{1+\varepsilon}{1-\varepsilon})$  we find that  $e^{\tanh^{-1}(\varepsilon)} = (\frac{1+\varepsilon}{1-\varepsilon})^{1/2}$  and  $e^{-\tanh^{-1}(\varepsilon)} = (\frac{1-\varepsilon}{1+\varepsilon})^{1/2}$ , hence  $\cosh(\tanh^{-1}(\varepsilon)) = \frac{1}{2}((\frac{1+\varepsilon}{1-\varepsilon})^{1/2} + (\frac{1-\varepsilon}{1+\varepsilon})^{1/2}) = \frac{1}{2}(\frac{(1+\varepsilon)+(1-\varepsilon)}{(1-\varepsilon^2)^{1/2}}) = \frac{1}{\sqrt{1-\varepsilon^2}}$ . Therefore,  $\psi_X^*(\varepsilon) = \frac{\varepsilon}{2} \log(\frac{1+\varepsilon}{1-\varepsilon}) + \frac{1}{2} \log(1-\varepsilon^2) = \frac{1+\varepsilon}{2} \log(1+\varepsilon) + \frac{1-\varepsilon}{2} \log(1-\varepsilon)$ .

(c) We have  $\psi_X(\lambda) = \lambda(p + \varepsilon) - \log(1 - p + pe^\lambda)$ . The maximiser of this is  $\lambda^* = \log(\frac{(1-p)(p+\varepsilon)}{p(1-(p+\varepsilon))})$  provided that  $p + \varepsilon < 1$ . Plugging in this value, after some algebra, gives the desired result. The result also extends to  $p + \varepsilon = 1$ : In this case  $\psi_X$  is increasing and  $\lim_{\lambda \rightarrow \infty} \lambda - \log(1 - p + pe^\lambda) = \lim_{\lambda \rightarrow \infty} \lambda - \log(pe^\lambda) = \log(1/p) = d(1, p)$ . For  $\varepsilon > 0$  so that  $p + \varepsilon > 1$ ,  $\psi_X^*(\varepsilon) = +\infty$  because as  $\lambda \rightarrow \infty$ ,  $\lambda(p + \varepsilon) - \log(1 - p + pe^\lambda) \sim \lambda(p + \varepsilon - 1) \rightarrow \infty$ .

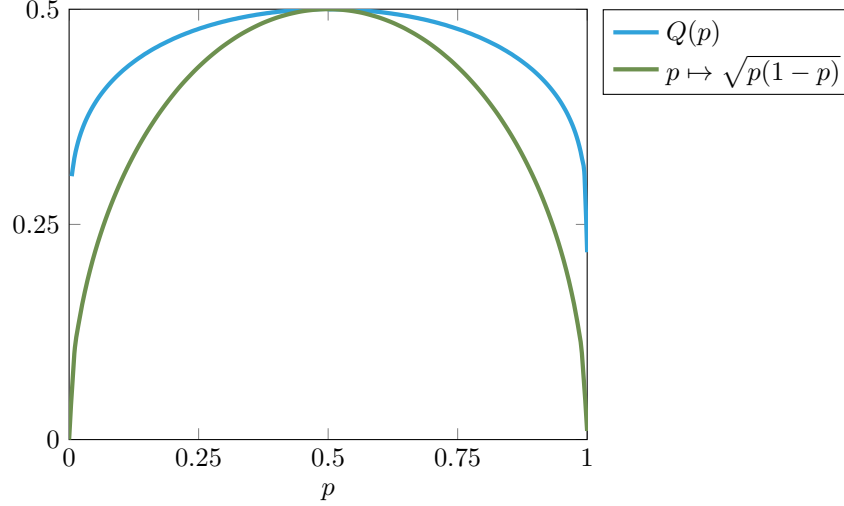
(d) Set  $\sigma = 1$  for simplicity. We have  $M_X(\lambda) = \frac{1}{\sqrt{2\pi}} \int \exp(-(x^2 - 2\lambda x)/2) dx = \frac{1}{\sqrt{2\pi}} \int \exp(-(x - \lambda)^2/2) \exp(\lambda^2/2) dx = \exp(\lambda^2/2)$ . Hence,  $f(\lambda) = \lambda\varepsilon - \log M_X(\lambda) = \lambda\varepsilon - \lambda^2/2$  and  $\sup_\lambda f(\lambda) = f(2\varepsilon) = \varepsilon^2/2$ .

(e) We need to calculate  $\lim_{n \rightarrow \infty} \frac{1}{n} \log(1 - \Phi(\varepsilon\sqrt{n/\sigma^2}))$ . By Eq. (5.3) we have  $1 - \Phi(x) \leq \sqrt{1/(2\pi x^2)} \exp(-x^2/2)$ . Further, by Eq. (13.4),  $1 - \Phi(x) \geq \exp(-x^2/2)/(\sqrt{\pi}(x/\sqrt{2} + \sqrt{x^2/2 + 2}))$ . Taking logarithm, plugging in  $x = \varepsilon\sqrt{n/\sigma^2}$ , dividing by  $n$  and taking  $n \rightarrow \infty$  gives

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log(1 - \Phi(\varepsilon\sqrt{n/\sigma^2})) = \varepsilon^2/(2\sigma^2).$$

When  $X$  is a Rademacher random variable,  $\psi_X^*(\varepsilon) = \frac{1+\varepsilon}{2} \log(1+\varepsilon) + \frac{1-\varepsilon}{2} \log(1-\varepsilon) \geq \varepsilon^2/2$  for any  $\varepsilon \in \mathbb{R}$ , with equality holding only at  $\varepsilon = 0$ . Hence, the question-marked equality cannot hold. (In fact, this is very easy to see also by noting that if  $X$  is supported on  $[-1, 1]$  then  $\hat{\mu}_n \in [-1, 1]$  almost surely and thus  $\mathbb{P}(\hat{\mu}_n > \varepsilon) = 0$  for any  $\varepsilon \geq 1$ , while the approximation from the CLT gives  $\varepsilon^2/(2\sigma^2)$ , a strictly larger value: The CLT can significantly overestimate tail probabilities. What goes wrong with the careless application of the (strong form) of the CLT is that  $\lim_{n \rightarrow \infty} \sup_x |f_n(x) - f(x)| = 0$  does not imply  $|\log f_n(x_n) - \log f(x_n)| = o(n)$  for all choices of  $\{x_n\}$ . For example, one can take  $f(x) = \exp(-x)$ ,  $f_n(x) = \exp(-x(1 + 1/n))$  so that  $\log f_n(x) - \log f(x) = -x/n$ . Then,  $|\log f_n(n^2) - \log f(n^2)| = n \neq o(n)$ . The same problem happens in the specific case that was investigated.

**5.12** Part (d) The plots of  $p \mapsto Q(p)$  and  $p \mapsto \sqrt{p(1-p)}$  are shown below:



As can be seen,  $\sqrt{p(1-p)} \leq Q(p)$  for  $p \in [0, 1]$ .

Part (e): Consider  $0 \leq \lambda < 4$  and  $\lambda \geq 4$  separately. In the latter case use  $\lambda^2 \geq 4\lambda$ . For the former case consider the extremes  $p = 1$  and  $p = 1/2$  and then use convexity. The general conclusion is that the subgaussianity constant may be misleadingly large when it comes to studying tails of distributions: Tail bounds (for the upper tail) only need bounds on the MGF for nonnegative values of  $\lambda$ !

### 5.13

(a) Using linearity of expectation  $\mathbb{E}[\hat{p}_n] = \mathbb{E}[\sum_{t=1}^n X_t/n] = \sum_{t=1}^n \mathbb{E}[X_t]/n = p$ . Similarly,  $\mathbb{V}[\hat{p}_n] = p(1-p)/n$ .

(b) The central limit theorem says that

$$\lim_{n \rightarrow \infty} [\mathbb{P}(\sqrt{n}(\hat{p}_n - p) \geq x) - \mathbb{P}(\sqrt{n}Z_n \geq x)] = 0 \quad \text{for all } x \in \mathbb{R}. \quad (5.1)$$

(c) This is an empirical question, the solution to which we omit. You should do this calculation directly using the binomial distribution.

(d. i) Let  $d(p, q) = p \log(p/q) + (1-p) \log((1-p)/(1-q))$  be the relative entropy between Bernoulli distributions with means  $p$  and  $q$ . By large deviation theory (see Exercise 5.10),

$$\begin{aligned} \mathbb{P}(\hat{p}_n \geq p + \Delta) &= \exp(-n(d_{\text{Ber}} + \varepsilon_n)) \quad \text{and} \\ \mathbb{P}(Z_n \geq p + \Delta) &= \exp(-n(d_{\text{Gauss}} + \xi_n)), \end{aligned}$$

where  $d_{\text{Ber}} = d(p + \Delta, p)$  and  $d_{\text{Gauss}} = \Delta^2/(2p(1-p))$  and  $(\varepsilon_n)_{n=1}^\infty$  and  $(\xi_n)_{n=1}^\infty$  satisfy  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$  and  $\lim_{n \rightarrow \infty} \xi_n = 0$ . It should be clear that  $\lim_{\delta \rightarrow 0} n_i(\delta, p, \Delta) = \infty$  for  $i \in \{\text{Ber}, \text{Gauss}\}$ . Hence, by inverting the above displays we arrive at

$$n_i(\delta, p, \Delta) = \left( \frac{1}{d_i} + o(1) \right) \log \left( \frac{1}{\delta} \right), \quad (5.2)$$



where the  $o(1)$  term vanishes as  $\delta$  tends to zero (see below for the precise argument). Therefore when  $\Delta = p = 1/10$ ,

$$\lim_{\delta \rightarrow 0} \frac{n_{\text{Ber}}(\delta, p, \Delta)}{n_{\text{Gauss}}(\delta, p, \Delta)} = \frac{d_{\text{Gauss}}}{d_{\text{Ber}}} = \frac{\Delta^2/(2p(1-p))}{d(p + \Delta, p)} \approx 1.2512.$$

It remains to see the validity of Eq. (5.2). This follows from an elementary but somewhat tedious argument. The precise claim is as follows: Let  $(p_n)$  be a sequence taking values in  $[0, 1]$ ,  $n(\delta) = \min\{n \geq 1 : p_n \leq \delta\}$  such that  $n(\delta) \rightarrow \infty$  as  $\delta \rightarrow 0$  and  $\log(1/p_n) = n(d + o(1))$ . We claim that from these it follows that  $n(\delta) = (1/d + o(1)) \log(1/\delta)$  as  $\delta \rightarrow 0$ . To show this it suffices to prove that for any  $\varepsilon > 0$ , for any  $\delta > 0$  small enough,  $n(\delta) \in [1/(d + \varepsilon) \log(1/\delta), 1/(d - \varepsilon) \log(1/\delta) + 1]$ . Fix  $\varepsilon > 0$ . Then, by our assumption on  $\log(1/p_n)$  there exist some  $n_0 > 0$  such that for any  $n \geq n_0$ ,  $\log(1/p_n) \in [n(d - \varepsilon), n(d + \varepsilon)]$ . Further, by our assumption on  $n(\delta)$ , there exists  $\delta_0 > 0$  such that for any  $\delta < \delta_0$ ,  $n(\delta) - 1 \geq n_0$ . Take some  $\delta < \delta_0$  and let  $n' = n(\delta)$ . By definition,  $p_{n'} \leq \delta < p_{n'-1}$  and hence  $\log(1/p_{n'}) \geq \log(1/\delta) > \log(1/p_{n'-1})$ . Since  $n' \geq n' - 1 \geq n_0$ , we also have that  $(n' - 1)(d - \varepsilon) \leq \log(1/p_{n'-1}) < \log(1/\delta) \leq \log(1/p_{n'}) \leq n'(d + \varepsilon)$ , from which it follows that  $\frac{\log(1/\delta)}{d + \varepsilon} \leq n' < \frac{\log(1/\delta)}{d - \varepsilon} + 1$ , finishing the proof.

- (d.ii) The central limit theorem only shows Eq. (5.1). In particular, you cannot choose  $x$  to depend on  $n$ . A second try is to use Berry-Esseen (Exercise 5.5) which warrants that

$$|\mathbb{P}(\hat{p}_n - p \geq \Delta) - \mathbb{P}(Z_n \geq \Delta)| = O(1/\sqrt{n}).$$

The problem is that this provides very little information in the regime where  $\Delta$  is fixed and  $n$  tends to infinity where both probabilities tend to zero exponentially fast and the error term washes away the comparison. In particular, for the inversion process to work, one needs nontrivial lower and upper bounds on  $\mathbb{P}(\hat{p}_n - p \geq \Delta)$  and the central limit theorem only asserts that this probability is in the range of  $[0, O(1/\sqrt{n})]$  (irrespective of the value of  $p$  and  $\Delta$ ), which does not lead to nontrivial bounds on  $n_{\text{Ber}}(\delta; p, \Delta)$ .

To summarise, the study of  $n_i(\delta, p, \Delta)$  as  $\delta$  tends to zero is a question about the large deviation regime, where the central limit theorem and Berry-Esseen do not provide meaningful information. To make use of the central limit theorem and Berry-Esseen, one needs to choose the deviation level  $x$  so that the probability  $\mathbb{P}(\sqrt{n}(\hat{p}_n - p) \geq x)$  is of a larger magnitude than  $O(1/\sqrt{n})$ , which is the range of ‘small deviations’.

As an aside, comparisons between normal and binomial distributions have been studied extensively. If you are interested, the most relevant lower bound for this discussion is Slud’s inequality [Slud, 1977].

## 5.14

- (a) We have  $g'(x) = \frac{x^2(\exp(x)-1)-2x(\exp(x)-1-x)}{x^4}$  so that  $x^3 g'(x) = h(x) = xe^x - 2e^x + 2 + x$ . We have  $h'(x) = xe^x - e^x + 1$  and  $h''(x) = xe^x$ . Hence,  $h'$  is increasing on  $(0, \infty)$  and decreasing on  $(-\infty, 0)$ . Since  $h(0) = 0$ , so  $\text{sign}(h(x)) = \text{sign}(x)$  and thus  $g'(x) > 0$  for  $x \neq 0$ .

(b) We have  $\exp(x) = 1 + x + g(x)x^2$ . Therefore,  $\mathbb{E}[\exp(X)] = 1 + \mathbb{E}[g(X)X^2] \leq 1 + \mathbb{E}[g(b)X^2] = 1 + g(b)\mathbb{V}[X]$ , where the last inequality used that  $g$  is increasing.

(c) Calculation – left to the reader.

(d) Let  $Z_t = X_t - \mathbb{E}X_t$  so that  $S = \sum_{t=1}^n Z_t$ . By the Cramér-Chernoff method, for any  $\lambda \geq 0$ ,

$$\mathbb{P}(S \geq \varepsilon) \leq \exp(-\lambda\varepsilon) \prod_{t=1}^n \mathbb{E}[\exp(\lambda Z_t)].$$

Using  $\mathbb{E}[\exp(\lambda Z_t)] \leq 1 + g(\lambda b)\lambda^2\mathbb{V}[Z_t] \leq \exp(g(\lambda b)\lambda^2\mathbb{V}[Z_t])$ , we get

$$\mathbb{P}\left(\sum_t Z_t \geq \varepsilon\right) \leq \exp(-\lambda\varepsilon + g(\lambda b)\lambda^2 v). \quad (5.3)$$

Differentiation shows that the exponent is minimised by  $\lambda = 1/b \log(1 + \alpha)$  where recall that  $\alpha = b\varepsilon/v$ . Plugging in this value we get (5.10) and then using the bound in Part ((c)) we get (5.11).

(e) We need to solve  $\delta = \exp\left(-\frac{\varepsilon^2}{2v(1+\frac{b\varepsilon}{3v})}\right)$  for  $\varepsilon \geq 0$ . Algebra gives that this is quadratic equation in  $\varepsilon$ : Using the abbreviation  $L = \log(1/\delta)$ , this quadratic equation is  $\varepsilon^2 - \frac{2}{3}bL\varepsilon - 2vL$ . The positive root is  $\varepsilon = \frac{1}{2}\left(\frac{2}{3}bL + \sqrt{(\frac{2}{3}bL)^2 + 8vL}\right)$ . Hence, with probability  $1 - \delta$ ,  $S \leq \varepsilon$ . Further upper bounding  $\varepsilon$  using  $\sqrt{|a|+|b|} \leq \sqrt{|a|} + \sqrt{|b|}$  gives that with probability  $1 - \delta$ ,  $S_n \leq \frac{2}{3}bL + \sqrt{2vL}$ , which is the desired inequality.

(f) We start by modifying the Cramér-Chernoff method. In particular, consider the problem of bounding the probability of event  $A$  where for a random vector  $X \in \mathbb{R}^d$  and a fixed vector  $x \in \mathbb{R}^d$ ,  $A$  takes the form  $A = \{X \geq x\}$ . Notice that for  $f : \mathbb{R}^d \rightarrow [0, \infty)$ ,  $\mathbb{P}(A) \leq \mathbb{E}[\mathbb{I}\{A\}e^{f(X)}] \leq \mathbb{E}[e^{f(X)}]$ . We use this with  $X = (S, -V)$  and  $x = (\varepsilon, v)$  so that  $A = \{S \geq \varepsilon, V \leq v\} = \{X \geq (\varepsilon, v)\}$ . Then, for  $\lambda > 0$  letting  $h(S, V) = \lambda S - g(\lambda b)\lambda^2 V$  we have on  $A$  that  $h(S, V) \geq h(\varepsilon, v)$  and so  $f(S, V) = h(S, V) - h(\varepsilon, v) \geq 0$  and  $\mathbb{P}(A) \leq e^{-h(\varepsilon, v)}\mathbb{E}[e^{h(S, V)}]$ . We have  $e^{h(S, V)} = U_1 \dots U_n$  where  $U_t = e^{\lambda Z_t - \lambda^2 g(\lambda b)\mathbb{E}_{t-1}[Z_t^2]}$  and  $Z_t = X_t - \mu_t = X_t - \mathbb{E}_{t-1}[X_t]$ . Furthermore, owing to  $\lambda > 0$ ,  $\lambda Z_t \leq \lambda b$ , hence

$$\begin{aligned} \mathbb{E}_{s-1}[U_s] &= e^{-\lambda^2 g(\lambda b)\mathbb{E}_{s-1}[Z_s^2]} \mathbb{E}_{s-1}[e^{\lambda Z_s}] \\ &\leq e^{-\lambda^2 g(\lambda b)\mathbb{E}_{s-1}[Z_s^2]} (1 + g(\lambda b)\mathbb{E}_{s-1}[(\lambda Z_s)^2]) \\ &= e^{-\lambda^2 g(\lambda b)\mathbb{E}_{s-1}[Z_s^2]} e^{\lambda^2 g(\lambda b)\mathbb{E}_{s-1}[Z_s^2]} = 1, \end{aligned}$$

and thus

$$\begin{aligned} \mathbb{E}[e^{g(S_n, V_n)}] &= \mathbb{E}\left[\prod_{t=1}^n U_t\right] = \mathbb{E}[U_1 \dots U_{n-1}\mathbb{E}_{n-1}[U_n]] \\ &\leq \mathbb{E}[U_1 \dots U_{n-1}] \leq \dots \leq 1. \end{aligned}$$

Thus,  $\mathbb{P}(A) \leq e^{-h(\varepsilon, v)}$ . Notice that the expression on the right-hand side is the same as in Eq. (5.3), finishing the proof.

**5.15** Let  $\alpha_t = \eta \mathbb{E}_{t-1}[(X_t - \mu_t)^2]$ . We use the Cramér-Chernoff method:

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

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

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

Therefore  $\exp(\eta \sum_{t=1}^n (X_t - \mu_t - \alpha_t))$  is a supermartingale, which completes the proof of Part (a). The proof of Part (b) follows in the same fashion.

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

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

Applying the Cramér-Chernoff method,

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

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

**5.17** The 1-norm can be re-written as

$$\|p - \hat{p}\|_1 = \max_{\lambda \in \{-1, 1\}^m} \langle \lambda, p - \hat{p} \rangle.$$

Next, let  $\lambda \in \{-1, 1\}^m$  be fixed. Then,

$$\langle \lambda, p - \hat{p} \rangle = \frac{1}{n} \sum_{t=1}^n \langle \lambda, p - e_{X_t} \rangle.$$

Now,  $|\langle \lambda, p - e_{X_t} \rangle| \leq \|\lambda\|_\infty \|p - e_{X_t}\|_1 \leq 2$  and  $\mathbb{E}[\langle \lambda, p - e_{X_t} \rangle] = 0$ . Then, by Hoeffding's bound,

$$\mathbb{P} \left( \langle \lambda, p - \hat{p} \rangle \geq \sqrt{\frac{2}{n} \log \left( \frac{1}{\delta} \right)} \right) \leq \delta.$$

Taking a union bound over all  $\lambda \in \{-1, 1\}^m$  shows that

$$\mathbb{P} \left( \max_{\lambda \in \{-1, 1\}^m} \langle \lambda, p - \hat{p} \rangle \geq \sqrt{\frac{2}{n} \log \left( \frac{2^m}{\delta} \right)} \right) \leq \delta.$$

**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  $\mathbb{E}[Z] \leq \sqrt{2 \sigma^2 \log(n)}$ . For Part (b), a union bound in combination with Theorem 5.3 suffices.

