

# Reinforcement Learning Theory

Paulo Rauber

2024

## 1 Asymptotic analysis

Consider a function  $f : \mathbb{N} \rightarrow \mathbb{R}$ .

**Definition 1.1.** For every  $m \in \mathbb{N}$ ,  $\inf_{n \geq m} f(n)$  is the largest  $r \in [-\infty, \infty]$  such that  $r \leq f(n)$  for every  $n \geq m$ .

**Definition 1.2.** For every  $m \in \mathbb{N}$ ,  $\sup_{n \geq m} f(n)$  is the smallest  $r \in [-\infty, \infty]$  such that  $r \geq f(n)$  for every  $n \geq m$ .

**Definition 1.3.** The limit inferior  $\liminf_{n \rightarrow \infty} f(n)$  is defined by

$$\liminf_{n \rightarrow \infty} f(n) = \lim_{m \rightarrow \infty} \inf_{n \geq m} f(n).$$

Since the function  $g$  given by  $g(m) = \inf_{n \geq m} f(n)$  is non-decreasing, the limit exists in  $[-\infty, \infty]$ .

**Proposition 1.1.** If  $z < \liminf_{n \rightarrow \infty} f(n)$ , then  $z < f(n)$  for all sufficiently large  $n \in \mathbb{N}$ .

**Proposition 1.2.** If  $z > \liminf_{n \rightarrow \infty} f(n)$ , then  $z > f(n)$  for infinitely many  $n \in \mathbb{N}$ .

**Definition 1.4.** The limit superior  $\limsup_{n \rightarrow \infty} f(n)$  is defined by

$$\limsup_{n \rightarrow \infty} f(n) = \lim_{m \rightarrow \infty} \sup_{n \geq m} f(n).$$

Since the function  $g$  given by  $g(m) = \sup_{n \geq m} f(n)$  is non-increasing, the limit exists in  $[-\infty, \infty]$ .

**Proposition 1.3.** If  $z > \limsup_{n \rightarrow \infty} f(n)$ , then  $z > f(n)$  for all sufficiently large  $n \in \mathbb{N}$ .

**Proposition 1.4.** If  $z < \limsup_{n \rightarrow \infty} f(n)$ , then  $z < f(n)$  for infinitely many  $n \in \mathbb{N}$ .

**Proposition 1.5.** For every  $m \in \mathbb{N}$ , the infimum, limit inferior, limit superior, and supremum are related by

$$\inf_{n \geq m} f(n) \leq \liminf_{n \rightarrow \infty} f(n) \leq \limsup_{n \rightarrow \infty} f(n) \leq \sup_{n \geq m} f(n).$$

**Definition 1.5.** The function  $f$  is said to converge in  $[-\infty, \infty]$  if and only if

$$\liminf_{n \rightarrow \infty} f(n) = \limsup_{n \rightarrow \infty} f(n).$$

**Definition 1.6.** The set of asymptotically positive function  $\mathcal{F}$  is defined by

$$\mathcal{F} = \{f : \mathbb{N} \rightarrow \mathbb{R} \mid \text{there is an } m \in \mathbb{N} \text{ such that } f(n) > 0 \text{ for every } n \geq m\}.$$

**Definition 1.7.** For every  $f \in \mathcal{F}$  and  $g \in \mathcal{F}$ , let  $(f/g) \in \mathcal{F}$  be given by

$$(f/g)(n) = \begin{cases} f(n)/g(n), & \text{if } g(n) \neq 0, \\ 0, & \text{if } g(n) = 0. \end{cases}$$

For convenience, we often write  $(f/g)(n)$  as  $f(n)/g(n)$ , since  $(f/g)(n) = f(n)/g(n)$  for all sufficiently large  $n \in \mathbb{N}$ .