**5.19** Let  $\mathcal{P}$  be the set of measures on  $([0, 1], \mathfrak{B}([0, 1]))$  and for  $q \in \mathcal{P}$  let  $\mu_q$  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

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

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  $q \in \mathcal{P}$ ,

$$F_n(q) \doteq \int_0^1 f_{n-1} \left( \frac{x - \mu_q}{1 - y} \right) dq(y) \leq f_n(x). \quad (5.4)$$

Let  $q \in \mathcal{P}$  have mean  $\mu$  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

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

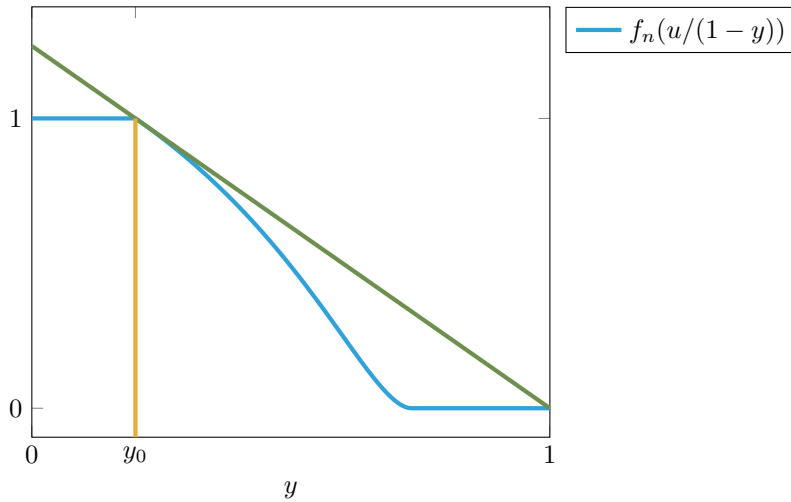
**Lemma 5.1.** *Suppose that  $n \geq 1$  and  $u \in (0, n]$  and  $y_0 = \max(0, 1-u)$ . Then*

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

*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  $f_n(u/(1-y)) = \mathbb{I}\{y \leq 1-u\}$ . 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

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

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.  $\square$



## Chapter 6 The Explore-then-Commit Algorithm

**6.2** If  $\Delta \leq 1/\sqrt{n}$  then from  $R_n \leq n\Delta$  we get  $R_n \leq \sqrt{n}$ . Now, if  $\Delta > 1/\sqrt{n}$  then from  $R_n \leq \Delta + \frac{4}{\Delta} \left(1 + \log_+ \left(\frac{n\Delta^2}{4}\right)\right) \leq \Delta + 4\sqrt{n} + \max_{x>0} \frac{1}{x} \log_+(nx^2/4)$ . A simple calculation shows that  $\max_{x>0} \frac{1}{x} \log_+(nx^2/4) = e^{-2}\sqrt{n}$ . Putting things together we get that  $R_n \leq \Delta + (4 + e^{-2})\sqrt{n}$  holds no matter the value of  $\Delta > 0$ .

**6.3** Assume for simplicity that  $n$  is even and let  $\Delta = \max\{\Delta_1, \Delta_2\}$  and

$$m = \min \left\{ n/k, \frac{4}{\Delta^2} \log \left( \frac{1}{\delta} \right) \right\}.$$

When  $2m = n$  the pseudo regret is bounded by  $\bar{R}_n \leq m\Delta$ . Now suppose that  $2m < n$ . Then

$$\mathbb{P}(T_2(n) > m) \leq \mathbb{P}(\hat{\mu}_2(2m) - \mu_2 - \hat{\mu}_1(2m) + \mu_1 \geq \Delta) \leq \exp \left( -\frac{m\Delta^2}{4} \right) \leq \delta.$$

Hence with probability at least  $1 - \delta$  the pseudo regret is bounded by

$$\bar{R}_n \leq m\Delta = \min \left\{ \frac{n\Delta}{2}, \frac{4}{\Delta} \log \left( \frac{1}{\delta} \right) \right\}.$$

**6.5** By slightly abusing notation, to reduce clutter we abbreviate  $R_n(\nu)$  to  $R_n$  and  $\Delta_\nu$  to  $\Delta$ .

- (a) For the first part we need to show that  $R_n \leq (\Delta + C)n^{2/3}$  when  $m = f(n)$  for a suitable chosen function  $f$  and  $C > 0$  is a universal constant. By Eq. (6.4),  $R_n \leq m\Delta + n\Delta \exp \left( -\frac{m\Delta^2}{4} \right) \leq m\Delta + n \max_{\Delta>0} \Delta \exp \left( -\frac{m\Delta^2}{4} \right) = m\Delta + 2n\sqrt{\frac{1}{m}} \exp \left( -\frac{1}{2} \right)$ , where the equality follows because  $\max_x x \exp(-cx^2)$  is at the value  $x^* = \sqrt{1/(2c)}$  and is equal to  $\sqrt{1/(2c)} \exp(-1/2)$  as a simple calculation shows (so,  $\Delta^* = \sqrt{4/(2m)} = \sqrt{2/m} = \sqrt{2/n^{2/3}} = \sqrt{2}n^{-1/3}$ ). That  $R_n \leq \Delta + Cn^{2/3}$  cannot hold follows because  $m\Delta/2 \leq R_n$ . Hence, if  $R_n \leq \Delta + Cn^{2/3}$  was also true, we would get that for any  $\Delta > 0$ ,  $m\Delta/2 \leq \Delta + Cn^{2/3}$  holds. Dividing both sides by  $\Delta$  and letting  $\Delta \rightarrow \infty$ , this would imply that  $m \leq 2$ . However, if  $m \leq 2$  then  $R_n = \Omega(n)$  on some instances: In particular, there is a positive probability that the arm chosen after trying both arms at most twice is the suboptimal arm.
- (b) For any fixed  $m$ ,  $\hat{\mu}_i(2m) - \mu_i$  is  $\sqrt{1/m}$  subgaussian. Hence defining  $G = \{|\hat{\mu}_i(2m) - \mu_i| \leq \sqrt{2 \log(n/\delta)/m}, i = 1, 2, m = 1, 2, \dots, \lfloor n/2 \rfloor\}$ , using  $n \geq 2\lfloor n/2 \rfloor$  union bounds, we have that  $\mathbb{P}(G) \geq 1 - \delta$ . Introduce  $w(m) = \sqrt{2 \log(n/\delta)/m}$ . Let  $M = \min\{1 \leq m \leq \lfloor n/2 \rfloor : |\hat{\mu}_1(2m) - \hat{\mu}_2(2m)| > 2w(m)\}$  (note that  $M = \infty$  if the condition is never met). Then on  $G$  if  $M < +\infty$  and say  $1 = \arg\max_i \hat{\mu}_i(2M)$  then  $\mu_1 \geq \hat{\mu}_1(2M) - w(M) > \hat{\mu}_2(2M) + 2w(M) - w(M) \geq \mu_2$  where the first and last inequalities used that we are on  $G$  and the middle one used the stopping condition and that we assumed that at stopping, arm one has the highest mean. Hence,

$R_n = \mathbb{P}(G^c) n\Delta/2 + \mathbb{E}[M\mathbb{I}\{G\}] \Delta/2 \leq \delta n + \mathbb{E}[M\mathbb{I}\{G\}] \Delta/2$ . We now show a bound on  $M$  on  $G$ . To reduce clutter assume that  $\mu_1 > \mu_2$ . Assume  $G$  holds and let  $m < M$ . Then,  $2w(m) \geq |\hat{\mu}_1(2m) - \hat{\mu}_2(2m)| \geq \hat{\mu}_1(2m) - \hat{\mu}_2(2m) \geq (\mu_1 - w(m)) - (\mu_2 + w(m)) = \Delta - 2w(m)$ . Reordering we see that  $4w(m) \geq \Delta$ , which, using the definition of  $w(m)$ , is equivalent to  $m \leq (4/\Delta)^2 2 \log(n/\delta)$ . Hence, on  $G$ ,  $M = 1 + \max\{m : 2w(i) \geq |\hat{\mu}_1(2i) - \hat{\mu}_2(2i)|, i = 1, 2, \dots, m\} \leq 1 + (4/\Delta)^2 2 \log(n/\delta)$ . Plugging this in and setting  $\delta = 1/n$ , we get  $R_n \leq \Delta + \frac{16}{\Delta} \log(n)$ .

(c) In addition to the said inequality, of course,  $R_n \leq n\Delta$  also holds. If  $\Delta \leq \sqrt{\log(n)/n}$ , we thus have  $R_n \leq \sqrt{n/\log(n)} \leq \sqrt{n \log(n)}$ . If  $\Delta > \sqrt{\log(n)/n}$ ,  $R_n \leq \Delta + C\sqrt{n \log(n)}$ . Combining the inequalities, we have  $R_n \leq \Delta + (C \vee 1)\sqrt{n \log(n)}$ .

(d) Change the definition of  $w(m)$  from Part (b) to  $w(m) = \sqrt{2 \log(n/(m\delta))/m}$ . Then,  $\mathbb{P}(G^c) \leq \frac{\delta}{n} \sum_{m=1}^{n/2} m \leq c\delta n$  for a suitable universal constant  $c > 0$ . We will choose  $\delta = 1/(cn^2)$  so that  $\mathbb{P}(G^c) \leq 1/n$ . Hence,  $w(m) = \sqrt{c' \log(n/m)/m}$  with a suitable universal constant  $c' > 0$ . With the same reasoning as in Part (b), we find that  $M \leq 1 + m^*$  where  $m^* = \max\{m \geq 1 : m \leq c' \log(n/m)/\Delta^2\}$ . A case analysis then gives that  $m^* \leq \frac{c'' \log(e \vee (\Delta^2 n))}{\Delta^2}$  for a suitable universal constant  $c'' > 0$ . Finishing as in Part (b),  $R_n = \mathbb{P}(G^c) n\Delta/2 + \mathbb{E}[M\mathbb{I}\{G\}] \Delta/2 \leq \delta n + \mathbb{E}[M\mathbb{I}\{G\}] \Delta/2 \leq \Delta + \frac{c'' \log(e \vee (\Delta^2 n))}{\Delta}$ .

(e) See the solution to Exercise 6.2.

## 6.6

(a) Let  $N_0 = 0$  and for  $\ell > 1$ , let  $N_\ell = \min(N_{\ell-1} + n_\ell, n)$ ,  $T_\ell = \{N_{\ell-1} + 1, \dots, N_\ell\}$ . The intervals  $(T_\ell)_{\ell=1}^{\ell_{\max}}$  are non-overlapping and policy  $\pi$  is used with horizon  $n_\ell$  on interval  $T_\ell$ . Since  $\nu$  is a stochastic environment,

$$R_n(\pi^*, \nu) = \sum_{\ell=1}^{\ell_{\max}} R_{|T_\ell|}(\pi(n_\ell), \nu) \leq \sum_{\ell=1}^{\ell_{\max}} \max_{1 \leq t \leq n_\ell} R_t(\pi(n_\ell), \nu) \leq \sum_{\ell=1}^{\ell_{\max}} f_{n_\ell}(\nu), \quad (6.1)$$

where the first inequality uses that  $|T_\ell| \leq n_\ell$  (in fact for  $\ell < \ell_{\max}$ ,  $|T_\ell| = n_\ell$ ), and the second inequality uses (6.10).

(b) We have  $\sum_{i=1}^{\ell} n_i = \sum_{i=0}^{\ell-1} 2^i = 2^\ell - 1$ , hence  $\ell_{\max} = \lceil \log_2(n+1) \rceil$  and  $2^{\ell_{\max}} \leq 2(n+1)$ . By Eq. (6.1), the assumption on  $f_n$  and the choice of  $(n_\ell)_\ell$ ,

$$\begin{aligned} R_n(\pi^*, \nu) &\leq \sum_{\ell=1}^{\ell_{\max}} \sqrt{2^{\ell-1}} \leq \frac{1}{\sqrt{2}-1} \sqrt{2^{\ell_{\max}}} \leq \frac{1}{\sqrt{2}-1} \sqrt{2n} \sqrt{1 + \frac{1}{n}} \\ &= 2(1 + \sqrt{2})\sqrt{n}. \end{aligned}$$

(c) By Eq. (6.1) we have

$$\begin{aligned} R_n(\pi^*, \nu) &\leq g(\nu) \sum_{\ell=1}^{\ell_{\max}} \log(2^{\ell-1}) = \log(2)g(\nu) \sum_{\ell=0}^{\ell_{\max}-1} \ell \\ &= \log(2)g(\nu) \frac{(\ell_{\max}-1)\ell_{\max}}{2} \leq Cg(\nu) \log^2(n+1) \end{aligned}$$

with some universal constant  $C > 0$ . Hence, the regret significantly worsens in this case. A better choice  $n_\ell = 2^{2^{\ell-1}}$ . With this,

$$\begin{aligned} R_n(\pi^*, \nu) &\leq g(\nu) \sum_{\ell=1}^{\ell_{\max}} \log(2^{2^{\ell-1}}) = \log(2)g(\nu) \sum_{\ell=0}^{\ell_{\max}-1} 2^\ell \leq \log(2)g(\nu) 2^{\ell_{\max}} \\ &\leq Cg(\nu) \log(n), \end{aligned}$$

with another universal constant  $C > 0$ .

(d) The power/advantage of the doubling trick is its generality. It shows, that, as long as we do not mind losing a constant factor, under mild conditions, adapting to an unknown horizon is not challenging. The first disadvantage is that  $\pi^*$  *does* lose a constant factor when perhaps there is no need to lose a constant factor. The second disadvantage is that oftentimes one can design algorithms such that the immediate expected regret at time  $t$  decreases as the time  $t$  increases. This is a highly desirable property and can often be met, as it was explained in the chapter. Yet, by applying the doubling trick this monotone decrease of the immediate expected regret *will* be lost, which will raise questions in the user.

## 6.8

(a) Using the definition of the algorithm and concentration for subgaussian random variables:

$$\begin{aligned} \mathbb{P}(1 \notin A_{\ell+1}, 1 \in A_\ell) &\leq \mathbb{P}\left(1 \in A_\ell, \text{exists } i \in A_\ell \setminus \{1\} : \hat{\mu}_{i,\ell} \geq \hat{\mu}_{1,\ell} + 2^{-\ell}\right) \\ &= \mathbb{P}\left(1 \in A_\ell, \text{exists } i \in A_\ell \setminus \{1\} : \hat{\mu}_{i,\ell} - \hat{\mu}_{1,\ell} \geq 2^{-\ell}\right) \\ &\leq k \exp\left(-\frac{m_\ell 2^{-2\ell}}{4}\right), \end{aligned}$$

where in the last final inequality we used (c) of Lemma 5.4 and Theorem 5.3.

(b) Again, concentration and the algorithm definition show that:

$$\begin{aligned} \mathbb{P}(i \in A_{\ell+1}, 1 \in A_\ell, i \in A_\ell) &\leq \mathbb{P}\left(1 \in A_\ell, i \in A_\ell, \hat{\mu}_{i,\ell} + 2^{-\ell} \geq \hat{\mu}_{1,\ell}\right) \\ &= \mathbb{P}\left(1 \in A_\ell, i \in A_\ell, (\hat{\mu}_{i,\ell} - \mu_i) - (\hat{\mu}_{1,\ell} - \mu_1) \geq \Delta_i - 2^{-\ell}\right) \\ &\leq \exp\left(-\frac{m_\ell (\Delta_i - 2^{-\ell})^2}{4}\right). \end{aligned}$$



(c) Let  $\delta \in (0, 1)$  be some constant to be chosen later and

$$m_\ell = 2^{4+2\ell} \log(\ell/\delta).$$

Then by Part (a),

$$\begin{aligned} \mathbb{P}(\text{exists } \ell : 1 \notin A_\ell) &\leq \sum_{\ell=1}^{\infty} \mathbb{P}(1 \notin A_{\ell+1}, 1 \in A_\ell) \\ &\leq k \sum_{\ell=1}^{\infty} \exp\left(-\frac{m_\ell 2^{2\ell}}{4}\right) \\ &\leq k\delta \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} \\ &= \frac{k\pi^2\delta}{6}. \end{aligned}$$

Furthermore, by Part (b),

$$\begin{aligned} \mathbb{P}(i \in A_{\ell_i+1}) &\leq \mathbb{P}(i \in A_{\ell_i+1}, i \in A_{\ell_i}, 1 \in A_{\ell_i}) + \mathbb{P}(1 \notin A_{\ell_i}) \\ &\leq \exp\left(-\frac{m_\ell(\Delta_i - 2^{-\ell_i})^2}{4}\right) + \frac{k\pi^2\delta}{6} \\ &\leq \exp\left(-\frac{m_\ell 2^{-2\ell_i}}{16}\right) + \frac{k\pi^2\delta}{6} \\ &\leq \delta \left(1 + \frac{k\pi^2}{6}\right). \end{aligned}$$

Choosing  $\delta = n^{-1}(1 + k\pi^2/6)^{-1}$  completes the result.

(d) For  $n < k$ , the result is trivial because all actions are tried at most once and hence  $R_n \leq \sum_{i:\Delta_i} \Delta_i$  which is below the desired bound provided that  $C > 1$ . Hence, assume that  $n \geq k$ . If there is no suboptimal action, the statement is again trivial. Otherwise, let  $i$  be a suboptimal action. Notice that  $2^{-\ell_i} \geq \Delta_i/4$  and hence  $2^{2\ell_i} \leq 16/\Delta_i^2$ . Furthermore,  $m_\ell \geq m_1 \geq 1$  for  $\ell \geq 1$ . Hence,

$$\begin{aligned} \mathbb{E}[T_i(n)] &\leq n\mathbb{P}(i \in A_{\ell_i+1}) + \sum_{\ell=1}^{\ell_i \wedge n} m_\ell \\ &\leq 1 + \sum_{\ell=1}^{\ell_i \wedge n} 2^{4+2\ell} \log\left(\frac{n}{\delta}\right) \\ &\leq 1 + C 2^{2\ell_i} \log(nk) \\ &\leq 1 + \frac{16C}{\Delta_i^2} \log(nk), \end{aligned}$$

where  $C > 1$  is a suitably large universal constant derived by naively bounding the logarithmic term and the geometric series. The result follows from upper bounding  $\log(nk) \leq 2\log(n)$  which

follows from  $k \leq n$  and the standard regret decomposition (Lemma 4.5).

(e) Briefly, the idea is to choose

$$m_\ell = C 2^{2\ell} \log \max \left( e, kn 2^{-2\ell} \right),$$

where  $C$  is a suitably large universal constant chosen so that

$$\mathbb{P}(1 \notin A_\ell) \leq \frac{2^{2\ell}}{n} \quad \text{and} \quad \mathbb{P}(i \in A_{\ell_i+1}) \leq \frac{2^{2\ell}}{n}.$$

From this it follows that

$$\mathbb{E}[T_i(n)] \leq \frac{16}{\Delta_i^2} + C \sum_{\ell=1}^{\ell_i} 2^{2\ell} \log \max \left( e, kn 2^{-2\ell} \right).$$

Bounding the sum by an integral and some algebraic gymnastics eventually leads to the desired result. Note, you have to justify the  $k$  in the logarithm is a lower order term. Argue by splitting the suboptimal arms into those with  $\Delta_i \leq \sqrt{k/n}$  and the rest.

(f) Using the analysis in Part (e) and letting  $\Delta = \sqrt{k/n}$ . Then

$$\begin{aligned} R_n &= \sum_{i=1}^k \Delta_i \mathbb{E}[T_i(n)] \\ &\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \frac{C'}{\Delta_i} \log \max \left( e, nk\Delta_i^2 \right) \\ &\leq C'' \sqrt{nk \log(k)}, \end{aligned}$$

where the constant  $C'$  is derived from  $C$  in Part (e) and the last inequality follows by considering the monotonicity properties of  $x \mapsto 1/x \log \max(e, nx^2)$ .

## Chapter 7 The Upper Confidence Bound Algorithm

### 7.1

(a) We have

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

(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

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

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

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

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

Let  $\varepsilon > 0$ . By Part (g),  $\mathbb{P}(\Delta \geq \varepsilon) \leq 5/(4n\varepsilon^2)$ , where we used the fact the  $2c/\sqrt{\pi} \leq 5/4$  when  $c = 11/10$ . 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  $\kappa_i$  and  $\Delta$ ,

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

The expectation of  $\kappa_i$  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  $\varepsilon$  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 minimiser 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  $p \in [n]$ ,

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

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

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

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

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

Thus,

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

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

**10.4** When the exponential family is in canonical form the mean of  $P_\theta$  is  $\mu(\theta) = \mathbb{E}_\theta[S] = A'(\theta)$ . 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  $\mu_{\inf} = \inf_{\theta \in \Theta} \mu(\theta)$  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  $\theta$ . Precisely, let  $X_1, \dots, X_n$  be independent and identically distributed from  $P_\theta$  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  $\theta$ ,

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

There is an irritating edge case that  $\hat{\mu}_n$  does not lie in the range of  $\mu : \Theta \rightarrow \mathbb{R}$ . 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  $X_1, \dots, X_n$  sampled from  $\mathbb{P}_\theta$  and  $\hat{s}_t = \frac{1}{t} \sum_{u=1}^t S(X_u)$  and  $\hat{\theta}_t = \hat{\theta}(\hat{s}_t)$ . Let  $\tilde{\theta} \in \Theta$  be such that  $\underline{d}(\tilde{\theta}, \theta) = \varepsilon > 0$ . Then

$$\mathbb{P}(\underline{d}(\hat{\theta}_t, \theta) \geq \varepsilon) \leq \mathbb{P}(\hat{s}_t \geq \bar{s}(\tilde{\theta})) \leq \exp(-td(\tilde{\theta}, \theta)) = \exp(-t\varepsilon), \quad (10.1)$$

where the second inequality follows from Part (e) of Exercise 34.5. Similarly

$$\mathbb{P}(\bar{d}(\hat{\theta}_t, \theta) \geq \varepsilon) \leq \exp(-t\varepsilon). \quad (10.2)$$

Define random variable  $\tau$  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  $\tau$  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  $\tau$  is bounded by

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

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  $\kappa$  is easily bounded using Eq. (10.2),

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

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

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

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

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

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

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

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

Let  $\mu(\theta) = \int_{\mathbb{R}} x dP_{\theta}(x)$  and  $\bar{s}(\theta) = \mathbb{E}_{\theta}[S] = A'(\theta)$  and  $\mathcal{S} = \{\bar{s}(\theta) : \theta \in \Theta\}$ . Define  $\hat{\theta} : \mathbb{R} \rightarrow \text{cl}(\Theta)$  by

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



The algorithm is a generalisation of KL-UCB. Let

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

which is the empirical estimator of the sufficient statistic. Like UCB, the algorithm plays  $A_t = t$  for  $t \in [k]$  and subsequently  $A_t = \operatorname{argmax}_i U_i(t)$ , 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  $X_1, \dots, X_n$  sampled from  $\mathbb{P}_\theta$  and  $\hat{s}_t = \frac{1}{t} \sum_{u=1}^t S(X_u)$  and  $\hat{\theta}_t = \hat{\theta}(\hat{s}_t)$ . Let  $\tilde{\theta} \in \Theta$  be such that  $\bar{d}(\tilde{\theta}, \theta) = \varepsilon > 0$ . Then

$$\mathbb{P}(\bar{d}(\hat{\theta}_s, \theta) \geq \varepsilon) \leq \mathbb{P}(\hat{s}_t \geq \bar{s}(\tilde{\theta})) \leq \exp(-td(\tilde{\theta}, \theta)) = \exp(-t\varepsilon). \quad (10.5)$$

Using an identical argument,

$$\mathbb{P}(\underline{d}(\hat{\theta}_t, \theta) \geq \varepsilon) \leq \exp(-t\varepsilon). \quad (10.6)$$

Define random variable  $\tau$  by

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

Then the expectation of  $\tau$  is bounded by

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

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  $\kappa$  is easily bounded Eq. (10.5) and Eq. (10.6):

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

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

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

which means that  $U_1(t) \geq \mu^*$  for all  $t \geq \tau$ . For suboptimal arms  $i$  let  $\kappa_i$  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

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

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

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

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

## Chapter 11 The Exp3 Algorithm

**11.2** Let  $\pi$  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** Let  $\hat{X}$  as stated in the problem and let  $f(x) = \sum_i P_i \hat{X}(i, x_i)$ . Let  $x \in \mathbb{R}^k$  be arbitrary and  $x' \in \mathbb{R}^k$  be such that  $x'_i = x_i$  except possibly for component  $j > 1$ . Note that  $f(x) = x_1 = x'_1 = f(x')$  and thus  $0 = f(x) - f(x') = P_j(\hat{X}(j, x_j) - \hat{X}(j, x'_j))$ . Dividing by  $P_j > 0$  implies that  $\hat{X}(j, x_j) = \hat{X}(j, x'_j)$ . Since  $x_j$  and  $x'_j$  were arbitrary,  $\hat{X}(j, \cdot) \equiv \text{const.}$  Call the value that  $\hat{X}(j, \cdot)$  is equal to  $a_j$ . Now, let  $x, x' \in \mathbb{R}^k$  be such that they agree on all components except possibly on component one and let  $x'_1 = 0$ . Further, let  $a_1 = \hat{X}(1, 0)$ . Then,  $x_1 - 0 = f(x) - f(x') = P_1(\hat{X}(1, x_1) - a_1)$ . Reordering gives that for any  $x_1 \in \mathbb{R}$ ,  $\hat{X}(1, x_1) = a_1 + x_1/P_1$ . Finally, let  $x$  be such that  $x_1 = 0$ . Then,  $0 = f(x) = \sum_i P_i a_i$ , finishing the proof.