**Definition 1.8.** If  $g \in \mathcal{F}$ , then the following subsets of  $\mathcal{F}$  are defined:

$$\begin{aligned} o(g) &= \left\{ f \in \mathcal{F} \mid \limsup_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0 \right\}, \\ O(g) &= \left\{ f \in \mathcal{F} \mid \limsup_{n \rightarrow \infty} \frac{f(n)}{g(n)} < \infty \right\}, \\ \Omega(g) &= \left\{ f \in \mathcal{F} \mid \liminf_{n \rightarrow \infty} \frac{f(n)}{g(n)} > 0 \right\}, \\ \omega(g) &= \left\{ f \in \mathcal{F} \mid \liminf_{n \rightarrow \infty} \frac{f(n)}{g(n)} = \infty \right\}, \\ \Theta(g) &= O(g) \cap \Omega(g). \end{aligned}$$

Consider a real number  $a > 0$ .

**Example 1.1.** Since  $\lim_{n \rightarrow \infty} an/n^2 = \limsup_{n \rightarrow \infty} an/n^2 = \liminf_{n \rightarrow \infty} an/n^2 = 0$ :

- $(n \mapsto an) \in o(n \mapsto n^2)$ , often written as  $an \in o(n^2)$ .
- $(n \mapsto an) \in O(n \mapsto n^2)$ , often written as  $an \in O(n^2)$ .
- $(n \mapsto an) \notin \Omega(n \mapsto n^2)$ , often written as  $an \notin \Omega(n^2)$ .
- $(n \mapsto an) \notin \omega(n \mapsto n^2)$ , often written as  $an \notin \omega(n^2)$ .
- $(n \mapsto an) \notin \Theta(n \mapsto n^2)$ , often written as  $an \notin \Theta(n^2)$ .

**Example 1.2.** Since  $\lim_{n \rightarrow \infty} n^2/an = \limsup_{n \rightarrow \infty} n^2/an = \liminf_{n \rightarrow \infty} n^2/an = \infty$ :

- $(n \mapsto n^2) \notin o(n \mapsto an)$ , often written as  $n^2 \notin o(an)$ .
- $(n \mapsto n^2) \notin O(n \mapsto an)$ , often written as  $n^2 \notin O(an)$ .
- $(n \mapsto n^2) \in \Omega(n \mapsto an)$ , often written as  $n^2 \in \Omega(an)$ .
- $(n \mapsto n^2) \in \omega(n \mapsto an)$ , often written as  $n^2 \in \omega(an)$ .
- $(n \mapsto n^2) \notin \Theta(n \mapsto an)$ , often written as  $n^2 \notin \Theta(an)$ .

**Example 1.3.** Since  $\lim_{n \rightarrow \infty} an^2/n^2 = \limsup_{n \rightarrow \infty} an^2/n^2 = \liminf_{n \rightarrow \infty} an^2/n^2 = a$ :

- $(n \mapsto an^2) \notin o(n \mapsto n^2)$ , often written as  $an^2 \notin o(n^2)$ .
- $(n \mapsto an^2) \in O(n \mapsto n^2)$ , often written as  $an^2 \in O(n^2)$ .
- $(n \mapsto an^2) \in \Omega(n \mapsto n^2)$ , often written as  $an^2 \in \Omega(n^2)$ .
- $(n \mapsto an^2) \notin \omega(n \mapsto n^2)$ , often written as  $an^2 \notin \omega(n^2)$ .
- $(n \mapsto an^2) \in \Theta(n \mapsto n^2)$ , often written as  $an^2 \in \Theta(n^2)$ .

**Proposition 1.6.** For every  $f \in \mathcal{F}$  and  $g \in \mathcal{F}$ , unless the product on the right side below is  $0 \cdot \infty$  or  $\infty \cdot 0$ ,

$$\limsup_{n \rightarrow \infty} f(n)g(n) \leq \left( \limsup_{n \rightarrow \infty} f(n) \right) \left( \limsup_{n \rightarrow \infty} g(n) \right).$$

**Proposition 1.7.** For every  $f \in \mathcal{F}$  and  $g \in \mathcal{F}$ , unless the product on the right side below is  $0 \cdot \infty$  or  $\infty \cdot 0$ ,

$$\liminf_{n \rightarrow \infty} f(n)g(n) \geq \left( \liminf_{n \rightarrow \infty} f(n) \right) \left( \liminf_{n \rightarrow \infty} g(n) \right).$$

**Proposition 1.8.** If  $f \in \mathcal{F}$  and  $\liminf_{n \rightarrow \infty} f(n) > 0$ , then

$$\limsup_{n \rightarrow \infty} \frac{1}{f(n)} = \frac{1}{\liminf_{n \rightarrow \infty} f(n)},$$

where  $1/\infty$  is used to denote 0 on the right side above.