**11.6** 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

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

(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

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

which combined with the previous part means that

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

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

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

(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  $\mathbb{E}[\hat{R}_n] < 0$ .

- (f) Since for  $n = 10^4$ , the probability of seeing a large regret is about  $1/65$  by the answer to the previous part, Exp3 was run  $m = 500$  times, which gives us a good margin to encounter large regrets. The results are shown in Fig. 11.4. As can be seen, as predicted by the theory a significant fraction of the cases, the regret is above  $n/4 = 2500$ . As seen from the figure, the mean regret is negative.

**11.7** First, note that if  $G = -\log(-\log(U))$  with  $U$  uniform on  $[0, 1]$  then

$$\mathbb{P}(G \leq g) = e^{-\exp(-g)}.$$

Now, the result follows from a long sequence of equalities:

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

## Chapter 12 The Exp3-IX Algorithm

### 12.1

- (a) We have  $\mu_t = \mathbb{E}_{t-1}[\hat{Y}_{ti}] = \frac{P_{ti}y_{ti}}{P_{ti}+\gamma}$ . Further,  $\mathbb{V}_{t-1}[\hat{Y}_{ti}] (= \mathbb{E}_{t-1}[(\hat{Y}_{ti} - \mu_t)^2]) = \frac{P_{ti}(1-P_{ti})y_{ti}^2}{(P_{ti}+\gamma)^2} \leq \frac{(P_{ti}+\gamma)y_{ti}}{(P_{ti}+\gamma)^2} = \frac{y_{ti}}{P_{ti}+\gamma}$ . For any  $\eta > 0$  such that  $\eta(\hat{Y}_{ti} - \mu_t) = \eta \frac{(A_{ti}-P_{ti})y_{ti}}{P_{ti}+\gamma} \leq 1$  almost surely for all  $t \in [n]$ ,

$$\hat{L}_{ni} - \sum_t \frac{P_{ti}y_{ti}}{P_{ti}+\gamma} \leq \eta \sum_t \frac{y_{ti}}{P_{ti}+\gamma} + \frac{1}{\eta} \log(1/\delta).$$

Choosing  $\eta = \gamma$ , the constraints  $\eta(\hat{Y}_{ti} - \mu_t) \leq 1$  are satisfied for  $t \in [n]$ . Plugging in this value and reordering gives the desired inequality.

- (b) We have  $\mu_t = \mathbb{E}_{t-1}[\sum_i \hat{Y}_{ti}] = \sum_i \frac{P_{ti}y_{ti}}{P_{ti}+\gamma}$ . Further,  $\mathbb{V}_{t-1}[\sum_i \hat{Y}_{ti}] \leq \mathbb{E}_{t-1}[(\sum_i \hat{Y}_{ti})^2] = \sum_i \mathbb{E}_{t-1}[\hat{Y}_{ti}^2] = \sum_i \frac{P_{ti}y_{ti}^2}{(P_{ti}+\gamma)^2} \leq \sum_i \frac{y_{ti}}{P_{ti}+\gamma}$ . To satisfy the constraint on  $\eta$  we calculate  $\eta(\sum_i \hat{Y}_{ti} - \mu_t) \leq \eta \sum_i \hat{Y}_{ti} =$

$\eta \sum_i \frac{A_{ti} y_{ti}}{P_{ti} + \gamma} \leq \frac{\eta}{\gamma} \sum_i A_{ti} = \frac{\eta}{\gamma}$ . Hence, any  $\eta \leq \gamma$  is suitable. Choosing  $\eta = \gamma$ , we get

$$\sum_i \hat{L}_{ni} - \sum_t \sum_i \frac{P_{ti} y_{ti}}{P_{ti} + \gamma} \leq \sum_t \sum_i \frac{\gamma y_{ti}}{P_{ti} + \gamma} + \frac{1}{\gamma} \log(1/\delta).$$

Reordering as before gives the desired result.

**12.4** We proceed in five steps.

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

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

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

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

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

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

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  $1 - \delta$

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

STEP 4: BOUNDING (C) For (C) we have

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

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  $1 - \delta$ .

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

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

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

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  $\omega$ ,

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

where the supremum is taken over all finite partitions of  $\mathbb{R}$  with rational-valued end-points. By the definition of a probability kernel it follows that the quantity inside the supremum on the right-hand side is  $\mathcal{F}$ -measurable as a function of  $\omega$  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

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

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

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

which is well defined for all  $x_1, \dots, x_{t-1} \in \mathbb{R}^{t-1}$  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

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

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

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

and assume that  $P(U_t = 1) = 1$  for all  $t$ . Then  $U = \cap_{t=1}^n U_t$  satisfies  $P(U) = 1$ . On  $U_t$  let  $F(x_t | x_1, \dots, x_{t-1}) = dP_t(\cdot | x_1, \dots, x_{t-1})/dQ_t(\cdot | x_1, \dots, x_{t-1})(x_t)$  and otherwise let  $F(x_t | x_1, \dots, x_{t-1}) = 0$ . 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

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

Hence  $\prod_{t=1}^n F(x_t | x_1, \dots, x_{t-1})$  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  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$  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  $d(\hat{\theta}, P_1) \geq \Delta - d(\hat{\theta}, P_0) \geq \Delta/2$ .

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

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  $D(\mathcal{R}(\mu), \mathcal{R}(\mu')) = \infty$ , 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  $\sum_{i=2}^{\infty} i2^{1-i} = 3$ . Hence

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

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

**16.7** Fix any policy  $\pi$ . If  $\pi$  is not a consistent policy for  $\mathcal{E}_{[0,b]}^k$  then  $\pi$  cannot have logarithmic regret, which contradicts both Eq. (16.7) and Eq. (16.8). Hence, we may assume that  $\pi$  is consistent for this environment class. We can thus apply Theorem 16.2. Let  $\mathcal{M}$  be the set of probability distributions supported on  $[0, b]$  and let  $\mathcal{M}'$  be the set of scaled Bernoulli distributions supported on  $[0, b]$ :  $P \in \mathcal{M}'$  if  $P = (1-p)\delta_0 + p\delta_b$  for  $p \in [0, 1]$  where  $\delta_x$  is the Dirac distribution supported on  $\{x\}$ . Then, thanks to Theorem 16.2, for any  $\nu \in \mathcal{E}_{[0,b]}^k$ ,

$$\liminf_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{\log(n)} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{d_{\inf}(P_i, \mu^*, \mathcal{M})} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{d_{\inf}(P_i, \mu^*, \mathcal{M}')},$$

where the latter inequality holds thanks to  $\mathcal{M}' \subset \mathcal{M}$ . Note that any  $P \in \mathcal{M}'$  is uniquely determined by its Bernoulli parameter  $p$ . Choose some  $\nu \in (\mathcal{M}')^k$ ,  $\nu = (P_i)$  and let  $p_i$  the Bernoulli parameter underlying  $P_i$ . Introduce  $p^* = \max_i p_i$ . Then,  $d_{\inf}(P_i, \mu^*, \mathcal{M}') = d(p_i, p^*)$  where  $d(p, q) (= D(\mathcal{B}(p), \mathcal{B}(q))) = p \log(p/q) + (1-p) \log((1-p)/(1-q))$  (cf. Definition 10.1). Furthermore,  $\Delta_i = b(p^* - p_i)$ . Hence, we conclude that

$$\limsup_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{\log(n)} \geq \liminf_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{\log(n)} \geq \sum_{i: p_i < p^*} \frac{b(p^* - p_i)}{d(p_i, p^*)}.$$

We will consider environments  $\nu = \nu_\delta$  given by the Bernoulli parameters  $((1+\delta)/2, \dots, (1+\delta)/2, (1 -$

$\delta)/2)$  for some  $\delta \in [0, 1]$ . Thus,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{\log(n)} &\geq \sum_{i: p_i < p^*} \frac{b(p^* - p_i)}{d(p_i, p^*)} = \frac{b\delta}{d((1-\delta)/2, (1+\delta)/2)} \\ &= \frac{b}{\log((1+\delta)/(1-\delta))}. \end{aligned}$$

Denote the right-hand side by  $f(\delta)$ . Noticing that  $\lim_{\delta \rightarrow 0} f(\delta) = \infty$  we immediately see that Eq. (16.7) cannot hold: As the action gap gets small, if we maintain the variance at constant (as in this example) then the regret blows up with the inverse action gap. To show that Eq. (16.8) cannot hold either, consider the case when  $\delta \rightarrow 1$ . The right-hand side of Eq. (16.8) is  $\sigma_k^2/\Delta_k = b(1-\delta^2)/(4\delta)$ . Now,  $f(\delta) \frac{\Delta_k}{\sigma_k^2} = \frac{4\delta}{(1-\delta^2)\log((1+\delta)/(1-\delta))} \rightarrow \infty$  as  $\delta \rightarrow 1$ . (Note that in this case the variance decreases to zero, while the gap is maintained. Since the algorithm does not know that the variance is zero, it has to pay a logarithmic cost).

## Chapter 17 High-Probability Lower Bounds

### 17.1

*Proof of Claim 17.5.* We have

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

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

*Proof of Claim 17.6.* Abbreviate  $\mathbb{P}_i$  to be the law of  $A_1, X_1, \dots, A_n, X_n$  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 minimises  $\mathbb{E}_1[T_j(n)]$ , which satisfies

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

Therefore by Theorem 14.2 and Lemma 15.1,

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

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

*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

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

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

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

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

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

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

## Chapter 18 Contextual Bandits

### 18.1

(a) By Jensen's inequality,

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

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  $X_i \leq \max_j X_j$  for any family of random variables  $(X_i)_i$ . Hence  $\mathbb{E}[X_i] \leq \mathbb{E}[\max_j X_j]$ .
- (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  $E_n^* \leq nk$  is trivial since  $\max_m E_{mi}^{(t)} \leq 1$ . To prove  $E_n^* \leq nM$ , let  $m_{ti}^* = \operatorname{argmax}_m E_{mi}^{(t)}$ . Then

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

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  $t \leq m$ . Since  $A_t$  is chosen uniformly at random,

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

The variance satisfies

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

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

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

where  $g(x) = (\exp(x) - 1 - x)/x^2$ , which for  $x \in (0, 1]$  satisfies  $g(x) \leq 1/2 + x/4$ . Suppose that

$m \geq 2k \log(2|\Phi|)$ . Then by following the argument in Exercise 5.18,

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

where the second inequality follows by choosing

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

Therefore the regret is bounded by

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

By tuning  $m$  it follows that

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

**18.9** Let  $C_1, \dots, C_n \in \mathcal{C}$  be the i.i.d. sequence of contexts and for  $k \in [n]$  let  $C_{1:k} = (C_1, \dots, C_k)$ . The algorithm is as follows: Following the hint, for the first  $m$  rounds the algorithm selects arms in an arbitrary fashion. The regret from this period is bounded by  $m$ . The algorithm then picks  $M = |\Phi_{C_{1:m}}|$  functions  $\Phi' = \{\phi_1, \dots, \phi_M\}$  from  $\Phi$  so that  $\Phi'|_{C_{1:m}} = \Phi_{C_{1:m}}$  and for the remaining  $n - m$  rounds uses Exp4 with the set  $\Phi'$ .

The regret of Exp4 for competing against  $\Phi'$  is  $\sqrt{4n \log(M)} \leq \sqrt{4n \log((em/d)^d)} = \sqrt{4nd \log(em/d)}$ , where the inequality follows from Sauer's lemma. It remains to show that the best expert in  $\Phi'$  achieves almost as much reward as the best expert in  $\Phi$ . For this, it suffices to show that for any  $\phi_* \in \Phi$ , the expert  $\phi'_* \in \Phi'$  that agrees with  $\phi_*$  on  $C_{1:m}$  agrees with  $\phi_*$  on most of the rest of the rounds. Let  $d^*$  be a positive integer to be chosen later. We will show that with high probability  $\phi'_*$  and  $\phi_*$  agree except for at most  $d^*$  rounds.

We need a few definitions: For  $a, b$  finite sequences of equal length  $k$  let  $d(a, b) = \sum_{i=1}^k \mathbb{I}\{a_i \neq b_i\}$  be their Hamming distance. For a sequence  $c \in \mathcal{C}^k$ , let  $\phi(c) = (\phi(c_1), \dots, \phi(c_k)) \in \{1, 2\}^k$ . For  $\phi, \phi' \in \Phi$ , let  $d_C^{(k)}(\phi, \phi') = d(\phi(C_{1:k}), \phi'(C_{1:k}))$ . For a permutation  $\pi$  on  $[n]$ , let  $C^\pi = (C_{\pi(1)}, C_{\pi(2)}, \dots, C_{\pi(n)})$ .

Let  $\pi$  be a random permutation on  $[n]$ , chosen uniformly from the set of all permutations on  $[n]$ , independently of  $C$ . We have

$$\begin{aligned} p &= \mathbb{P}\left(d_C^{(n)}(\phi_*, \phi'_*) \geq d^*\right) \leq \mathbb{P}\left(\exists \phi, \phi' \in \Phi : d_C^{(n)}(\phi, \phi') \geq d^*, d_C^{(m)}(\phi, \phi') = 0\right) \\ &= \mathbb{P}\left(\exists \phi, \phi' \in \Phi : d_{C^\pi}^{(n)}(\phi, \phi') \geq d^*, d_{C^\pi}^{(m)}(\phi, \phi') = 0\right) \\ &= \mathbb{E}\left[\mathbb{P}\left(\exists \phi, \phi' \in \Phi : d_{C^\pi}^{(n)}(\phi, \phi') \geq d^*, d_{C^\pi}^{(m)}(\phi, \phi') = 0 \mid C\right)\right], \end{aligned}$$

where the first equality is the definition of  $p$ , the second holds by the exchangeability of  $(C_1, \dots, C_n)$ . Now, by a union bound and because  $\pi$  and  $C$  are independent,

$$\begin{aligned} & \mathbb{P} \left( \exists \phi, \phi' \in \Phi : d_{C^\pi}^{(n)}(\phi, \phi') \geq d^*, d_{C^\pi}^{(m)}(\phi, \phi') = 0 | C \right) \\ & \leq \sum_{a, b \in \Phi_{C_{1:n}}} \mathbb{P} (d(a^\pi, b^\pi) \geq d^*, d(a_{1:m}^\pi, b_{1:m}^\pi) = 0) . \end{aligned}$$

By Sauer's lemma,  $|\Phi_{C_{1:n}}| \leq (en/d)^d$ . For any fixed  $a, b \in \{1, 2\}^n$ ,  $p(a, b) = \mathbb{P}(d(a^\pi, b^\pi) \geq d^*, d(a_{1:m}^\pi, b_{1:m}^\pi) = 0)$  is the probability that all bits in a random subsequence of length  $m$  of a bit sequence of length  $n$  with at least  $d^*$  one bits has only zero bits. As the probability of a randomly chosen bit to be equal to zero is  $1 - d^*/n$ ,  $p(a, b) \leq \left(1 - \frac{d^*}{n}\right)^m \leq \exp(-d^*m/n)$ . Choosing  $d^* = \lceil \frac{n}{m} \log \left( \frac{(en/d)^{2d}}{\delta} \right) \rceil$  we find that  $p \leq \delta$ .

Choosing  $\delta = 1/n$ , we get that

$$\begin{aligned} R_n & \leq 1 + m + d^* + \sqrt{4nd \log \left( \frac{em}{d} \right)} \\ & \leq 2 + m + \frac{n}{m} \left( \log(n) + (2d) \log \left( \frac{en}{d} \right) \right) + \sqrt{4nd \log \left( \frac{em}{d} \right)} . \end{aligned}$$

Letting  $m = \sqrt{dn}$  gives the desired bound.

## Chapter 19 Stochastic Linear Bandits

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

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

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

$$x \log(1 + y(3x \log(1 + xy))) \leq x \log(1 + 3x^2 y^2) \leq x \log(1 + xy)^3 = 3x \log(1 + xy).$$

Since  $z - x \log(1 + yz)$  is decreasing for  $z \geq 3x \log(1 + xy)$  it follows that

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

#### 19.4

(a) Let  $A = B + uu^\top$ . Then, for  $x \neq 0$ ,

$$\frac{\|x\|_A^2}{\|x\|_B^2} = 1 + \frac{(x^\top u)^2}{\|x\|_B^2} \leq 1 + \frac{\|x\|_B^2 \|u\|_{B^{-1}}^2}{\|x\|_B^2} = 1 + \|u\|_{B^{-1}}^2 = \frac{\det A}{\det B}. \quad (19.1)$$

where the inequality follows by Cauchy-Schwartz, and the equality follows because  $\det(A) = \det(B) \det(I + B^{-1/2}uu^\top B^{-1/2}) = \det(B)(1 + \|u\|_{B^{-1}}^2)$  per the proof of Lemma 19.4. In the general case,  $C = A - B \succeq 0$  can be written using its eigendecomposition as  $C = \sum_{i=1}^k u_i u_i^\top$  with  $0 \leq k \leq d$ . Letting  $A_j = B + \sum_{i \leq j} u_i u_i^\top$ ,  $0 \leq j \leq k$  note that  $A = A_k$  and  $B = A_0$  and thus by (19.1),

$$\begin{aligned} \frac{\|x\|_A^2}{\|x\|_B^2} &= \frac{\|x\|_{A_k}^2}{\|x\|_{A_0}^2} = \frac{\|x\|_{A_k}^2}{\|x\|_{A_{k-1}}^2} \frac{\|x\|_{A_{k-1}}^2}{\|x\|_{A_{k-2}}^2} \cdots \frac{\|x\|_{A_1}^2}{\|x\|_{A_0}^2} \\ &\leq \frac{\det A_k}{\det A_{k-1}} \frac{\det A_{k-1}}{\det A_{k-2}} \cdots \frac{\det A_1}{\det A_0} = \frac{\det A}{\det B}. \end{aligned}$$

(b) We need some notation. As before, for  $t \geq 0$ , we let  $V_t(\lambda) = V_0 + \sum_{s \leq t} A_s A_s^\top$ , where  $A_s$  is the action taken in round  $s$ . Let  $\tau_0 = 0$ . For  $t \geq 1$ , we let  $\tau_t \in [t]$  denote the round index when the phase that contains round  $t$  starts. That is,

$$\tau_t = \begin{cases} \tau_{t-1}, & \text{if } \det V_{t-1}(\lambda) \leq (1 + \varepsilon) \det V_{\tau_{t-1}-1}(\lambda); \\ t, & \text{otherwise.} \end{cases}$$

Further,

$$A_t = \begin{cases} A_{t-1}, & \text{if } \tau_t = \tau_{t-1}; \\ \operatorname{argmax}_{a \in \mathcal{A}} \text{UCB}_t(a), & \text{otherwise.} \end{cases}$$

Here,  $\text{UCB}_t(a)$  is the upper confidence bound based on all the data available up the beginning of round  $t$ .

Let the event when  $\theta_* \in \cap_{t \in [n]} \mathcal{C}_t$  hold. Define  $\tilde{\theta}_t \in \mathcal{C}_t$  so that if  $\tilde{A}_t = \operatorname{argmax}_{a \in \mathcal{A}} \text{UCB}_t(a)$  then  $\langle \tilde{\theta}_t, \tilde{A}_t \rangle = \max_{a \in \mathcal{A}} \text{UCB}_t(a)$ . Letting  $a^* = \operatorname{argmax}_{a \in \mathcal{A}} \langle \theta_*, a \rangle$  and noting that  $\max_a \text{UCB}_{\tau_t}(a) =$



$\text{UCB}_{\tau_t}(A_{\tau_t})$  and that  $A_t = A_{\tau_t}$ ,

$$\langle \theta_*, a^* \rangle \leq \text{UCB}_{\tau_t}(a_*) \leq \text{UCB}_{\tau_t}(A_{\tau_t}) = \langle \tilde{\theta}_{\tau_t}, A_{\tau_t} \rangle = \langle \tilde{\theta}_{\tau_t}, A_t \rangle.$$

Hence,

$$r_t = \langle \theta_*, a^* - A_t \rangle \leq \langle \tilde{\theta}_{\tau_t} - \theta_*, A_t \rangle \leq \|\tilde{\theta}_{\tau_t} - \theta_*\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}}.$$

Now, by Part (a), thanks to  $V_{\tau_t-1} \preceq V_{t-1}$  and that  $\theta_*, \tilde{\theta}_{\tau_t} \in \mathcal{C}_{\tau_t}$  and that  $\mathcal{C}_{\tau_t}$  is contained in an ellipsoid of radius  $\beta_{\tau_t} \leq \beta_t$  and “shape” determined by  $V_{\tau_t-1}$ ,

$$\|\tilde{\theta}_{\tau_t} - \theta_*\|_{V_{t-1}} \leq \|\tilde{\theta}_{\tau_t} - \theta_*\|_{V_{\tau_t-1}} \left( \frac{\det V_{t-1}}{\det V_{\tau_t-1}} \right)^{1/2} \leq 2\sqrt{(1+\varepsilon)\beta_t}.$$

Combined with the previous inequality, we see that

$$r_t \leq 2\sqrt{(1+\varepsilon)\beta_t} \|A_t\|_{V_{t-1}^{-1}},$$

which is the same as Eq. (19.10) except that  $\beta_t$  is replaced with  $(1+\varepsilon)\beta_t$ . This implies in turn that  $\hat{R}_n = \sum_{t=1}^n \leq \hat{R}_n((1+\varepsilon)\beta_n)$ .

**19.5** We only present in detail the solution for the first part.

- (a) Partition  $\mathcal{C}$  into  $m$  equal length subintervals, call these  $\mathcal{C}_1, \dots, \mathcal{C}_m$ . 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  $i \in [m]$ ,

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

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 discretisation:

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

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  $|C_t - c| \leq 1/m$ . It remains to bound the error due to learning. Note that  $\mathbb{E}[X_t \mid [C_t] = i] = \tilde{r}(i, A_t)$ . 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 28.13. For  $i \in [m]$ , let  $T_i = \{t \in [n] \mid [C_t] = i\}$ ,  $N_i = |T_i|$  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),  $\mathbb{E}[R_{ni}] \leq C\mathbb{E}[\sqrt{k \log(k) N_i}]$  and thus

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

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

$$R_n \leq \frac{Ln}{m} + C\sqrt{k \log(k) mn}.$$

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  $\varepsilon$  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  $\varepsilon$  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.

## 19.6

(a) We need to show that  $L_t(\theta_*) \leq \beta_t^{1/2}$  for all  $t$  with high probability. By definition,

$$\begin{aligned} L_t(\theta_*) &= \left\| g_t(\theta_*) - \sum_{s=1}^t X_s A_s \right\|_{V_t^{-1}} \\ &= \left\| \lambda \theta_* + \sum_{s=1}^t (\mu(\langle \theta_*, A_s \rangle) A_s - (\mu(\langle \theta_*, A_s \rangle) A_s + \eta_s) A_s) \right\|_{V_t^{-1}} \\ &= \left\| \lambda \theta_* + \sum_{s=1}^t A_s \eta_s \right\|_{V_t^{-1}} \\ &\leq \left\| \sum_{s=1}^t A_s \eta_s \right\|_{V_t^{-1}} + \sqrt{\lambda} \|\theta_*\|_2. \end{aligned}$$

The result now follows by Theorem 20.4.

(b) Let  $g'_t$  denote the derivative of  $g_t$  (which exists by assumption). By the mean value theorem, there exists a  $\xi$  on the segment connecting  $\theta$  and  $\theta'$  such that

$$\begin{aligned} g_t(\theta) - g_t(\theta') &= g'_t(\xi)(\theta - \theta') \\ &= \left( \lambda I + \sum_{s=1}^t \mu'(\langle \xi, A_s \rangle) A_s A_s^\top \right) (\theta - \theta') \\ &= M_t(\theta - \theta'), \end{aligned}$$

where  $M_t = \lambda I + \sum_{s=1}^t \mu'(\langle \xi, A_s \rangle) A_s A_s^\top \succeq c_1 V_t$ . Hence,

$$\|g_t(\theta) - g_t(\theta')\|_{V_t^{-1}} = \|\theta - \theta'\|_{M_t V_t^{-1} M_t} \geq c_1 \|\theta - \theta'\|_{V_t}.$$

(c) Let  $\tilde{\theta}_t$  be such that  $\mu(\langle \tilde{\theta}_t, A_t \rangle) = \max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \mu(\langle \theta, a \rangle)$ . Using the fact that  $\theta_* \in \mathcal{C}_t$ ,

$$\begin{aligned} r_t &= \mu(\langle \theta_*, A_t^* \rangle) - \mu(\langle \theta_*, A_t \rangle) \leq \mu(\langle \tilde{\theta}_t, A_t \rangle) - \mu(\langle \theta_*, A_t \rangle) \leq c_2 \langle \tilde{\theta}_t - \theta_*, A_t \rangle \\ &\leq c_2 \|\tilde{\theta}_t - \theta_*\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}} \\ &\leq \frac{c_2}{c_1} \|g_{t-1}(\tilde{\theta}_t) - g_{t-1}(\theta_*)\|_{V_{t-1}^{-1}} \|A_t\|_{V_{t-1}^{-1}} \\ &\leq \frac{c_2}{c_1} (L_{t-1}(\tilde{\theta}_t) + L_{t-1}(\theta_*)) \|A_t\|_{V_{t-1}^{-1}} \\ &\leq \frac{2c_2 \beta_{t-1}^{1/2}}{c_1} \|A_t\|_{V_{t-1}^{-1}}. \end{aligned}$$

- (d) By Part (a), with probability at least  $1 - \delta$ ,  $\theta_* \in \mathcal{C}_t$  for all  $t$ . Hence, by Part (c), with probability at least  $1 - \delta$ ,

$$\begin{aligned}\hat{R}_n &= \sum_{t=1}^n r_t \leq \frac{2c_2\beta_n^{1/2}}{c_1} \sum_{t=1}^n (1 \wedge \|A_t\|_{V_{t-1}^{-1}}) \\ &\leq \frac{2c_2}{c_1} \sqrt{2nd\beta_n \log \left(1 + \frac{nL^2}{d}\right)},\end{aligned}$$

where we used Lemma 19.4 and the same argument as in the proof of Theorem 19.2.

## 19.7

- (a) Let  $\xi_1, \dots, \xi_d$  be the eigenvalues of  $V_t$  in increasing order. By the Courant–Fischer min-max theorem,  $\delta_i = \xi_i - \lambda_i \geq 0$ . Hence,

$$\frac{\det(V_t)}{\det(V_0)} = \prod_{i=1}^d \frac{\xi_i}{\lambda_i} = \prod_{i=1}^d \left(1 + \frac{\delta_i}{\lambda_i}\right).$$

A trace argument shows that  $\sum_{i=1}^d \delta_i \leq nL^2$ , which, assuming that  $d_{\text{eff}} < d$ , implies that

$$\begin{aligned}\log \left( \frac{\det(V_t)}{\det(V_0)} \right) &= \sum_{i=1}^d \log \left( 1 + \frac{\delta_i}{\lambda_i} \right) \\ &= \sum_{i=1}^{d_{\text{eff}}} \log \left( 1 + \frac{\delta_i}{\lambda_i} \right) + \sum_{i=d_{\text{eff}}+1}^d \frac{\delta_i}{\lambda_{d_{\text{eff}}+1}} \quad (\log(1+x) \leq x, x \geq 0, \lambda_i \text{ increasing}) \\ &\leq \sum_{i=1}^{d_{\text{eff}}} \log \left( 1 + \frac{nL^2}{\lambda} \right) + \frac{nL^2}{\lambda_{d_{\text{eff}}+1}} \\ &\leq 2d_{\text{eff}} \log \left( 1 + \frac{nL^2}{\lambda} \right).\end{aligned}$$

When  $d_{\text{eff}} = d$ , the desired inequality trivially holds.

- (b) By Theorem 20.4 in the next chapter, it holds with probability least  $1 - \delta$  that for all  $t$ ,

$$\left\| \sum_{s=1}^t A_s X_s \right\|_{V_t^{-1}} \leq \sqrt{2 \log \left( \frac{1}{\delta} \right) + \log \left( \frac{\det(V_t)}{\det(V_0)} \right)}.$$

On this event,

$$\begin{aligned}
\|\hat{\theta}_t - \theta\|_{V_t} &= \left\| V_t^{-1} \sum_{s=1}^t A_s A_s^\top \theta - \theta + V_t^{-1} \sum_{s=1}^t A_s \eta_s \right\|_{V_t} \\
&\leq \|V_t^{-1} V_0 \theta\|_{V_t} + \left\| \sum_{s=1}^t A_s \eta_s \right\|_{V_t} \\
&\leq \|\theta\|_{V_0} + \sqrt{2 \log \left( \frac{1}{\delta} \right) + \log \left( \frac{\det(V_t)}{\det(V_0)} \right)} \\
&\leq m + \sqrt{2 \log \left( \frac{1}{\delta} \right) + \log \left( \frac{\det(V_t)}{\det(V_0)} \right)}.
\end{aligned}$$

(c) Using again the argument as in the proof of Theorem 19.2,

$$r_t = \langle A_t^* - A_t, \theta_* \rangle \leq 2\beta_n^{1/2} \min\{1, \|A_t\|_{V_{t-1}^{-1}}\}.$$

Then, using Lemma 19.4,

$$\begin{aligned}
\sum_{t=1}^n r_t &\leq 2\beta_n^{1/2} \sqrt{n \sum_{t=1}^n \min\{1, \|A_t\|_{V_{t-1}^{-1}}^2\}} \\
&\leq 2\beta_n^{1/2} \sqrt{2n \log \left( \frac{\det V_t}{\det V_0} \right)} \\
&\leq 2\beta_n^{1/2} \sqrt{2nd_{\text{eff}} \log \left( 1 + \frac{nL^2}{\lambda} \right)}.
\end{aligned}$$

- (d) The result follows by combining the result of the last part and that of Part (a) while choosing  $\delta = 1/n$ .
- (e) The bound improves when  $d_{\text{eff}} \ll d$ . For this,  $V_0$  should have a large number of large eigenvalues while maintaining  $\sum_i \lambda_i \langle v_i, \theta_* \rangle^2 = \|\theta_*\|_{V_0}^2 \leq m^2$ : Choosing  $V_0$  is betting on the directions  $(v_i)$  in which  $\theta^*$  will have small components as in those directions we can choose  $\lambda_i$  to be large. This way, one can have a finer control than just choosing features of some fixed dimension: The added flexibility is the main advantage of non-uniform regularisation. However, this is not for free: Non-uniform regularisation may increase  $\|\theta_*\|_{V_0}$  for some  $\theta_*$  which may lead to worse regret. In particular, when  $\|\theta_*\|_{V_0} \leq m$  is not satisfied, the confidence set will still contain  $\theta_*$ , but the coverage probability  $1 - \delta$  will worsen to about  $1 - \delta'$  where  $\delta' \approx \delta \exp((\|\theta_*\|_{V_0} - m)_+^2)$ , which increases the  $\delta n$  term in the regret to  $\delta' n$ . Thus, the degradation is smooth, but can be quite harsh. Since the confidence radius is probably conservative, the degradation may be less noticeable than what one would expect based on this calculation.

**19.8** For the last part, note that it suffices to store the inverse matrix  $U_t = V_t^{-1}(\lambda)$ ,  $S_t = \sum_{s \leq t} A_s X_s$ , together with  $d_t = \log \frac{\det V_t(\lambda)}{\det V_0(\lambda)}$ . Choosing  $V_0 = \lambda I$ , we have  $U_0 = \lambda^{-1} I$ ,  $S_0 = 0$  and  $d_0 = 0$ . Then,

for  $t \geq 0$  and  $a \in \mathbb{R}^d$  we have

$$\text{UCB}_t(a) = \langle \hat{\theta}_t, a \rangle + \sqrt{\beta_t} \|a\|_{U_{t-1}}$$

where

$$\begin{aligned}\hat{\theta}_t &= U_{t-1} S_{t-1}, \\ \sqrt{\beta_t} &= m_2 \sqrt{\lambda} + \sqrt{2 \log(1/\delta) + d_t},\end{aligned}$$

and where after receiving  $(A_t, X_t)$ , the following updates need to be executed:

$$\begin{aligned}U_t &= U_{t-1} - \frac{(U_{t-1} A_t)(U_{t-1} A_t)^\top}{1 + A_t^\top U_{t-1} A_t}, \\ d_t &= d_{t-1} + \log(1 + A_t^\top U_{t-1} A_t), \\ S_t &= S_{t-1} + A_t X_t.\end{aligned}$$

Here, the update of  $d_t$  follows from Eq. (19.9). Note that the  $O(d^2)$  operations are the calculation of  $\hat{\theta}_t$  and the update of  $U_t$ .

## Chapter 20 Confidence Bounds for Least Squares Estimators

**20.1** From the definition of the design  $V_n = mI$  and the  $i$ th 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  $i$ th block starting in round  $t_i = (i-1)m + 1$ . For all  $i \in \{1, \dots, d\}$  we set  $A_{t_i} = e_i$ . For  $t \in \{t_i, \dots, t_i + m - 1\}$  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,

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

Therefore

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

### 20.3

- (a) If  $C \subset \mathcal{A}$  is an  $\varepsilon$ -covering then it is also an  $\varepsilon'$ -covering with any  $\varepsilon' \geq \varepsilon$ . Hence,  $\varepsilon \rightarrow N(\varepsilon)$  is a decreasing function of  $\varepsilon$ .
- (b) The inequality  $M(2\varepsilon) \leq N(\varepsilon)$  amounts to showing that any  $2\varepsilon$  packing has a cardinality at most the cardinality of any  $\varepsilon$  covering. Assume this does not hold, that is, there is a  $2\varepsilon$  packing  $P \subset \mathcal{A}$  and an  $\varepsilon$ -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  $2\varepsilon$ -packing.

If  $M(\varepsilon) = \infty$ , the inequality  $N(\varepsilon) \leq M(\varepsilon)$  is trivially true. Otherwise take a maximum  $\varepsilon$ -packing  $P$  of  $\mathcal{A}$ . This packing is automatically an  $\varepsilon$ -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  $\varepsilon$ -cover of  $\mathcal{A}$ . Then from the definition of cover and the additivity of volume,  $\text{vol}(\mathcal{A}) \leq \sum_{x \in C} \text{vol}(B(x, \varepsilon)) = N\varepsilon^d \text{vol}(B)$ . 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  $\varepsilon$ -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,  $M \text{vol}(\frac{\varepsilon}{2}B) \leq \text{vol}(\mathcal{A} + \frac{\varepsilon}{2}B)$ .

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  $\text{vol}(cX) = c^d \text{vol}(X)$ . 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  $\text{vol}(\mathcal{A} + \varepsilon/2B) \leq \text{vol}(rB + \varepsilon/2B) = \text{vol}((r + \varepsilon/2)B) < +\infty$ , 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.

**20.4** The result follows from Part (c) of Exercise 20.3 by taking  $\mathcal{A} = B = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ , which shows that the covering number  $N(B, \varepsilon) \leq (3/\varepsilon)^d$ , from which the result follows.

**20.5** 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, Lemma 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  $0 \not\leq 0$ . 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  $x$  (though a measurable choice can be constructed using separability of  $\mathbb{R}^d$  and continuity of  $x \mapsto M_t(x)$ ).

**20.8** 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  $\mathbb{E}[M_t] = M_0 = 1$ , 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.

## 20.9

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

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

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

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

**20.10** 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  $\mathcal{G}_1 \subset \dots \subset \mathcal{G}_n$  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  $\tau$  with  $\tau \leq n$  almost surely,  $\mathbb{P}(M_\tau \geq 1/\delta) \leq \delta$ . Next we need a bound on  $M_\tau$ . 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.

**20.11**

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

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

Then, applying the Cramér-Chernoff trick,

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

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,

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

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

**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  $\lambda_i$  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  $\text{span}(\mathcal{A}_n) = \mathbb{R}^d$  and  $\lim_{n \rightarrow \infty} d(\mathcal{A}, \mathcal{A}_n) = 0$  where  $d$  is the Hausdorff metric. Then let  $\pi_n$  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

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

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

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

Taking the limit as  $n$  tends to infinity shows that

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

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

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

Notice that  $\pi_n$  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  $(\pi_n)$  has a cluster point  $\pi^*$ , which represents a distribution on  $\mathcal{A}$  with support at most  $d(d+1)/2$ . The previous display shows that  $g(\pi^*) \leq d$ . The fact that  $g(\pi^*) \geq d$  follows from the same argument as the proof of Theorem 21.1.

**21.5** Let  $\pi$  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  $\pi^*$  is a minimiser of  $f$ ,

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

Rearranging shows that  $\|a\|_{V(\pi)^{-1}}^2 \geq d$ . 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

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

Lemma 15.1 gives that

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

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

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

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

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

## 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  $\mathcal{A} = \{e_1, \dots, e_{d-1}\}$ . 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  $\mathcal{A} = \{e_1, \dots, e_{d-1}\}$ . Then for small values of  $\Delta$  adding  $e_d$  decreases the asymptotic regret.

## Chapter 26 Foundations of Convex Analysis

**26.2** 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  $\langle a, x \rangle + b \leq f(x)$  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

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

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.

### 26.3

(a) Using the definition,

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

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

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

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

### 26.9

(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,  $\text{dom}(\nabla f^*) = \mathbb{R}^d$ .

(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  $\text{int}(\text{dom}(f))$  and

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

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

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

### 26.13

(a) Let  $g(z) = f(z) - \langle z - y, \nabla f(y) \rangle$ . Then by definition

$$z \in \text{argmin}_A g(z),$$

which exists by the assumption that  $A$  is compact. By convexity of  $A$  and the first-order optimality condition it follows that

$$\nabla_{x-z} f(z) - \langle x - z, \nabla f(y) \rangle = \nabla_{x-z} g(z) \geq 0.$$

Therefore

$$\begin{aligned} \nabla_{x-z} f(z) &\geq \langle x - z, \nabla f(y) \rangle = \langle x - y, \nabla f(y) \rangle - \langle z - y, \nabla f(y) \rangle \\ &= \nabla_{x-y} f(y) - \nabla_{z-y} f(y). \end{aligned}$$

Substituting the definition of the Bregman divergence shows that

$$D_f(x, y) \geq D_f(x, z) + D_f(z, y)$$

The proof fails when  $f$  is not differentiable at  $y$  because the map  $v \mapsto \nabla_v f(y)$  need not be linear.

(b) Consider the function  $f(x, y) = -(xy)^{1/4}$  and let  $y = (0, 0)$  and  $x = (1, 0)$  and  $A = \{(t, 1 - t) : t \in [0, 1]\}$ . Then  $D_f(x, y) = D_f(z, y) = 0$ , but  $D_f(x, z) = \infty$ .

**26.14** Parts (a) and (b) are immediate from convexity and the definitions. For Part (c), we have

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle,$$

which is linear as a function of  $x$ . Note that here we used differentiability of  $f$  at  $y$ . An example showing that differentiability at  $y$  is necessary occurs when  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  is given by

$$f(x) = \max\{1, \|x\|_1\}.$$

Then consider  $y = (2, 0)$  and  $x = (0, 2)$  and  $z = (0, -2)$ . Then  $D_f(z, y) = D_f(x, y) = 0$ , but  $D_f((x + z)/2, y) = D_f(\mathbf{0}, y) = 1 \geq (D_f(x, y) + D_f(z, y))/2$ .

**26.15** The first part follows immediately from Taylor's theorem. The second part takes a little work. To begin, abbreviate  $\|x - y\|_z = \|x - y\|_{\nabla^2 f(z)}$  and for  $t \in (0, 1)$  let

$$\delta(x, y, t) = D_f(x, y) - \frac{1}{2}\|x - y\|_{tx+(1-t)y}^2,$$

which is continuous on  $\text{int}(\text{dom}(f)) \times \text{int}(\text{dom}(f)) \times [0, 1]$ . Let  $(a, b) \subset (0, 1)$  and consider

$$A = \bigcup_{\delta \in (0, b-a) \cap \mathbb{Q}} \bigcap_{\varepsilon \in (0, 1) \cap \mathbb{Q}} \bigcup_{t \in U \cap \mathbb{Q}} \{(x, y) : \delta(x, y, t) \leq \varepsilon\}.$$

As we mentioned already, Taylor's theorem ensures there exists a  $t \in [0, 1]$  such that  $\delta(x, y, t) = 0$  for all  $(x, y) \in \text{int}(\text{dom}(f)) \times \text{int}(\text{dom}(f))$ . By the continuity of  $\delta$  it follows that  $(x, y) \in A$  if and only if there exists a  $t \in (a, b)$  such that  $\delta(x, y, t) = 0$ . Since  $(x, y) \mapsto \delta(x, y, t)$  is measurable for each  $t$  it follows that  $A$  is measurable. Let  $T(x, y) = \{t : \delta(x, y, t) = 0\}$ . Then by the Kuratowski–Ryll–Nardzewski measurable selection theorem (theorem 6.9.4, [Bogachev 2007](#)) there exists a measurable function  $\tau : \text{int}(\text{dom}(f)) \times \text{int}(\text{dom}(f)) \rightarrow (0, 1)$  such that  $\tau(x, y) \in T(x, y)$  for all  $(x, y) \in \text{dom}(f) \times \text{dom}(f)$ . Therefore  $g(x, y) = \tau(x, y)x + (1 - \tau(x, y))y$  is measurable and the result is complete.

## Chapter 27 Exp3 for Adversarial Linear Bandits

**27.1** Let

$$\tilde{P}_t(a) = \frac{\exp(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a))}{\sum_{a' \in \mathcal{A}} \exp(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a'))}.$$

Then,

$$P_t = (1 - \gamma)\tilde{P}_t + \gamma\pi. \quad (27.1)$$

Let  $\hat{L}_n(a) = \sum_{t=1}^n \hat{Y}_t(a)$ ,  $\hat{L}_n = \sum_{t=1}^n \langle P_t, \hat{Y}_t \rangle$  and  $\tilde{L}_n = \sum_{t=1}^n \langle \tilde{P}_t, \hat{Y}_t \rangle$ , where we abuse  $\langle \cdot, \cdot \rangle$  by defining  $\langle p, y \rangle = \sum_{a \in \mathcal{A}} p(a)y(a)$  for  $p, y : \mathcal{A} \rightarrow \mathbb{R}$ . Then,

$$R_n = \max_{a \in \mathcal{A}} R_n(a) \quad \text{where} \quad R_n(a) = \mathbb{E} \left[ \sum_{t=1}^n \langle A_t, y_t \rangle - \langle a, y_t \rangle \right].$$

As in the proof of Theorem 11.1,

$$R_n(a) = \mathbb{E} \left[ \hat{L}_n - \hat{L}_n(a) \right].$$



Now, by (27.1),

$$\hat{L}_n = (1 - \gamma)\tilde{L}_n + \gamma \sum_{t=1}^n \langle \pi, \hat{Y}_t \rangle.$$

Repeating the steps of the proof of Theorem 11.1 shows that, thanks to  $\eta \hat{Y}_t(a) \geq -1$ ,

$$\begin{aligned} \tilde{L}_n &\leq \hat{L}_n(a) + \frac{\log k}{\eta} + \eta \sum_{t=1}^n \langle \tilde{P}_t, \hat{Y}_t^2 \rangle \\ &\leq \hat{L}_n(a) + \frac{\log k}{\eta} + \frac{\eta}{1 - \gamma} \sum_{t=1}^n \langle P_t, \hat{Y}_t^2 \rangle, \end{aligned} \quad (27.2)$$

where  $\hat{Y}_t^2$  denotes the function  $a \mapsto \hat{Y}_t^2(a)$  and the second inequality used that  $\tilde{P}_t = \frac{P_t - \gamma\pi}{1 - \gamma} \leq \frac{P_t}{1 - \gamma}$ . Now,

$$\begin{aligned} \hat{L}_n - \hat{L}_n(a) &\leq \gamma \sum_{t=1}^n \langle \pi, \hat{Y}_t \rangle + (1 - \gamma)\hat{L}_n(a) + \frac{\log k}{\eta} + \eta \sum_{t=1}^n \langle P_t, \hat{Y}_t^2 \rangle - \hat{L}_n(a) \\ &= \frac{\log k}{\eta} + \eta \sum_{t=1}^n \langle P_t, \hat{Y}_t^2 \rangle + \gamma \sum_{t=1}^n \langle \pi - e_a, \hat{Y}_t \rangle, \end{aligned}$$

where  $e_a(a') = \mathbb{I}\{a = a'\}$ . Now, thanks to  $-1 \leq \langle a, y_t \rangle \leq 1$ ,

$$\mathbb{E} \left[ \langle \pi - e_a, \hat{Y}_t \rangle \right] = \langle \pi - e_a, y_t \rangle \leq 2.$$

Putting things together,

$$R_n \leq \max_a R_n(a) \leq \frac{\log k}{\eta} + 2\gamma n + \eta \sum_{t=1}^n \mathbb{E} \left[ \langle P_t, \hat{Y}_t^2 \rangle \right],$$

thus, finishing the proof.

**27.4** Note that it suffices to show that  $\|x\|_{B^{-1}}^2 - \|x\|_{A^{-1}}^2 = \|x\|_{B^{-1} - A^{-1}}^2 \geq 0$  for any  $x \in \mathbb{R}^d$ . 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  $\mathcal{L} = \{y \in \mathbb{R}^d : \|y\|_V \leq 1\}$ . Let  $Tx = V^{1/2}x$  and note that  $T^{-1}\mathcal{L} = T\mathcal{A} = B = \{u \in \mathbb{R}^d : \|u\|_2 \leq 1\}$ . Then let  $\mathcal{U}$  be an  $\varepsilon$ -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  $\text{span}(\mathcal{L}) = \mathbb{R}^d$ . Let  $\mathcal{E} = \{y \in \mathbb{R}^d : \|y\|_V \leq 1\}$  be the ellipsoid of maximum volume contained by  $\text{cl}(\mathcal{L})$ . Then let

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

which satisfies  $\mathcal{A} \subseteq \mathcal{E}_*$ . Since  $\text{span}(\mathcal{L}) = \mathbb{R}^d$  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  $\Pi(x) \in \arg\min_{x' \in \mathcal{A}} \|x - x'\|_{\mathcal{E}}$ . 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  $\text{cl}(\text{co}(\mathcal{A}))$ , which by the previous part has size at most  $(6d/\varepsilon)^d$ . The result follows by choosing  $\mathcal{C} = \{\Pi(x) : x \in \bar{\mathcal{C}}\}$  where  $\Pi$  is the projection onto  $\mathcal{A}$  with respect to  $\|\cdot\|_{\mathcal{E}}$  where  $\mathcal{E}$  is the maximum volume ellipsoid contained by  $\text{cl}(\text{co}(\mathcal{A}))$ .

**27.8** Consider the case when  $d = 1$ ,  $k = n$  and  $\mathcal{A} = \{1, -1, \varepsilon, \varepsilon/2, \varepsilon/4, \dots\}$  for suitably small  $\varepsilon$ . Then, for  $t = 1$ ,  $P_t$  is uniform on  $\mathcal{A}$  and hence  $Q_t^{-1} \approx 2/k$  and with probability  $1 - 2/k$ ,  $|\hat{Y}_t| \approx 1/k = 1/n$ . If  $\eta$  is small, then the algorithm will barely learn. If  $\eta$  is large, then it learns quickly that either 1 or  $-1$  is optimal, but is too unstable for small regret.

**27.9** We can copy the proof presented for the finite-action case in the solution to Exercise 27.1 in an almost verbatim manner: The minor change is that is that we need to replace the sums over the action space with integrals. In particular, here we have

$$\tilde{P}_t(a) = \frac{\exp(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a))}{\int_{\mathcal{A}} \exp(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a')) da'}$$

and  $\langle p, y \rangle = \int_{\mathcal{A}} p(a)y(a)da$  for  $p, y : \mathcal{A} \rightarrow \mathbb{R}$ . Now up to (27.2) everything is the same. Recall that the inequality in this display was obtained by using the steps of the proof of Theorem 11.1. Here, we need add a little detail because we need to change this inequality slightly.

We argue as follows: 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  $W_0 = \text{vol}(\mathcal{A})$  and

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

Following the proof of Theorem 11.1, thanks to  $-\eta \hat{Y}_t(a) \leq 1$ ,

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

Therefore,

$$\log W_n \leq \log(\text{vol}(\mathcal{A})) - \eta \sum_{t=1}^n \langle \tilde{P}_t, \hat{Y}_t \rangle + \eta^2 \sum_{t=1}^n \langle \tilde{P}_t, \hat{Y}_t^2 \rangle.$$

Recalling that  $\tilde{L}_n = \sum_{t=1}^n \langle \tilde{P}_t, \hat{Y}_t \rangle$ , a rearrangement of the previous display gives

$$\tilde{L}_n \leq \frac{1}{\eta} \log \left( \frac{\text{vol}(\mathcal{A})}{W_n} \right) + \eta \sum_{t=1}^n \langle \tilde{P}_t, \hat{Y}_t^2 \rangle. \quad (27.3)$$

Let  $a^* = \text{argmin}_{a \in \mathcal{A}} \sum_{t=1}^n \langle y_t, a \rangle$ . Note that

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

By adding and subtracting  $\hat{L}_n(a^*)$  to the right-side of Eq. (27.3) and using the last identity and the definition of  $W_n$ , we get

$$\tilde{L}_n \leq \hat{L}_n(a^*) + \frac{1}{\eta} \log(K_n) + \eta \sum_{t=1}^n \langle \tilde{P}_t, \hat{Y}_t^2 \rangle,$$

where

$$K_n = \frac{\text{vol}(\mathcal{A})}{\int \exp \left( -\eta \sum_{t=1}^n (\hat{Y}_t(a) - \hat{Y}_t(a^*)) \right) da}.$$

which is the inequality that replaces (27.2). In particular, the only difference between (27.2) and the above display is that in the above display  $\log(k)$  got replaced by  $\log(K_n)$ . From here, we can

follow the steps of the proof of Exercise 27.1 up to the end, to get

$$R_n \leq \frac{\mathbb{E}[\log(K_n)]}{\eta} + 2\gamma n + \eta \mathbb{E} \left[ \sum_{t=1}^n \langle P_t, \hat{Y}_t^2 \rangle \right],$$

The result is completed by noting that  $\gamma = \eta d$  and the same argument as in the proof of Theorem 27.1 to bound

$$\eta \mathbb{E} \left[ \sum_{t=1}^n \langle P_t, \hat{Y}_t^2 \rangle \right] \leq \eta d n.$$

**27.11** Throughout  $\|\cdot\| = \|\cdot\|_2$  is the standard Euclidean norm. By translating  $\mathcal{K}$  we may assume that  $x^* = \mathbf{0}$ . Note that  $\sup_{x,y \in \mathcal{K}} \langle x - y, u \rangle = \sup_{x \in \mathcal{K}} \langle x, u \rangle$  (the last equality uses that  $x^* = \mathbf{0}$ ). Clearly, the claim is equivalent to

$$\int_{\mathcal{K}} \exp(-\langle x, u \rangle) \frac{dx}{\text{vol}(\mathcal{K})} \geq \underbrace{\exp \left( -d(1 + \log_+ \sup_{x \in \mathcal{K}} \langle x, u \rangle / d) \right)}_{=: g(\sup_{x \in \mathcal{K}} \langle x, u \rangle)}.$$

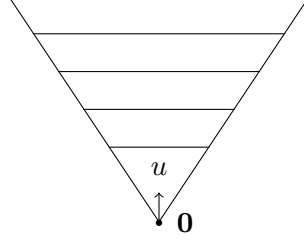
We claim that it suffices to show the result for the case when  $\|u\| = 1$ . Indeed, if the claim was true for  $\|u\| = 1$  then for any  $u \neq 0$  it would follow that

$$\begin{aligned} \frac{\int_{\mathcal{K}} \exp(-\langle x, u \rangle) dx}{\text{vol}(\mathcal{K})} &= \frac{\int_{\mathcal{K}} \exp(-\langle x\|u\|, u/\|u\| \rangle) dx}{\text{vol}(\mathcal{K})} = \frac{\int_{\|u\|\mathcal{K}} \exp(-\langle y, u/\|u\| \rangle) dy}{\|u\|^d \text{vol}(\mathcal{K})} \\ &= \frac{\int_{\|u\|\mathcal{K}} \exp(-\langle y, u/\|u\| \rangle) dy}{\text{vol}(\|u\|\mathcal{K})} \geq g\left(\sup_{x \in \|u\|\mathcal{K}} \langle x, u/\|u\| \rangle\right) = g\left(\sup_{x \in \mathcal{K}} \langle x, u \rangle\right). \end{aligned}$$

Hence, it remains to show the claim for vectors  $u$  such that  $\|u\| = 1$ . With an entirely similar reasoning we can show that it suffices to show the claim for the case when  $\text{vol}(\mathcal{K}) = 1$ . Hence, from now on we will assume these.

Introduce  $\alpha = \sup_{x \in \mathcal{K}} \langle x, u \rangle$ . For  $t \in [0, \alpha]$ , define  $f(t)$  to be the volume of the slice  $\mathcal{K}_t = \{x \in \mathcal{K} : \langle x, u \rangle = t\}$  with respect to the  $(d-1)$ -form. Since  $\text{vol}(\mathcal{K}) = 1$ ,  $f(t) > 0$  for  $0 < t < \alpha$  and  $1 = \int_0^\alpha f(t) dt$ . Clearly, we also have  $\int_{\mathcal{K}} \exp(-\langle x, u \rangle) dx = \int_0^\alpha f(t) \exp(-t) dt$ . Now, since  $\mathcal{K}$  is convex, for any  $t \in (0, \alpha)$ ,  $f(t) \geq \text{vol}_{t-1}(q/t\mathcal{K}_q)$  for any  $t \leq q \leq \alpha$ .

Since  $e^{-t}$  decreasing, a rearrangement argument shows that the function  $f$  that minimises  $\int_0^\alpha f(t) \exp(-t) dt$  and which satisfies the above properties is  $f(t) = (tf(\alpha))^{d-1}$  for a suitable value of  $f(\alpha)$  so that  $\int_0^\alpha f(t) dt = 1$  (we want the function to increase as fast as possible). Note that  $f(t)$  gives the volume of the  $t\tilde{\mathcal{K}}_\alpha$  for a suitable set  $\tilde{\mathcal{K}}_\alpha$ , as shown in the figure below:



The whole triangle is  $\tilde{\mathcal{K}} = \{t\tilde{\mathcal{K}}_\alpha : t \in [0, \alpha]\}$  with  $x^* = \mathbf{0}$  at the bottom corner. The thin lines represent  $t\tilde{\mathcal{K}}_\alpha$  for different values of  $t$ , which are  $(d-1)$ -dimensional subsets of  $\tilde{\mathcal{K}}$  that lie in affine spaces with normal vector  $u$ .

From the constraint  $\int_0^\alpha f(t)dt = 1$  we get  $f(\alpha)^{d-1} = (\int_0^\alpha t^{d-1}dt)^{-1}$ . We calculate

$$\begin{aligned} \int_{\mathcal{K}} \exp(-\langle x, u \rangle) dx &\geq \int_0^\alpha \exp(-t) (tf(\alpha))^{d-1} dt = \frac{\int_0^\alpha \exp(-t) t^{d-1} dt}{\int_0^\alpha t^{d-1} dt} \\ &\geq \min \left\{ 1, (d/\alpha)^d \right\} / e^d = g(\alpha), \end{aligned}$$

where the final inequality follows because  $\int_0^\alpha t^{d-1} dt = \alpha^d/d$  and

$$\int_0^\alpha t^{d-1} \exp(-t) dt \geq \frac{1}{e^d} \int_0^{\alpha \wedge d} t^{d-1} dt = \frac{\min(\alpha, d)^d}{e^d d}.$$

## Chapter 28 Follow-the-Regularised-Leader and Mirror Descent

**28.1** The mapping  $a \mapsto D_F(a, b)$  is the sum of Legendre function  $F$  and a linear function, which is clearly Legendre. Hence,  $\Phi$  is Legendre. Suppose now that  $c \in \partial \text{int}(\mathcal{D})$  and let  $d \in \mathcal{A} \cap \text{int}(\mathcal{D})$  be arbitrary. Then, the map  $\alpha \mapsto \Phi(\alpha c + (1-\alpha)d)$  must be decreasing. And yet,

$$\begin{aligned} \frac{d}{d\alpha} \Phi(\alpha c + (1-\alpha)d) &= \langle \nabla \Phi(\alpha c + (1-\alpha)d), c - d \rangle \\ &= \langle y + \nabla F(\alpha c + (1-\alpha)d), c - d \rangle, \end{aligned}$$

which converges to infinity as  $\alpha$  tends to one by Proposition 26.7 and is a contradiction.

**28.5** 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

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

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

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

where the inequality follows by the first-order optimality condition applied to  $a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \operatorname{dom}(F)} \Phi_t(a)$  and  $a_{t+2}$ . Substituting this into Eq. (28.1) completes the proof.

## 28.10

(a) For the first relation, direct calculation shows that  $\tilde{P}_{t+1,i} = P_{ti} \exp(-\eta \hat{Y}_{ti})$  and

$$\begin{aligned}
D_F(P_t, \tilde{P}_{t+1}) &= \sum_{i=1}^k P_{ti} \log \left( \frac{P_{ti}}{\tilde{P}_{t+1,i}} \right) - \sum_{i=1}^k P_{ti} + \sum_{i=1}^k \tilde{P}_{t+1,i} \\
&= \sum_{i=1}^k P_{ti} \left( \exp(-\eta \hat{Y}_{ti}) - 1 + \eta \hat{Y}_{ti} \right).
\end{aligned}$$

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  $\operatorname{diam}_F(\mathcal{P}_{k-1}) = \max_{p,q \in \mathcal{P}_{k-1}} F(p) - F(q) \leq \log(k)$ .

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

**28.11** 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

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

Summing completes the proof. For the second part use the generalised Pythagorean theorem (Exercise 26.13) and positivity of the Bregman divergence to argue that  $D(a, \tilde{a}_{t+1}) \geq D(a, a_{t+1})$ .

## 28.12

(a) We use the same argument as in the solution to Exercise 28.5. 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 28.5, but now we have to keep track of the changing potentials:

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

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

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

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  $F_t(a) = (F(a) - \min_{b \in \mathcal{A}} F(b))/\eta_t$  immediately gives the result.

### 28.13

(a) Apply your solution to Exercise 26.11.

(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 28.12 combined with the fact that  $\text{diam}_F(\mathcal{A}) = \log(k)$ .

(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 26.13 with  $H = \nabla^2 f(q) = \text{diag}(1/q)$  for some  $q \in [P_t, P_{t+1}]$  we have

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

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

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

Therefore

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

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

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

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

### 28.14

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

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



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

$$\sum_{t=1}^n \eta_t y_{tA_t}^2 \leq 2\sqrt{k \sum_{t=1}^n y_{tA_t}^2 \log(n/k)}. \quad (28.2)$$

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

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

For the remainder of the proof let  $p \in [1/n, 1] \cap \mathcal{P}_{k-1}$  be arbitrary. Notice that  $F(p) - \min_{q \in \mathcal{P}_{k-1}} F(q) \leq k \log(n/k)$ . By the result in Exercise 28.12,

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

If  $P_{t+1, A_t} \geq P_{tA_t}$ , then  $\langle P_t - P_{t+1}, \hat{Y}_t \rangle \leq 0$ . Now suppose that  $P_{t+1, A_t} \leq P_{tA_t}$ . By Theorem 26.12, there exists a  $\xi \in [P_t, P_{t+1}]$  such that

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

By the definition of  $\eta_n$  and Eq. (28.2),

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

Therefore by Eq. (28.3),

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

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 in  $R_n$  shows that for a suitably large universal constant  $C$ ,

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

**28.15** The first parts are mechanical and are skipped.

(e) By Part (c), Theorem 28.5 and Theorem 26.13,

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

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

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

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

**28.16**

(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,

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

Taking the limit as  $n$  tends to infinity shows that

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

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

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

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 generalisation of the above argument and the analysis in Proposition 28.6 shows

that

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

In the same manner,

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

Hence

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

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  $(x_n, y_n)$  may not have a cluster point that is a saddle point.

**28.17** 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  $\sup_{y \in Y} \inf_{x \in X} f(x, y) = -\infty$ . For a bounded example, consider  $X = Y = [1, \infty)$  and

$$f(x, y) = y/(x + y).$$

## 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. Indeed,

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

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

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

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  $\eta$ .

### 29.4

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

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

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

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

which, given that  $\bar{A}_t$  is  $\mathcal{F}_{t-1}$ , implies that

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

Combining with Eq. (29.1) shows that

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

Then we need to check that  $\eta\|\hat{\theta}_t\|_2 = \eta\|dE_t A_t Y_t / (1 - \|\bar{A}_t\|_2)\|_2 \leq \eta d / (1 - r) \leq 1/2$ . 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  $\Omega(d\sqrt{n} + \varepsilon n)$ . The  $\sqrt{d}$  factor cannot be improved greatly, but the argument is more complicated [Lattimore and Szepesvári, 2019].

## Chapter 30 Combinatorial Bandits

### 30.4

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

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

Taking the expectation of both sides yields the result.

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

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

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.

**30.5** Since  $\mathcal{A}$  is compact,  $\text{dom}(\phi) = \mathbb{R}^d$ . Let  $\mathcal{D}$  be the set of points  $x$  at which  $\phi$  is differentiable. Then, as noted in the hint,  $\lambda(\mathbb{R}^d \setminus \mathcal{D}) = 0$  where  $\lambda$  is the Lebesgue measure. Since  $Q \ll \lambda$ , then  $Q(\mathbb{R}^d \setminus \mathcal{D}) = 0$  as well. Define  $a(x) = \arg\max_{a \in \mathcal{A}} \langle a, x \rangle$ . Let  $v \in \mathbb{R}^d$  be non-zero. Then, by the second part of the hint, the directional derivative of  $\phi$  is

$$\nabla_v \phi(x) = \max_{a \in \mathcal{A}(x)} \langle a, v \rangle.$$

By the last part of the hint, for  $x \in \mathcal{D}$  this implies that  $\mathcal{A}(x)$  is a singleton and thus  $\nabla \phi(x) = a(x)$ . Let  $q = \frac{dQ}{d\lambda}$  be the density of  $Q$  with respect to the Lebesgue measure. Then, for any  $v \in \mathbb{R}^d$ ,

$$\begin{aligned} \nabla_v \int_{\mathbb{R}^d} \phi(x+z) q(z) dz &= \int_{\mathbb{R}^d} \nabla_v \phi(x+z) q(z) dz \\ &= \int_{\mathcal{D}+\{x\}} \nabla_v \phi(x+z) q(z) dz \\ &= \int_{\mathcal{D}+\{x\}} \langle a(x+z), v \rangle q(z) dz \\ &= \left\langle \int_{\mathcal{D}+\{x\}} a(x+z) q(z) dz, v \right\rangle, \end{aligned}$$

where the exchange of limit (hidden in the derivative) and integral is justified by the dominated convergence theorem. By the last part of the hint, since  $v \in \mathbb{R}^d$  was arbitrary, it follows that  $\nabla \int_{\mathbb{R}^d} \phi(x+z) q(z) dz$  exists and is equal to  $\int_{\mathcal{D}+\{x\}} a(x+z) q(z) dz = \mathbb{E}[a(x+Z)]$ .

### 30.6

(a) To show that  $F$  is well defined we need to show that  $F^*$  is the Fenchel dual of a unique proper convex closed function. Let  $g = (F^*)^*$ . It is not hard to see that  $F^*$  is a proper convex function,  $\text{dom}(F^*) = \mathbb{R}^d$ , and hence the epigraph of  $F^*$  is closed. Then, by the hint,  $g^* = (F^*)^{**} = F^*$ , hence  $F^*$  is the Fenchel dual of  $g$ . By the hint, the Fenchel dual of  $g$  is a proper convex closed function, so we can take  $F = g^*$ . It remains to show that there is only a single proper convex closed function whose Fenchel dual is  $F^*$ . To show this let  $g, h$  be proper convex closed functions such that  $g^* = h^* = F^*$ . Then  $g = g^{**} = (F^*)^* = h^{**} = h$ , hence,  $F$  is uniquely defined.

- (b) By Part (c) of Theorem 26.6, it suffices to show that  $F^*$  is Legendre. As noted earlier, the domain of  $F^*$  is all of  $\mathbb{R}^d$ . From Exercise 30.5, it follows that  $F^*$  is everywhere differentiable. Part (c) of the definition of Legendre functions is automatically satisfied since  $\partial\mathbb{R}^d = \emptyset$ , hence it remains to prove that  $F^*$  is strictly convex.

To prove this, we need to prove that for all  $x \neq y$ ,

$$F^*(y) > F^*(x) + \langle y - x, \nabla F^*(x) \rangle.$$

Let  $a(x) = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, x \rangle$ , with ties broken arbitrarily and let  $q = \frac{dQ}{d\lambda}$  be the density of  $Q$  with respect to the Lebesgue measure. Recalling the definitions and the result of Exercise 30.5,

$$\begin{aligned} & F^*(y) - F^*(x) - \langle y - x, \nabla F^*(x) \rangle \\ &= \int_{\mathbb{R}^d} \phi(y + z) q(z) dz - \int_{\mathbb{R}^d} \phi(x + z) q(z) dz - \left\langle y - x, \int_{\mathbb{R}^d} a(x + z) q(z) dz \right\rangle \\ &= \int_{\mathbb{R}^d} \langle y + z, a(y + z) - a(x + z) \rangle q(z) dz \\ &= \int_{\mathbb{R}^d} \langle u, a(u) - a(u + \delta) \rangle q(u - y) du, \end{aligned}$$

where  $\delta = x - y$ . Clearly the term  $f(u) = \langle a(u) - a(u + \delta), u \rangle$  is nonnegative for any  $u \in \mathbb{R}^d$ . Since by assumption  $q > 0$ , it suffices to show that  $f$  is strictly positive over a neighborhood of zero that has positive volume. The assumption that  $\operatorname{span}(\mathcal{A}) = \mathbb{R}^d$  means that  $\langle a(-\delta/2) - a(\delta/2), -\delta/2 \rangle = \varepsilon > 0$ . To see this, notice that

$$\begin{aligned} \operatorname{co}(\mathcal{A}) &\subset \{x : \langle x, \delta/2 \rangle \leq \langle a(\delta/2), \delta/2 \rangle\} \quad \text{and} \\ \operatorname{co}(\mathcal{A}) &\subset \{x : \langle x, -\delta/2 \rangle \leq \langle a(-\delta/2), -\delta/2 \rangle\} = \{x : \langle x, \delta/2 \rangle \geq \langle a(-\delta/2), \delta/2 \rangle\}. \end{aligned}$$

Were it the case that  $\langle a(-\delta/2), \delta/2 \rangle = \langle a(\delta/2), \delta/2 \rangle$ , then  $\operatorname{co}(\mathcal{A})$  would be a subset of a  $(d-1)$ -dimensional hyperplane, contradicting the assumption that  $\operatorname{span}(\mathcal{A}) = \mathbb{R}^d$ . Let  $u = -\delta/2$  and  $\|\cdot\| = \|\cdot\|_2$  and  $\operatorname{diam}(\mathcal{A}) = \operatorname{diam}_{\|\cdot\|}(\mathcal{A})$ . Then

$$\begin{aligned} \langle a(-\delta/2 + v), -\delta/2 \rangle &= \langle a(v - \delta/2), v - \delta/2 \rangle - \langle a(v - \delta/2), v \rangle \\ &\geq \langle a(-\delta/2), v - \delta/2 \rangle - \|v\| \operatorname{diam}(\mathcal{A}) \\ &\geq \langle a(-\delta/2), -\delta/2 \rangle - 2\|v\| \operatorname{diam}(\mathcal{A}). \end{aligned}$$



Similarly,  $\langle a(v + \delta/2), \delta/2 \rangle \geq \langle a(\delta/2), \delta/2 \rangle - 2\|v\|\text{diam}(\mathcal{A})$  and hence

$$\begin{aligned}
f(u + v) &= \langle a(u + v) - a(u + \delta + v), u + v \rangle \\
&\geq \langle a(u + v) - a(u + \delta + v), u \rangle - 2\|v\|\text{diam}(\mathcal{A}) \\
&= \langle a(v - \delta/2) - a(v + \delta/2), -\delta/2 \rangle - 2\|v\|\text{diam}(\mathcal{A}) \\
&= \langle a(v - \delta/2), -\delta/2 \rangle + \langle a(\delta/2 + v), \delta/2 \rangle - 2\|v\|\text{diam}(\mathcal{A}) \\
&\geq \langle a(-\delta/2), -\delta/2 \rangle + \langle a(\delta/2), \delta/2 \rangle - 6\|v\|\text{diam}(\mathcal{A}) \\
&= \langle a(-\delta/2) - a(\delta/2), -\delta/2 \rangle - 6\|v\|\text{diam}(\mathcal{A}) \\
&= \varepsilon - 6\|v\|\text{diam}(\mathcal{A}).
\end{aligned}$$

Thus, for sufficiently small  $\|v\|$ , it holds that  $f(u + v) \geq \varepsilon/2$  and the claim follows.

- (c) We need to show that  $\text{int}(\text{dom}(F)) = \text{int}(\text{co}(\mathcal{A}))$ . By the first two parts of the exercise,  $F$  is Legendre, and hence by Part (a) of Theorem 26.6 and by Exercise 30.5, we have

$$\text{int}(\text{dom}(F)) = \nabla F^*(\mathbb{R}^d) = \left\{ \int_{\mathbb{R}^d} a(x + z)q(z)dz : x \in \mathbb{R}^d \right\}.$$

Clearly, this is a subset of  $\text{int}(\text{co}(\mathcal{A}))$ . To establish the equality, by convexity of  $\text{co}(\mathcal{A})$  it suffices to show that for any extreme point  $a \in \mathcal{A}$  and  $\varepsilon > 0$  there exists an  $x$  such that  $\|\nabla F^*(x) - a\| \leq \varepsilon$ . To show this, choose a vector  $x_0 \in \mathbb{R}^d$  so that  $a(x_0 + v) = a$  for any  $v$  in the unit ball centered at zero. Such a vector exist because of the conditions on  $\mathcal{A}$ . Let  $K_\varepsilon$  be a closed ball centered at zero such that  $Q(K_\varepsilon) \geq 1 - \varepsilon/(\max_{a \in \mathcal{A}} \|a\|)$ . This exist because  $Q(\mathbb{R}^d) = 1$ . Let  $r$  be the radius of  $K_\varepsilon$ . Pick any  $c > r_\varepsilon$ . Then, for any  $v \in K_\varepsilon$ ,  $a(cx_0 + v) = a(x_0 + v/c) = a$  and hence  $\nabla F^*(cx_0) = s + \int_{K_\varepsilon} a(cx_0 + z)q(z) = s + a$ , where  $\|s\| \leq \varepsilon$ , finishing the proof.

**30.8** Following the advice, assume that the learner plays  $m$  bandits in parallel, each having  $k = d/m$  actions. Let  $R_{ni}$  be the regret of the learner in the  $i$ th bandit. Then,  $R_{ni} \geq c\sqrt{nk} = c\sqrt{nd/m}$  for some universal constant  $c > 0$ . Further, if  $R_n$  is the regret of the learner,  $R_n = \sum_{i=1}^m R_{ni}$ . Hence,  $R_n \geq c\sqrt{ndm}$ .

An alternative to this is to emulate a  $k = d/m$ -armed bandit with scaled rewards: For this imagine that the  $d$  items (components of the combinatorial action) are partitioned into  $k$  parts, each having  $m$  items in it. Unlike in multi-task bandits, the learner needs to choose a part and receives feedback for all the items in it. Hence, the the rewards received belong to the  $[0, m]$  interval and we also get  $R_n \geq cm\sqrt{nk} = c\sqrt{ndm}$ .

## Chapter 31 Non-stationary Bandits

**31.1** As suggested, Exp4 is used with each element of  $\Gamma_{nm}$  identified with one expert. Consider an arbitrary enumeration of  $\Gamma_{nm} = \{a^{(1)}, \dots, a^{(G)}\}$  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  $E_{g,j}^t = \mathbb{I}\{a_t^{(g)} = j\}$ ,  $j \in [k]$ . 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  $t = 1, \dots, n-1$ . 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,  $G \leq k^m \sum_{s=0}^{m-1} \binom{n-1}{s} = k^m \Phi_{m-1}(n-1) \leq k^m \Phi_m(n)$ . Now note that for  $n \geq m$ ,  $0 \leq m/n \leq 1$ , hence

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

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

**31.3** Use the construction and analysis in Exercise 11.6 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

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

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

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

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

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

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

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

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

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

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

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

For the second term that makes up (A),

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

Hence (A)  $\leq nm^2$  and  $R_n \leq nm\mathbb{P}(F_n) + m\ell + \sqrt{4m^3\ell n \log\left(\frac{c\sqrt{n}}{\delta}\right)}$  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 optimisation problems and so

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

where we used the positive homogeneity of  $f(\alpha)$  and the  $\ell_1$  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

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

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

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

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

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

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.

### 33.5

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

$$\liminf_{n \rightarrow \infty} \frac{-\log(\mathbb{P}_{\nu\pi}(\Delta_{A_{n+1}} > 0))}{\log(n)} > 1 + \varepsilon. \quad (33.1)$$

Notice that if Eq. (33.1) were not true then we would be done. Then let  $\nu'$  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,

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

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

$$\mathbb{P}_{\nu\pi}(A_{n+1} \neq 1) + \mathbb{P}_{\nu'\pi}(A_{n+1} \neq i) \geq \frac{1}{2} \left(\frac{1}{n}\right)^{1+\varepsilon+\varepsilon_n},$$

where  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ . Using Eq. (33.1) shows that

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

which implies that

$$\liminf_{n \rightarrow \infty} \frac{-\log(\mathbb{P}_{\nu'\pi}(A_{n+1} \neq i))}{\log(n)} \leq \limsup_{n \rightarrow \infty} \frac{-\log(\mathbb{P}_{\nu'\pi}(A_{n+1} \neq i))}{\log(n)} \leq 1 + \varepsilon.$$

(b) No. Consider the algorithm that plays UCB on rounds  $t \notin \{2^{k^2} : k \in \mathbb{N}\}$  and otherwise plays round-robin.

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

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

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

### 33.6

(a) Assume without loss of generality that arm 1 is unique in  $\nu$ . By the work in Part (a) of Exercise 33.4,  $\alpha^*(\nu) = \operatorname{argmax}_{\alpha \in \mathcal{P}_{k-1}} \Phi(\nu, \alpha)$  with

$$\begin{aligned} \Phi(\nu, \alpha) &= \frac{1}{2} \inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^k \alpha_i (\mu_i(\nu) - \mu_i(\tilde{\nu}))^2 \\ &= \frac{1}{2} \min_{i>1} \frac{\alpha_1 \alpha_i \Delta_i^2}{\alpha_1 + \alpha_i} = \frac{1}{2} \min_{i>1} f_i(\alpha_1, \alpha_i) \end{aligned}$$

where  $\Phi(\nu, \alpha) = 0$  if  $\alpha_i = 0$  for any  $i$  and the last equality serves as the definition of  $f_i$ . The function  $\Phi(\nu, \cdot)$  is the minimum of a collection of concave functions and hence concave. Abbreviate  $\alpha^* = \alpha^*(\nu)$  and notice that  $\alpha^*$  must equalize the functions  $(f_i)$  so that  $f_i(\alpha_1^*, \alpha_i^*)$  is constant for  $i > 1$ . Hence, for all  $i > 1$ ,

$$\frac{1}{2} \frac{\alpha_1^* \alpha_i^* \Delta_i^2}{\alpha_1^* + \alpha_i^*} = \max_{\alpha \in \mathcal{P}_{k-1}} \Phi(\nu, \alpha) = \Phi(\nu).$$

Rearranging shows that

$$\alpha_i^* = \frac{2\alpha_1^* \Phi(\nu)}{\Delta_i^2 \alpha_1^* - 2\Phi(\nu)}.$$

Therefore

$$\alpha_i^* + \sum_{i=2}^k \frac{2\alpha_1^* \Phi(\nu)}{\Delta_i^2 \alpha_1^* - 2\Phi(\nu)} = 1.$$

The solutions to this equation are the roots of a polynomial and by the fundamental theorem of algebra, either this polynomial is zero or there are finitely many roots. Since the former is clearly not true, we conclude there are at most finitely many maximisers. Yet concavity of the objective means that the number of maximisers is either one or infinite. Therefore there is a unique maximiser.

- (b) Notice that  $i^*(\xi) = i^*(\nu)$  whenever  $d(\xi, \nu)$  is sufficiently small. Hence, by the previous part, the function  $\Phi(\cdot, \cdot)$  is continuous at  $(\nu, \alpha)$  for any  $\alpha$ . Suppose that  $\alpha^*(\cdot)$  is not continuous at  $\nu$ . Then there exists a sequence  $(\nu_n)_{n=1}^\infty$  with  $\lim_{n \rightarrow \infty} d(\nu_n, \nu) = 0$  and for which  $\liminf_{n \rightarrow \infty} \|\alpha^*(\nu) - \alpha^*(\nu_n)\|_\infty > 0$ . By compactness of  $\mathcal{P}_{k-1}$ , the sequence  $\alpha^*(\nu_n)$  has a cluster point  $\alpha_\infty^*$ , which by assumption must satisfy  $\alpha^*(\nu) \neq \alpha_\infty^*$ . And yet, taking limits along an appropriate subsequence,  $\Phi(\alpha^*(\nu), \nu) = \lim_{n \rightarrow \infty} \Phi(\alpha^*(\nu), \nu_n) \leq \lim_{n \rightarrow \infty} \Phi(\alpha^*(\nu_n), \nu_n) = \Phi(\alpha_\infty^*, \nu)$ . Therefore by Part (a),  $\alpha^*(\nu) = \alpha_\infty^*$ , which is a contradiction.

- (c) We'll be a little lackadaisical about constants here. Define random variable

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

which by the usual concentration analysis and union bounding satisfies  $\mathbb{P}(\Lambda \geq x) \leq 1/x$ . Therefore

$$\mathbb{E}[\log(\Lambda)^2] = \int_0^\infty \mathbb{P}(\Lambda \geq \exp(x^{1/2})) dx \leq \int_0^\infty \exp(-x^{1/2}) dx = 2.$$

By the definition of  $\lambda$ ,

$$\tau_\nu(\varepsilon) \leq 1 + \max \left\{ t : \sqrt{\frac{2 \log(\Lambda kt(t+1))}{\min_i T_i(t)}} > \varepsilon \right\}.$$

The forced exploration in the algorithm means that  $T_i(t) = \Omega(\sqrt{t})$  almost surely and hence

$$\mathbb{E}[\tau_\nu(\varepsilon)] = O\left(\mathbb{E}[\log(\Lambda)^2]\right) = O(1).$$

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

- (e) 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

$$t \geq \max \left\{ \frac{2k\tau_\alpha(\varepsilon/(2k))}{\varepsilon}, \frac{16k^2}{\varepsilon^2} \right\}.$$



Then the definition of the algorithm implies that

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

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

$$T_i(t) \geq t - \sum_{j \neq i} T_j(t) \geq t - \sum_{j \neq i} t \left( \alpha_j^*(\nu) + \frac{\varepsilon}{k} \right) \geq t(\alpha_i^*(\nu) - \varepsilon).$$

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

$$\mathbb{E} \left[ \max \left\{ \frac{2k\tau_\alpha(\varepsilon/(2k))}{\varepsilon}, \frac{16k^2}{\varepsilon^2} \right\} \right] < \infty.$$

(f) Given  $\varepsilon > 0$  let  $\tau_\beta(\varepsilon) = 1 + \max \{t : t\Phi(\nu, \alpha^*(\nu)) < \beta_t(\delta) + \varepsilon t\}$  and

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

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

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

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  $\delta \rightarrow 0$  and using the previous parts shows that for any sufficiently small  $\varepsilon > 0$ ,

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

Continuity of  $\Phi(\cdot, \cdot)$  at  $(\nu, \alpha^*(\nu))$  ensures that  $\lim_{\varepsilon \rightarrow 0} u(\varepsilon) = 0$  and the result follows since  $c^*(\nu) = 1/\Phi(\nu, \alpha^*(\nu))$ . Note that taking the limit as  $\delta \rightarrow 0$  only works because the policy does not depend on  $\delta$ . Hence the expectations of  $\tau_\nu(\varepsilon)$ ,  $\tau_\alpha(\varepsilon)$  and  $\tau_T(\varepsilon)$  do not depend on  $\delta$ .

### 33.7

(a) Recalling the definitions,

$$H_1(\mu) = \sum_{i=1}^k \min \left\{ \frac{1}{\Delta_{\min}^2}, \frac{1}{\Delta_i^2} \right\} \quad \text{and} \quad H_2(\mu) = \max_{i: \Delta_i > 0} \frac{i}{\Delta_i^2},$$

where  $\Delta_1 \leq \Delta_2 \leq \dots \leq \Delta_k$ . Now, for any  $i \in [k]$  with  $\Delta_i > 0$ ,

$$H_1(\mu) \geq \sum_{j=1}^i \min \left\{ \frac{1}{\Delta_{\min}^2}, \frac{1}{\Delta_j^2} \right\} \geq \sum_{j=1}^i \min \left\{ \frac{1}{\Delta_{\min}^2}, \frac{1}{\Delta_i^2} \right\} \geq \frac{i}{\Delta_i^2}.$$

Therefore  $H_1(\mu) \geq H_2(\mu)$ . For the second inequality, let  $i_{\min} = \min\{i : \Delta_i > 0\}$ . Then

$$\begin{aligned} H_1(\mu) &= \frac{i_{\min}}{\Delta_{\min}^2} + \sum_{i=i_{\min}+1}^k \frac{1}{i} \frac{i}{\Delta_i^2} \\ &\leq \left( 1 + \sum_{i=i_{\min}+1}^k \frac{1}{i} \right) H_2(\mu) \leq (1 + \log(k)) H_2(\mu). \end{aligned}$$

The result follows because  $i_{\min} > 1$  and because  $\sum_{i=3}^k \frac{1}{i} \leq \log(k)$ .

(b) When  $\Delta_2 = \dots = \Delta_k > 0$  it holds that  $H_1(\mu) = H_2(\mu)$ . For the other direction let  $\Delta_i = \sqrt{i}$  for  $i \geq 2$  so that  $i/\Delta_i^2 = 1 = H_2(\mu)$  and

$$H_1(\mu) = 1 + \sum_{i=3}^k \frac{1}{i} = L = L H_2(\mu).$$

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

## Chapter 34 Foundations of Bayesian Learning

### 34.4

(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  $\theta$  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  $x \in N$ ,

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

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.

**34.5** 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:

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

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  $\theta \in \text{int}(\text{dom}(A))$  there exists a neighborhood  $N$  of  $\theta$  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  $\varepsilon$  it follows by the dominated convergence theorem that the derivative and integral can be exchanged.

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

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

(d) This is another straightforward calculation:

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

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

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

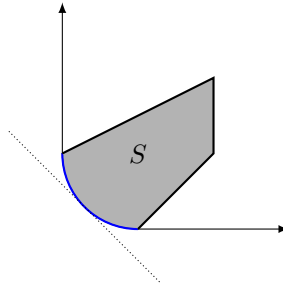
A symmetric calculation shows that for  $\theta' < \theta$ ,

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

**34.13** Let  $\pi^*$  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  $\pi$ ,

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

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



### 34.14

- (a) Suppose that  $\pi$  is not admissible. Then there exists another policy  $\pi'$  with  $\ell(\pi', \nu) \leq \ell(\pi, \nu)$  for all  $\nu \in \mathcal{E}$ . Clearly  $\pi'$  is also Bayesian optimal. But  $\pi$  was unique, which is a contradiction.
- (b) Suppose that  $\Pi = \{\pi_1, \pi_2\}$  and  $\mathcal{E} = \{\nu_1, \nu_2\}$  and  $\ell(\pi, \nu_1) = 0$  for all  $\pi$  and  $\ell(\pi_i, \nu_2) = \mathbb{I}\{i = 2\}$ . Then any policy is Bayesian optimal for  $Q = \delta_{\nu_1}$ , but  $\pi_2$  is dominated by  $\pi_1$ .
- (c) Suppose that  $\pi$  is not admissible. Then there exists another policy  $\pi'$  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.

- (d) Repeat the previous solution with the restriction to the support.

**34.15** Let  $\Pi$  be the set of all policies and  $\Pi_D = \{e_1, \dots, e_N\}$  the set of all deterministic policies, which is finite. A policy  $\pi \in \Pi$  can be viewed as a probability measure on  $(\Pi_D, \mathfrak{B}(\Pi_D))$ , which is the essence of Kuhn's theorem on the equivalence of behavioral and mixed strategies in extensive form games. Note that since  $\Pi_D$  is finite, probability measures on  $(\Pi_D, \mathfrak{B}(\Pi_D))$  can be viewed as distributions in  $\mathcal{P}_{N-1}$ . In this way  $\Pi$  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  $\Pi$  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  $\mathcal{E}_j = \{\nu_1, \dots, \nu_j\}$ . Using the notation of the previous exercise, let  $R_{n,j}(\pi) = (R_n(\pi, \nu_1), \dots, R_n(\pi, \nu_j))$ ,  $S_j = R_{n,j}(\Pi) \subset \mathbb{R}^j$  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  $\pi^{\text{adm}} \in \Pi$  be an admissible policy. Then  $R_{n,j}(\pi^{\text{adm}}) \in \lambda(S_j)$  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  $\text{BR}_n(\pi^{\text{adm}}, Q_j) \leq \min_{\pi} \text{BR}_n(\pi, Q_j)$ . 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  $\pi$ ,

$$\lim_{i \rightarrow \infty} \text{BR}_n(\pi, Q_i) = \lim_{i \rightarrow \infty} \int_{\mathcal{E}} R_n(\pi, \nu) dQ_i(\nu) = \text{BR}_n(\pi, Q).$$

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

**34.16** Clearly

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

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  $\Pi$  be the space of all policies

with the discrete topology and

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

which is a convex subspace of the topological vector space of all signed measures on  $(\Pi, 2^\Pi)$  with the weak\* topology. Let  $\mathcal{L} : \mathcal{P} \times \mathcal{Q} \rightarrow [0, n]$  be defined by

$$\mathcal{L}(S, Q) = - \int_{\Pi} \int_{\mathcal{E}} R_n(\pi, \nu) dQ(\nu) dS(\pi) = - \int_{\mathcal{E}} R_n(\pi_S, \nu) dQ(\nu),$$

where  $\pi_S$  a policy such that  $\mathbb{P}_{\nu\pi_S} = \int_{\Pi} \mathbb{P}_{\nu\pi} dS(\pi)$ , which is defined in Exercise 4.4. The regret is bounded in  $[0, n]$  and the discrete topology on  $\Pi$  means that all functions from  $\Pi$  to  $\mathbb{R}$  are continuous, including  $\pi \mapsto \int_{\mathcal{E}} R_n(\pi, \nu) Q(d\nu)$ . By the definition of the weak\* topology on  $\mathcal{P}$  it holds that  $\mathcal{L}(\cdot, Q)$  is continuous in its first argument for all  $Q$ . The integral over  $\Pi$  with respect to  $S \in \mathcal{P}$  is a finite sum and  $\nu \mapsto R_n(\pi, \nu)$  is continuous for all  $\pi$  by the result in Exercise 14.4. Therefore  $\mathcal{L}$  is continuous and linear in both arguments. By Sion's theorem (Theorem 28.12),

$$- \max_{Q \in \mathcal{Q}} \text{BR}_n^*(Q) = \min_{Q \in \mathcal{Q}} \sup_{S \in \mathcal{P}} \mathcal{L}(S, Q) = \sup_{S \in \mathcal{P}} \min_{Q \in \mathcal{Q}} \mathcal{L}(S, Q) = - \inf_{S \in \mathcal{P}} R_n^*(\pi_S, \mathcal{E}),$$

Therefore

$$\max_{Q \in \mathcal{Q}} \text{BR}_n^*(Q) = \inf_{S \in \mathcal{P}} R_n^*(\pi_S, \mathcal{E}) = \inf_{\pi \in \Pi} R_n^*(\pi, \mathcal{E}) = R_n^*(\mathcal{E}).$$

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

## Chapter 35 Bayesian Bandits

**35.1** Let  $\pi$  be the policy of MOSS from Chapter 9, which for any 1-subgaussian bandit  $\nu$  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  $C' = Ck$ ,

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

The first part follows since  $\cap_n \mathcal{E}_n = \emptyset$  and thus  $\lim_{n \rightarrow \infty} Q(\mathcal{E}_n) = 0$  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  $n \in \{1, 2, 4, 8, \dots\}$  and denote them  $Q_1, Q_2, \dots$ . 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  $\tau$  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 \[2006\]](#),

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

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



that

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

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

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

where the last equality follows from the monotone convergence theorem. Next, let  $V_n = v(S_n)$ . Note that  $\lim_{n \rightarrow \infty} \mathbb{E}_x[U_n] = \mathbb{E}_x[\lim_{n \rightarrow \infty} U_n] = \mathbb{E}_x[U_\infty]$ , where the exchange of the limit and expectation is justified by the dominated convergence theorem because  $\sup_n |U_n|$  is  $\mathbb{P}_x$  integrable. By definition,  $V_n \geq U_n$ . Hence,

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

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  $V_\infty = \lim_{n \rightarrow \infty} V_n = U_\infty$   $\mathbb{P}_x$ -a.s. Then

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

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

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

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  $\tau$  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

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

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

**35.6** Fix  $x \in \mathcal{S}$  and let

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

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

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

which implies that

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

Moving now to (b), first note that  $v_\gamma(x) \geq 0$  for any  $\gamma \in \mathbb{R}$  because when  $\tau = 1$ ,  $\mathbb{E}_x \left[ \sum_{t=1}^{\tau-1} \alpha^{t-1} (r(S_t) - \gamma) \right] = 0$ . 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  $\tau \geq 2$ ,

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

which implies that

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

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

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

finishing the proof.

**35.7** We want to apply Theorem 35.3. The difficulty is that Theorem 35.3 considers the case where the reward depends only on the current state, while here the reward accumulates. The solution is to augment the state space to include the history. We use the convention that if  $x \in \mathcal{S}^n$  and  $y \in \mathcal{S}^m$ , then  $xy \in \mathcal{S}^{n+m}$  is the concatenation of  $x$  and  $y$ . In particular, the  $i$ th component of  $xy$  is  $x_i$  if  $i \leq n$  and  $y_{i-n}$  if  $i > n$ . We will also denote by  $x_{1:n}$  the sequence  $(x_1, \dots, x_n)$  formed from  $x_1, \dots, x_n$ . Recall that  $\mathcal{S}^*$  is the set of all finite sequences with elements in  $\mathcal{S}$  and let  $\mathcal{G}^*$  be the  $\sigma$ -algebra given by

$$\mathcal{G}^* = \sigma \left( \bigcup_{n=0}^{\infty} \mathcal{G}^n \right).$$

For  $n \geq 1$ , let  $S^n = (S_1, \dots, S_n)$ . The sequence  $(S^n)_{n=1}^{\infty}$  is a Markov chain on the space of finite sequences  $(\mathcal{S}^*, \mathcal{G}^*)$  with probability kernel characterized by

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

where  $B_1, \dots, B_{n+1} \in \mathcal{G}$  are measurable. Note that for measurable  $f : \mathcal{S}^* \rightarrow \mathbb{R}$  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 the  $\mathcal{G}^*/\mathfrak{B}(\mathbb{R})$ -measurable function  $u : \mathcal{S}^* \rightarrow \mathbb{R}$  by

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

Notice that the value of  $u(x_{1:n})$  does not depend on  $x_n$ . Let  $\mathbb{P}_{x_{1:n}}$  be the probability measure carrying  $(S^n)_{n=1}^{\infty}$  for which  $\mathbb{P}_{x_{1:n}}(S^n = x_{1:n}) = 1$ . As usual, let  $\mathbb{E}_{x_{1:n}}$  be the expectation with respect to  $\mathbb{P}_{x_{1:n}}$ . Now let  $U_t = u(S^t)$  and define

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

The definitions ensure that for any  $x_{1:n} \in \mathcal{S}^*$ ,  $x \in \mathcal{S}$  and  $\gamma \in \mathbb{R}$ ,

$$v_{\gamma}(x) = \bar{v}_{\gamma}(x) \quad \text{and} \quad \bar{v}_{\gamma}(x_{1:n}) = u(x_{1:n}) + \alpha^n \bar{v}_{\gamma}(x_n). \quad (35.3)$$

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,  $U = \lim_{n \rightarrow \infty} U_n$  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  $u(xy) = r(x) - \gamma$ , and so by (35.3),

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

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) with  $x_{1:n} = x$ . Notice finally that by (35.3), for any  $x_{1:n} \in \mathcal{S}^*$ ,

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

which means that  $\tau = \min\{t \geq 2 : \alpha^{t-1} v_\gamma(S_t) = 0\} = \min\{t \geq 2 : g(S_t) \leq \gamma\}$ , where we used the fact  $v_\gamma(x) = 0 \Leftrightarrow G(x) \leq \gamma$ .

## Chapter 36 Thompson Sampling

**36.3** We need to show that

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

holds almost surely. For specificity, let  $r : \mathcal{E} \rightarrow [k]$  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  $\nu_t \sim Q_{t-1}(\cdot) = Q(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$ . We have

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

**36.5** We have

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

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

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

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  $\infty$  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

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

Summing,

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

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

**36.13** Let  $\pi$  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  $\tilde{X}_t$  to  $\pi$ . 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  $R_n^*([0, 1]^{n \times k}) \leq R_n^*(\{0, 1\}^{n \times k})$ . The other direction is obvious.

## Chapter 37 Partial Monitoring

**37.3** It suffices to show that  $y$  has no component in any direction  $z$  that is perpendicular to  $x$ . Let  $z$  be such a direction. Without loss of generality either  $z^\top \mathbf{1} = 0$  or  $z^\top \mathbf{1} = 1$ . Assume first that  $z^\top \mathbf{1} = 0$ . Take some  $u \in \ker'(x)$ . Then,  $u + z \in \ker'(x)$  also holds. Since  $\ker'(x) \subset \ker'(y)$ ,  $z^\top y = (u + z)^\top y - u^\top y = 0$ . Assume now that  $z^\top \mathbf{1} = 1$ . Then  $z \in \ker'(x) \subset \ker'(y)$  and hence  $z^\top y = 0$ .

**37.10** An argument that *almost* works is to choose  $\omega \in \text{ri}(C_b)$  arbitrarily and find the first  $\nu$  along the chord connecting  $\omega$  and  $\lambda$  for which  $\nu \in C_b \cap C_c$  for some cell  $c$ . The minor problem is that  $C_b \cap C_c \neq \emptyset$  does not imply that  $b$  and  $c$  are neighbours, because there could be a third non-duplicate Pareto optimal action  $d$  with  $\nu \in C_b \cap C_c \cap C_d$ . This issue is resolved by making an ugly dimension argument. Let  $\mathcal{D}$  be the set of  $\omega \in \mathcal{P}_{d-1}$  for which at least three non-duplicate Pareto optimal cells intersect, which has dimension at most  $\dim(\mathcal{D}) \leq d - 3$ . Since  $b$  is Pareto optimal,  $\dim(\text{ri}(C_b)) = d - 1$ . Meanwhile, the dimension of those  $\omega \in \text{ri}(C_b)$  such that  $[\omega, \lambda] \cap \mathcal{D} \neq \emptyset$  has dimension at most  $d - 2$ . Hence, there exists an  $\omega \in \text{ri}(C_b)$  such that  $[\omega, \lambda] \cap \mathcal{D} = \emptyset$  and for this choice the initial argument works.

### 37.12

(a) Assume without loss of generality that  $\Sigma = [m]$ . Given an action  $c \in [k]$ , let  $S_c$  be as in the

proof of Theorem 37.12 and  $S \in \mathbb{R}^{km \times d}$  be the matrix obtained by stacking  $(S_c)_{c=1}^k$ :

$$S = \begin{pmatrix} S_1 \\ \vdots \\ S_k \end{pmatrix}.$$

As in the proof of Theorem 37.12, by the definition of global observability, for any pair of neighbouring actions  $a, b$ , it holds that  $\ell_a - \ell_b \in \text{im}(S^\top)$  and hence

$$\ell_a - \ell_b = S^\top U(\ell_a - \ell_b),$$

where  $U$  is the Moore–Penrose pseudo-inverse of  $S^\top$ . Then,

$$\|U(\ell_a - \ell_b)\|_\infty \leq \|U(\ell_a - \ell_b)\|_2 \leq \|U\|_2 \|\ell_a - \ell_b\|_2 \leq d^{1/2} \|U\|_2.$$

The largest singular value of  $U$  is the square root of the reciprocal of the smallest non-zero eigenvalue of  $S^\top S \in \{0, \dots, k\}^{d \times d}$ . Let  $(\lambda_i)_{i=1}^p$  be the non-zero eigenvalues of  $S^\top S$ , in decreasing order. Recall that for square matrix  $A$ , the product of its non-zero eigenvalues is a coefficient of the characteristic polynomial. Since  $S^\top S$  has entries in  $\{0, \dots, k\}$ , the characteristic equation has integer coefficients. Since  $S^\top S$  is positive definite, its non-zero eigenvalues are all positive and it follows that  $\prod_{i=1}^p \lambda_i \geq 1$ . If  $p = 1$ , then we are done. Suppose that  $p > 1$ . By the arithmetic–geometric mean inequality,

$$\frac{1}{\lambda_p} \leq \prod_{i=1}^{p-1} \lambda_i \leq \left( \frac{\text{trace}(S^\top S)}{p-1} \right)^{p-1} \leq \left( \frac{dk}{p-1} \right)^{p-1} \leq k^d.$$

Hence,  $\|U\|_2 \leq k^{d/2}$  and the result follows.

(b) Repeat the argument above, but restrict  $S$  to stacking  $(S_c)_{c \in \mathcal{N}_{ab}}$ .

(c) For non-degenerate locally observable games,  $\mathcal{N}_e = \{a, b\}$  for all  $e = (a, b) \in E$ . Given a Pareto optimal action  $a$ , let  $V_a = \{(a, \Phi_{ai}) : i \in [d]\}$ . Let  $a, b$  be neighbouring actions and  $f \in \mathcal{E}_{ab}^{\text{loc}}$ , which exists by the assumption that the game is locally observable. Define  $C = \{((a, \Phi_{ai}), (b, \Phi_{bi})) : i \in [d]\} \subset V_a \times V_b$ , which makes  $(V_a \cup V_b, C)$  a bipartite graph. We define a new function  $g : [k] \times \Sigma \rightarrow \mathbb{R}$ . First, let  $g(c, \sigma) = 0$  whenever  $\sigma \notin \{\Phi_{ci} : i \in [d]\}$  or  $c \notin \{a, b\}$ . Next, let  $(V'_a \cup V'_b, C')$  be a connected component. Then, by the conditions on  $f$ ,

$$\max_{n', n'' \in V'_a \cup V'_b} |f(n') - f(n'')| \leq s \doteq 2(m-1) + 1.$$

Letting  $c$  be the midpoint of the interval that the values of  $f$  restricted to  $V'_a \cup V'_b$  fall into,



define

$$f'(n) = \begin{cases} f(n) - c, & \text{if } n \in V'_a; \\ f(n) + c, & \text{if } n \in V'_b. \end{cases}$$

Clearly we still have  $f' \in \mathcal{E}_{ab}^{\text{loc}}$  and  $f'|_{V'_a \cup V'_b}$  takes values in  $[-s/2, s/2] = [-m, m]$ . Now repeat the procedure with  $f'$  and the next connected component of the graph until all connected components are processed. Let the resulting function be  $g$ . Then,  $g \in \mathcal{E}_{ab}^{\text{loc}}$  and also  $\|g\|_\infty \leq m$ .

**37.13** Let  $\Sigma = \{\clubsuit, \heartsuit\}$  and  $\ell_1 = (1, 0, 0, 0, \dots, 0, 0)$  and  $\ell_2 = (0, 1, 1, 1, \dots, 1, 1)$ . For all other actions  $a > 2$  let  $\ell_a = \mathbf{1}$ . Hence  $\ell_1 - \ell_2 = (1, -1, -1, -1, \dots, -1, -1)$ . Let  $d = 2k - 1$ . Let  $\Phi \in \Sigma^{k \times d}$  be the matrix such that the first row is  $\clubsuit$  and  $\heartsuit$ , alternating and for  $a > 1$ , let

$$\Phi_{ai} = \begin{cases} \clubsuit & \text{if } i \leq 2(a-1) \\ \heartsuit & \text{if } i = 1 + 2(a-1) \\ \clubsuit & \text{otherwise and } i \text{ is odd} \\ \heartsuit & \text{otherwise and } i \text{ is even.} \end{cases}$$

For example, when  $k = 4$ , then

$$\Phi = \begin{pmatrix} \clubsuit & \heartsuit & \clubsuit & \heartsuit & \clubsuit & \heartsuit & \clubsuit \\ \clubsuit & \clubsuit & \heartsuit & \heartsuit & \clubsuit & \heartsuit & \clubsuit \\ \clubsuit & \clubsuit & \clubsuit & \clubsuit & \heartsuit & \heartsuit & \clubsuit \\ \clubsuit & \clubsuit & \clubsuit & \clubsuit & \clubsuit & \clubsuit & \heartsuit \end{pmatrix}.$$

Suppose that  $f \in \mathcal{E}_{12}^{\text{glo}}$ . Then,

$$2 = \ell_{11} - \ell_{21} + \ell_{22} - \ell_{21} = \sum_{a=1}^k f(a, \Phi_{a1}) - \sum_{a=1}^k f(a, \Phi_{a2}) = f(1, \clubsuit) - f(1, \heartsuit).$$

Furthermore, for  $a > 1$  and  $i = 2a$ ,

$$\sum_{b=1}^k f(b, \Phi_{bi}) - \sum_{b=1}^k f(b, \Phi_{b,i+1}) = f(a, \clubsuit) - f(a, \heartsuit) + \sum_{b < a} (f(b, \heartsuit) - f(b, \clubsuit))$$

Hence,

$$f(a, \clubsuit) - f(a, \heartsuit) = \sum_{b < a} (f(b, \clubsuit) - f(b, \heartsuit)).$$

By induction it follows that  $f(a, \clubsuit) - f(a, \heartsuit) = 2^{a-1}$  for all  $a > 1$ . Finally, note that 1 and 2 are non-duplicate and Pareto optimal. Since all other actions are degenerate, these actions are neighbours. The game is globally observable because all columns have distinct patterns.

**37.14** For bandit games with  $\Phi = \mathcal{L}$ , let  $p = q$  and

$$f(a, \sigma)_b = \sigma \mathbb{I}\{a = b\}.$$

We have  $\sum_{a=1}^k f(a, \Phi_{ai})_b = \sum_{a=1}^k f(a, \mathcal{L}_{ai})_b = \sum_{a=1}^k \mathcal{L}_{ai} \mathbb{I}\{a = b\} = \mathcal{L}_{bi}$  and thus  $f \in \mathcal{E}^{\text{vec}}$ . Then, using that  $\exp(-x) + x - 1 \leq x^2/2$  for  $x \geq 0$ ,

$$\begin{aligned} \text{opt}_q^*(\eta) &\leq \max_{i \in [d]} \left( \frac{(p - q)^\top \mathcal{L} e_i}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^k p_a \Psi_q \left( \frac{\eta f(a, \Phi_{ai})}{p_a} \right) \right) \\ &\leq \max_{i \in [d]} \left( \frac{1}{2} \sum_{a=1}^k p_a \sum_{b=1}^k \frac{q_b \mathcal{L}_{ai}^2 \mathbb{I}\{a = b\}}{p_a^2} \right) \\ &= \frac{k}{2}. \end{aligned}$$

The full information game is similar. As before, let  $p = q$ , but choose  $f(a, \sigma)_b = p_a \mathcal{L}_{b\sigma}$ . We have  $\sum_{a=1}^k f(a, \Phi_{ai})_b = \sum_{a=1}^k f(a, i)_b = \sum_{a=1}^k p_a \mathcal{L}_{bi} = \mathcal{L}_{bi}$  and thus  $f \in \mathcal{E}^{\text{vec}}$ . Then, again using that  $\exp(-x) + x - 1 \leq x^2/2$  for  $x \geq 0$ ,

$$\text{opt}_q^*(\eta) \leq \max_{i \in [d]} \frac{1}{2} \sum_{a=1}^k p_a \sum_{b=1}^k q_b \mathcal{L}_{bi}^2 = \frac{1}{2}.$$

## 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  $\pi' = (\pi'_1, \pi'_2, \dots)$  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  $n \rightarrow \infty$ , 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 maximised for small indices until  $|U|$  is exhausted. In particular, if  $(n_k)_{0 \leq k \leq m}$  denotes the optimal solution ( $n_k = 0$  for  $k > m$ ) then  $n_0 = A^0, \dots, n_{m-1} = A^k$ ,  $0 \leq n_m = |U| - \sum_{k=0}^{m-1} A^k (= |U| - \frac{A^m - 1}{A - 1}) < A^m$ . Hence,  $|U| < A^m + \frac{A^m - 1}{A - 1} \leq 2A^m$ , implying that  $m \geq \log_A(|U|/2)$ . Thus,

$$\begin{aligned} d^*(\mu_0, U) &= \sum_{k=0}^{m-1} k A^k + m n_m \\ &= m|U| + \sum_{k=0}^{m-1} (k - m) A^k \\ &\stackrel{(a)}{=} m|U| + \frac{m}{A - 1} - \frac{A^{m+1} - A}{(A - 1)^2} \\ &\geq |U| \left( m - 1 - \frac{1}{A - 1} \right) \\ &\geq |U| (\log_A(|U|) - 3), \end{aligned}$$

where step (a) follows since  $|U| < \frac{A^m - 1}{A - 1}$ . Choosing  $U = \mathcal{S}$ , we see that the expected minimum time to reach a random state in  $\mathcal{S}$  is lower bounded by  $\log_A(S) - 3$ . 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  $\lim_{n \rightarrow \infty} B_n = B$  and  $\lim_{n \rightarrow \infty} C_n = C$ . It suffices to show that  $B = C$ . From Part (b),  $B_m + \frac{1}{n_m}(P^{n_m} - I) = B_m P = P B_m$ . Taking the limit as  $m$  tends to infinity and using the fact that  $P^n$  is  $[0, 1]$ -valued we see that  $B = B P = P B$ . Similarly,  $C = C P = P C$  and it follows that  $B = B P^i = P^i B$  and  $C = C P^i = P^i C$  hold for any  $i \geq 0$ . Hence  $B = B C_m = C_m B$  and  $C = C B_m = B_m C$  for any  $m \geq 1$ . Taking limit as  $m$  tends to infinity shows that  $B = B C = C B$  and  $C = C B = B C$ , which together imply that  $B = C$ .
- (d) We have already seen in the proof of Part (c) that  $P^* = P^* P = P P^*$ . 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  $i = 1, \dots, n$  and dividing by  $n$  and using the fact that  $B^i = P^i - P^*$  for all  $i \geq 1$ ,

$$I - \frac{1}{n} \sum_{i=1}^n P^i + P^* = (I - B) H_n, \quad (38.1)$$

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  $H_\infty$  we find that  $I = (I - B) H_\infty$  and thus  $I - B$  is invertible and its inverse  $H$  is equal to  $H_\infty$ .

- (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  $H - \lim_{n \rightarrow \infty} U_n = P^*$ . From the definition of  $U$  we have  $U = \lim_{n \rightarrow \infty} U_n$ .

- (g) This follows immediately because

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{i-1} P^k (r - \rho) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{i-1} (P^k - P^*) r = U r.$$

- (h) One way to prove this is to note that by the previous part  $v = U r = (H - P^*) r$ , hence  $r = H^{-1}(v + \rho)$ . Now,  $H^{-1}(v + \rho) = (I - P + P^*)(v + \rho) = v - P v + P^* v + \rho - P \rho + P^* \rho = v - P v + \rho$ , where we used that  $P^* v = P^* U r$  and  $P^* U = P^*(H - P^*) = P^* H - P^* = (P^* - P^* H^{-1}) H = 0$  and that  $P \rho = P P^* r = P^* r = P^* \rho = P^* P^* r$ .

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  $\lim_{n \rightarrow \infty} \bar{v}_n = v$ . Then,

$v_{k+1} = Pv_k + (r - \rho)$ . Taking the average of these over  $k = 1, \dots, n$  we get

$$\frac{1}{n} ((n+1)\bar{v}_{n+1} - v_1) = P\bar{v}_n + (r - \rho).$$

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  $|\max_x f(x) - \max_y g(y)| \leq \max_x |f(x) - g(x)|$ . Then for  $v, w \in \mathbb{R}^S$ ,

$$\begin{aligned} \|T_\gamma v - T_\gamma w\|_\infty &\leq \max_{s \in \mathcal{S}} \max_{a \in \mathcal{A}} \gamma |\langle P_a(s), v - w \rangle| \\ &\leq \max_{s \in \mathcal{S}} \max_{a \in \mathcal{A}} \gamma \|P_a(s)\|_1 \|v - w\|_\infty \\ &= \gamma \|v - w\|_\infty. \end{aligned}$$

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  $v = r_\pi + \gamma P_\pi v$ . Solving for  $v$  completes the result.

(e) If  $\pi$  is a memoryless policy, it is trivial to see that  $v_\gamma^\pi = r_\pi + \gamma P_\pi v_\gamma^\pi$ . Let  $\pi^*$  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  $\pi$ ,  $v_\gamma^\pi \leq v$ . If  $\pi_t$  is the memoryless policy used in time step  $t$  when following  $\pi$ ,  $v_\gamma^\pi = \sum_{t=1}^{\infty} \gamma^{t-1} P^{(t-1)} r_{\pi_t}$ , where  $P^{(0)} = I$  and for  $t \geq 1$ ,  $P^{(t)} = P_{\pi_1} \dots P_{\pi_t}$ . For  $n \geq 1$ , let  $v_{\gamma,n}^\pi = \sum_{t=1}^n \gamma^{t-1} P^{(t-1)} r_{\pi_t}$ . It is easy to see that  $v_{\gamma,1}^\pi = r_{\pi_1} \leq T\mathbf{0}$ . Assume that for some  $n \geq 1$ ,

$$\sup_{\pi \text{ Markov}} v_{\gamma,n}^\pi \leq T^n \mathbf{0}, \quad (38.2)$$

Notice that  $f \leq g$  implies  $Tf \leq Tg$ . Hence,  $Tv_{\gamma,n}^\pi \leq T^{n+1} \mathbf{0}$ . Further,  $v_{\gamma,n+1}^\pi = r_{\pi_0} + \gamma P_{\pi_0} v_{\gamma,n}^{\pi'}$ , where  $\pi'$  is the Markov policy obtained from  $\pi$  by discarding  $\pi_0$  (that is, if  $\pi = (\pi_0, \pi_1, \pi_2, \dots)$ ,  $\pi' = (\pi_1, \pi_2, \dots)$ ). This shows that (38.2) holds for all  $n \geq 1$ . Letting  $n \rightarrow \infty$ , the right-hand side converges to  $v$ , while the left-hand side converges to  $v_\gamma^\pi$ . Hence,  $v_\gamma^\pi \leq v$ .

### 38.9

(a) Let  $0 \leq \gamma < 1$ . 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  $P_\gamma^*$  is right stochastic. By the compactness of the space of right stochastic matrices,  $(P_\gamma^*)_\gamma$  has at least one cluster point  $A$  as  $\gamma \rightarrow 1-$ . 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  $\lim_{\gamma \rightarrow 1-} P_\gamma^* = P^*$ .

(b) Since  $I - P + P^*$  is invertible and  $PP^* = P^* = P^*P^*$ , the required claim is equivalent to

$$\lim_{\gamma \rightarrow 1-} (I - P + P^*) \left( \frac{P_\gamma^* - P^*}{1 - \gamma} \right) = I - P^*.$$

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  $\gamma$  tends to one from below and using Part (a).

### 38.10

- (a) Let  $(\gamma_n)$  be an arbitrary increasing sequence with  $\gamma_n < 1$  for all  $n$  and  $\lim_{n \rightarrow \infty} \gamma_n = 1$ . Let  $v_n$  be the fixed point of  $T_{\gamma_n}$  and  $\pi_n$  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  $n_1 < n_2 < \dots$  and policy  $\pi$  such that  $\pi_{n_k} = \pi$  for all  $k$ .
- (b) For arbitrary  $\pi$  and  $0 \leq \gamma < 1$ , let  $v_\gamma^\pi$  be the value function of policy  $\pi$  in the  $\gamma$ -discounted MDP,  $v^\pi$  the value function of  $\pi$  and  $\rho^\pi$  its gain. Let  $U_\pi$  be the deviation matrix underlying  $P_\pi$ . Define  $f_\gamma^\pi$  by

$$v_\gamma^\pi = \frac{\rho^\pi}{1 - \gamma} + v_\pi + f_\gamma^\pi. \quad (38.3)$$

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  $\|f_\gamma^\pi\|_\infty \rightarrow 0$  as  $\gamma \rightarrow 1$ .

Fix now  $\pi$  to be the policy whose existence is guaranteed by the previous part. By Part (e) of Exercise 38.8,  $\pi$  is  $\gamma_n$ -discount optimal for all  $n \geq 1$ . Suppose that  $\rho^\pi$  is not a constant. In any case,  $\rho^\pi$  is piecewise constant on the recurrent classes of the Markov chain with transition probabilities  $P^\pi$ . 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  $\rho^\pi$  is the largest and take a policy  $\pi'$  that is identical to  $\pi$  over  $R$ , while  $\pi'$  is set up such that it gets to  $R$  with probability one. Such a  $\pi'$  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  $\gamma \geq \gamma^*$ ,

$$v_{\gamma}^{\pi'}(s) > v_{\gamma}^{\pi}(s). \quad (38.4)$$

If this was true and  $n$  is large enough so that  $\gamma_n \geq \gamma^*$  then, since  $\pi$  is  $\gamma_n$ -discount optimal,  $v_{\gamma_n}^{\pi}(s) \geq v_{\gamma_n}^{\pi'}(s) > v_{\gamma_n}^{\pi}(s)$ , which is a contradiction.

Hence, it remains to show (38.4). By the construction of  $\pi'$ ,  $\rho^{\pi'}(s) = \rho_*^{\pi} > \rho^{\pi}(s)$ . From (38.3),

$$v_{\gamma}^{\pi'}(s) = \frac{\rho^{\pi'}(s)}{1 - \gamma} + v_{\pi'}(s) + f_{\gamma}^{\pi'}(s) > \frac{\rho^{\pi}(s)}{1 - \gamma} + v_{\pi}(s) + f_{\gamma}^{\pi}(s) = v_{\gamma}^{\pi}(s),$$

where the inequality follows by taking  $\gamma \geq \gamma^*$  for some  $\gamma^* \in (0, 1)$ . The existence of any appropriate  $\gamma^*$  follows because  $f_{\gamma}^{\pi'}(s), f_{\gamma}^{\pi}(s) \rightarrow 0$ , while  $1/(1 - \gamma) \rightarrow \infty$  as  $\gamma \rightarrow 1$ .

- (c) Since  $v$  is the value function of  $\pi$  and  $\rho$  is its gain, by (38.4) we have  $\rho \mathbf{1} + v = r_{\pi} + P_{\pi}v$ . Let  $\pi'$  be an arbitrary stationary policy and  $v_n$  be as before. Let  $f_n = f_{\gamma_n}$  where  $f_{\gamma}$  is defined by (38.3). Note that  $\|f_n\|_{\infty} \rightarrow 0$  as  $n \rightarrow \infty$ . 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

$$\rho \mathbf{1} + v = r_{\pi} + P_{\pi} v. \quad (38.5)$$

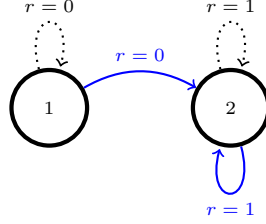
Since  $\pi'$  was arbitrary,  $\rho + v(s) \geq \max_a r_a(s) + \langle P_a(s), v \rangle$  holds for all  $s \in \mathcal{S}$ . This combined with (38.5) shows that the pair  $(\rho, v)$  satisfies the Bellman optimality equation as required.

**38.11** Clearly the optimal policy is to take action STAY in any state and this policy has gain  $\rho^* = 0$ . Pick any solution  $(\rho, v)$  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  $T : \mathbb{R}^S \rightarrow \mathbb{R}^S$  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  $v \in \mathbb{R}^S$  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  $\pi$  it holds that

$$v(s_2) \leq v(s_1) + (\rho^* - \min_{s,a} r_a(s)) \mathbb{E}^\pi[\tau_{s_2} | S_1 = s_1].$$

The remainder of the proof is devoted to proving this result for fixed  $s_1, s_2 \in \mathcal{S}$  and memoryless policy  $\pi$ . Abbreviate  $\tau = \tau_{s_2}$  and let  $\mathbb{E}[\cdot]$  denote the expectation with respect to the measure induced by the interaction of  $\pi$  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  $\bar{T} : \mathbb{R}^S \rightarrow \mathbb{R}^S$  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  $r_\pi(s) - \rho^* \geq \min_{s,a} r_a(s) - \rho^*$  and  $Tv = v$  it follows that  $(\bar{T}v)(s) \leq (Tv)(s) = v(s)$ . Notice that for  $u \leq w$  it holds that  $\bar{T}u \leq \bar{T}w$ . Then by induction we have  $\bar{T}^n v \leq v$  for all  $n \in \mathbb{N}^+$ . By unrolling the recurrence we have

$$v(s_1) \geq (\bar{T}^n v)(s_1) = \mathbb{E} \left[ -(\rho^* - \min_{s,a} r_a(s))(n \wedge \tau) + v(S_{\tau \wedge n}) \right].$$

Taking the limit as  $n$  tends to infinity shows that  $v(s_1) \geq v(s_2) - (\rho^* - \min_{s,a} r_a(s)) \mathbb{E}[\tau]$ , which completes the result.

### 38.14

- (a) It is clear that Algorithm 27 returns TRUE if and only if  $(\rho, v)$  is feasible for Eq. (38.6). Note that feasibility can be written in the compact form  $\rho \mathbf{1} v \geq Tv$ . It remains to show that when  $(\rho, v)$  is not feasible then  $u = (1, e_s - P_{a_s^*}(s))$  is such that for any  $(\rho', v')$  feasible,  $\langle (\rho', v'), u \rangle > \langle (\rho, v), u \rangle$ .



For this, we have  $\langle (\rho', v'), u \rangle = \rho' + v'(s) - \langle P_{a_s^*}(s), v' \rangle \geq r_{a_s^*}(s)$ , where the inequality used that  $(\rho', v')$  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  $-\varepsilon \leq v(\tilde{s}) \leq \varepsilon$ . Then add the  $\varepsilon$  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  $\langle a_k, x \rangle \geq b_k$  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  $\phi$  or  $\phi_k$  that did not return true.

### 38.16

- (a) Let  $\pi = (\pi_1, \pi_2, \dots)$  be an arbitrary Markov policy where  $\pi_t$  is the policy followed in time step  $t$ . Using the notation and techniques from the proof of Theorem 38.2,

$$\begin{aligned} P_\pi^{(t-1)} r_{\pi_t} &= P_\pi^{(t-1)} (r_{\pi_t} + P_{\pi_t} v - P_{\pi_t} v) \leq P_\pi^{(t-1)} (r_{\tilde{\pi}} + P_{\tilde{\pi}} v - P_{\pi_t} v) \\ &\leq P_\pi^{(t-1)} ((\rho + \varepsilon) \mathbf{1} + v - P_{\pi_t} v) = (\rho + \varepsilon) \mathbf{1} + P_\pi^{(t-1)} v - P_\pi^{(t)} v. \end{aligned}$$

Taking the average and then the limit shows that  $\bar{\rho}^\pi(s) \leq \rho + \varepsilon$  for all  $s \in \mathcal{S}$ . By the claim in Exercise 38.2,  $\rho^* \leq \rho + \varepsilon$ .

- (b) We have  $r_{\tilde{\pi}}(s) + \langle P_{\tilde{\pi}}(s), v \rangle \geq \max_a r_a(s) + \langle P_a(s), v \rangle - \varepsilon' \geq \rho + v(s) - (\varepsilon + \varepsilon')$ . Therefore,

$$P_{\tilde{\pi}}^{t-1} r_{\tilde{\pi}} = P_{\tilde{\pi}}^{t-1} (r_{\tilde{\pi}} + P_{\tilde{\pi}} v - P_{\tilde{\pi}} v) \geq P_{\tilde{\pi}}^{t-1} ((\rho - (\varepsilon + \varepsilon')) \mathbf{1} + v - P_{\tilde{\pi}} v).$$

Taking the average and the limit again shows that  $\rho^{\tilde{\pi}}(s) \geq \rho - (\varepsilon + \varepsilon')$ . The claim follows from combining this with the previous result.

- (c) Let  $\varepsilon = v + \rho \mathbf{1} - T v$ , which by the first constraint satisfies  $\varepsilon \geq 0$ . Let  $\pi^*$  be an optimal policy satisfying the requirements of the theorem statement and  $\pi$  be the greedy policy with respect to  $v$ . Then

$$P_{\pi^*}^t r_{\pi^*} \leq P_{\pi^*}^t (r_\pi + P_\pi v - P_{\pi^*} v) = P_{\pi^*}^t (\rho \mathbf{1} + v - \varepsilon - P_{\pi^*} v).$$

Hence  $\rho^* \mathbf{1} = \rho^{\pi^*} \mathbf{1} \leq \rho \mathbf{1} - P_{\pi^*}^* \varepsilon$ , which means that  $P_{\pi^*}^* \varepsilon \leq \delta \mathbf{1}$ . By the definition of  $\tilde{s}$  there exists a state  $s$  with

$$\delta \geq (P_{\pi^*} \varepsilon)(s) = \sum_{s' \in \mathcal{S}} P_{\pi^*}(s, s') \varepsilon(s') \geq P_{\pi^*}(s, \tilde{s}) \varepsilon(\tilde{s})$$

and so  $\varepsilon(\tilde{s}) \leq \delta |\mathcal{S}|$ . Notice that  $\tilde{v} = v - \varepsilon + \varepsilon(\tilde{s}) \mathbf{1}$  also satisfies the constraints in Eq. (38.7) and hence

$$\langle v - \varepsilon + \varepsilon(\tilde{s}) \mathbf{1}, \mathbf{1} \rangle \geq \langle v, \mathbf{1} \rangle,$$

which implies that  $\langle \varepsilon, \mathbf{1} \rangle \leq |\mathcal{S}| \varepsilon(\tilde{s}) \leq |\mathcal{S}|^2 \delta$ . Hence  $(\rho, v)$  approximately satisfies the Bellman optimality equation.

**38.17** Define operator  $T : \mathbb{R}^{\mathcal{S}} \rightarrow \mathbb{R}^{\mathcal{S}}$  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  $\pi$  that always chooses the dashed action.
- (b) The optimistic rewards are given by

$$\begin{aligned} \tilde{r}_{k, \text{STAY}}(s) &= \frac{1}{2} + \sqrt{\frac{L}{2(1 \vee T_{k-1}(s, \text{STAY}))}} \\ \tilde{r}_{k, \text{GO}}(s) &= \sqrt{\frac{L}{2(1 \vee T_{k-1}(s, \text{GO}))}}. \end{aligned}$$

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  $T_{k-1}(S_t, \text{STAY}) > T_{k-1}(S'_t, \text{STAY})$  and  $\tilde{r}_{k, \text{STAY}}(S_t) < 1$ . Then  $\tilde{r}_{k, \text{STAY}}(S'_t) > \tilde{r}_{k, \text{STAY}}(S_t)$ . 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 maximise 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,

$$\begin{aligned} \mathbb{P}(F) &\leq \mathbb{P}\left(\exists u \in \mathbb{N}, s, a \in \mathcal{S} \times \mathcal{A} : \|P - \hat{P}_{u,a}(s)\|_1 \geq \sqrt{\frac{2S \log(4SAu(u+1)/\delta)}{u}}\right) \\ &\leq \sum_{s,a \in \mathcal{S} \times \mathcal{A}} \sum_{u=1}^{\infty} \mathbb{P}\left(\|P - \hat{P}_{u,a}(s)\|_1 \geq \sqrt{\frac{2S \log(4SAu(u+1)/\delta)}{u}}\right) \\ &\leq \sum_{s,a \in \mathcal{S} \times \mathcal{A}} \sum_{u=1}^{\infty} \frac{\delta}{2SAu(u+1)} = \frac{\delta}{2}. \end{aligned}$$

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  $A_0 = A_1 = \dots = A_{m-1} = 1$ . 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

$$\begin{aligned} \tilde{R}_k &= \sum_{t \in E_k} (-v_k(S_t) + \langle P_{k,A_t}(S_t), v_k \rangle + \tilde{r}_{k,A_t}(S_t) - r_{A_t}(S_t)) \\ &\leq \sum_{t \in E_k} (\langle P_{A_t}, v_k \rangle - v_k(S_t)) + \frac{D}{2} \sum_{t \in E_k} \|P_{k,A_t}(S_t) - P_{A_t}(S_t)\|_1 \\ &\quad + \sum_{t \in E_k} \sqrt{\frac{2L}{1 \vee T_{\tau_k-1}(S_t, A_t)}}. \end{aligned}$$

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,

$$\begin{aligned} \sum_{k=1}^K \sum_{t \in E_k} \sqrt{\frac{2L}{1 \vee T_{\tau_k-1}(S_t, A_t)}} &= \sqrt{2L} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^K \frac{T_{(k)}(s, a)}{\sqrt{1 \vee T_{\tau_k-1}(s, a)}} \\ &\leq (\sqrt{2} + 1) \sqrt{2LSAn}. \end{aligned}$$

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

$$\hat{R}_n \leq CD(M)S \sqrt{nA \log \left( \frac{nSA}{\delta} \right)}.$$

### 38.24

- (a) An easy calculation shows that the depth  $d$  of the tree is bounded by  $2 + \log_A S$ , 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 maximised 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  $\pi$  be a policy that traverses the tree to a decision state with  $\varepsilon(s, a) = 0$ . We will show that for this policy

$$\mathbb{E}[\tau \mid S_1 = s_g] \leq D.$$

Let  $X_1, X_2, \dots$  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  $i$ th 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

$$\mathbb{E}[X_i] = 1/\delta \qquad \mathbb{E}[M] = 2.$$

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  $\tau$  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) We need to prove that  $R_{nj} \geq c_3 \Delta D \mathbb{E}_j[T_\sigma - T_j]$  where  $R_{nj}$  is the expected regret of  $\pi$  in MDP  $M_j$  over  $n$  rounds and  $c_3 > 0$  is a universal constant. The idea is to write the total reward incurred using episodes punctuated by visits to  $s_0$  and note that the expected lengths of these

episodes are the same regardless of the policy used.

For the formal argument, we start by rewriting the expected regret  $R_{nj}$  in a more suitable form. For this we introduce episodes indexed by  $k \in [n]$ . The  $k$ th episode starts at the time  $\tau_k$  when  $s_0$  is visited the  $k$ th time (in particular,  $\tau_1 = 0$ ). In the  $k$ th episode, a leaf is reached in time step  $t = \tau_k + d$ . Let  $E_k$  be the indicator that the the state-action pair of this time step is *not* the  $j$ th state-action in  $\mathcal{L} = \{(s_1, a_1), \dots, (s_p, a_p)\}$ :  $E_k = \mathbb{I}\{(S_{\tau_k+d}, A_{\tau_k+d}) = (s_j, a_j)\}$ . Note that this is an indicator of a “regret-inducing choice”.

In time steps  $\mathcal{T}_k = \{\tau_k + d + 1, \dots, \tau_{k+1}\}$ , the state is one of  $s_g$  and  $s_b$ . Let  $H_k = |\mathcal{T}_k \cap [n]|$  be the number of time steps in  $\mathcal{T}_k$  that happen before round  $n$  is over. For clarity, we add subindex  $\pi$  to  $\mathbb{E}_j$  and  $\mathbb{P}_j$  to make it explicit that these depend on  $\pi$ . Further, in MDP  $M_j$ , let the values of  $\varepsilon(s, a)$  be denoted by  $\varepsilon_j(s, a)$ .

By construction, the total expected reward incurred up to time step  $n$  by policy  $\pi$  is

$$\begin{aligned}
V_j^\pi &:= \sum_{k=1}^n \mathbb{E}_{j,\pi}[H_k \mathbb{I}\{S_{\tau_k+d+1} = s_g\}] \\
&= \sum_{k=1}^n \sum_{(s,a) \in \mathcal{L}} \mathbb{E}_{j,\pi}[H_k \mathbb{I}\{S_{\tau_k+d+1} = s_g\} | S_{\tau_k+d} = s, A_{\tau_k+d} = a] \mathbb{P}_{j,\pi}(S_{\tau_k+d} = s, A_{\tau_k+d} = a) \\
&\quad \text{(by the law of total probability)} \\
&= \sum_{k=1}^n \sum_{(s,a) \in \mathcal{L}} \mathbb{E}_{j,\pi}[H_k] \mathbb{P}_{j,\pi}(S_{\tau_k+d+1} = s_g | S_{\tau_k+d} = s, A_{\tau_k+d} = a) \mathbb{P}_{j,\pi}(S_{\tau_k+d} = s, A_{\tau_k+d} = a) \\
&\quad \text{(conditioning, Markov property)} \\
&= \sum_{k=1}^n \sum_{(s,a) \in \mathcal{L}} \mathbb{E}_{j,\pi}[H_k] \left( \frac{1}{2} + \varepsilon_j(s, a) \right) \mathbb{P}_{j,\pi}(S_{\tau_k+d} = s, A_{\tau_k+d} = a) \quad (M_j \text{ definition}) \\
&= \sum_{k=1}^n \sum_{p \neq j} \mathbb{E}_{j,\pi}[H_k] \left( \frac{1}{2} \right) \mathbb{P}_{j,\pi}(S_{\tau_k+d} = s_p, A_{\tau_k+d} = a_p) \\
&\quad + \left( \frac{1}{2} + \Delta \right) \sum_{k=1}^n \mathbb{E}_{j,\pi}[H_k] \mathbb{P}_{j,\pi}(S_{\tau_k+d} = s_j, A_{\tau_k+d} = a_j).
\end{aligned}$$

Now, note that  $\mathbb{E}_{j,\pi}[H_k] = \lambda_k$ , regardless of policy  $\pi$  and index  $j$ . If  $\pi_j^*$  is the optimal policy for MDP  $M_j$ ,  $\mathbb{P}_{j,\pi_j^*}(S_{\tau_k+d} = s_j, A_{\tau_k+d} = a_j) = 1$ . Let  $\rho_k^\pi = \mathbb{P}_{j,\pi}((S_{\tau_k+d}, A_{\tau_k+d}) \neq (s_j, a_j))$ . Hence,

$$\begin{aligned}
R_{nj} &= V_j^{\pi^*} - V_j^\pi \\
&= (1/2 + \Delta) \sum_{k=1}^n \lambda_k - \frac{1}{2} \sum_{k=1}^n \lambda_k \rho_k^\pi - \left( \frac{1}{2} + \Delta \right) \sum_{k=1}^n \lambda_k (1 - \rho_k^\pi) \\
&= \sum_{k=1}^n \lambda_k \left( (1/2 + \Delta) - \frac{1}{2} \rho_k^\pi - \left( \frac{1}{2} + \Delta \right) (1 - \rho_k^\pi) \right) \\
&= \Delta \sum_{k=1}^n \lambda_k \rho_k^\pi \geq \Delta \sum_{k=1}^{m-1} \lambda_k \rho_k^\pi \geq c \Delta D \sum_{k=1}^{m-1} \rho_k^\pi,
\end{aligned}$$

where  $m = \lceil n/D - 1 \rceil$  and the last inequality uses that  $\lambda_k \geq cD$  for  $k \in [m - 1]$  with some universal constant  $c > 0$ , the proof of which is left to the reader. Now, by definition,

$$\begin{aligned} \mathbb{E}_{j,\pi}[T_\sigma - T_j] &= \sum_{k=1}^n \mathbb{P}_{j,\pi}((S_{\tau_k+d}, A_{\tau_k+d}) \neq (s_j, a_j), \tau_k + d < \tau) \\ &\leq \sum_{k=1}^{m-1} \mathbb{P}_{j,\pi}((S_{\tau_k+d}, A_{\tau_k+d}) \neq (s_j, a_j)) \\ &= \sum_{k=1}^m \rho_k^\pi, \end{aligned}$$

where the second equality is because  $\tau = n \wedge \tau_m$  and thus for  $k \geq m$ ,  $\tau_k + d \geq \tau$ . Putting together the last two inequalities finishes the proof.

## Bibliography

- V. I. Bogachev. *Measure theory*, volume 2. Springer Science & Business Media, 2007. [64]
- O. Kallenberg. *Foundations of modern probability*. Springer-Verlag, 2002. [12, 55, 90]
- T. Lattimore and Cs. Szepesvári. Learning with good feature representations in bandits and in RL with a generative model. arXiv:1911.07676, 2019. [78]
- T. Needham. A visual explanation of Jensen’s inequality. *American Mathematical Monthly*, 100(8): 768–771, 1993. [62]
- G. Peskir and A. Shiryaev. *Optimal stopping and free-boundary problems*. Springer, 2006. [96]
- E. V. Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, pages 404–412, 1977. [17]