*Proof.* If  $\liminf_{n \rightarrow \infty} f(n) = \infty$ , then  $\lim_{n \rightarrow \infty} f(n) = \infty$  and  $\limsup_{n \rightarrow \infty} 1/f(n) = \lim_{n \rightarrow \infty} 1/f(n) = 0$ .

If  $\liminf_{n \rightarrow \infty} f(n) < \infty$ , consider the function  $g$  given by  $g(m) = \inf_{n \geq m} f(n) < \infty$ , which is non-decreasing. Because  $\lim_{m \rightarrow \infty} g(m) = \liminf_{n \rightarrow \infty} f(n) > 0$ , there is an  $N \in \mathbb{N}$  such that  $g(m) > 0$  for every  $m \geq N$ , which also implies  $f(n) > 0$  for every  $n \geq N$ . For every  $m \in \mathbb{N}$ , since the smaller the denominator the larger the fraction,

$$\sup_{n \geq \max(N, m)} \frac{1}{f(n)} = \frac{1}{\inf_{n \geq \max(N, m)} f(n)}.$$

By taking the limit when  $m \rightarrow \infty$ , since both sides are non-increasing with respect to  $m$ ,

$$\limsup_{n \rightarrow \infty} \frac{1}{f(n)} = \lim_{m \rightarrow \infty} \sup_{n \geq \max(N, m)} \frac{1}{f(n)} = \lim_{m \rightarrow \infty} \frac{1}{\inf_{n \geq \max(N, m)} f(n)} = \frac{1}{\liminf_{n \rightarrow \infty} f(n)}.$$

□

**Proposition 1.9.** If  $f \in \mathcal{F}$  and  $\limsup_{n \rightarrow \infty} f(n) < \infty$ , then

$$\liminf_{n \rightarrow \infty} \frac{1}{f(n)} = \frac{1}{\limsup_{n \rightarrow \infty} f(n)},$$

where  $1/0$  is used to denote  $\infty$  on the right side above.

*Proof.* If  $\limsup_{n \rightarrow \infty} f(n) = 0$ , then  $\lim_{n \rightarrow \infty} f(n) = 0$  and  $\liminf_{n \rightarrow \infty} 1/f(n) = \lim_{n \rightarrow \infty} 1/f(n) = \infty$ .

If  $\limsup_{n \rightarrow \infty} f(n) > 0$ , consider the function  $g$  given by  $g(m) = \sup_{n \geq m} f(n) > 0$ , which is non-increasing. Because  $\lim_{m \rightarrow \infty} g(m) = \limsup_{n \rightarrow \infty} f(n) < \infty$ , there is an  $N \in \mathbb{N}$  such that  $g(m) < \infty$  for every  $m \geq N$ . For every  $m \in \mathbb{N}$ , since the larger the denominator the smaller the fraction,

$$\inf_{n \geq \max(N, m)} \frac{1}{f(n)} = \frac{1}{\sup_{n \geq \max(N, m)} f(n)}.$$

By taking the limit when  $m \rightarrow \infty$ , since both sides are non-decreasing with respect to  $m$ ,

$$\liminf_{n \rightarrow \infty} \frac{1}{f(n)} = \lim_{m \rightarrow \infty} \inf_{n \geq \max(N, m)} \frac{1}{f(n)} = \lim_{m \rightarrow \infty} \frac{1}{\sup_{n \geq \max(N, m)} f(n)} = \frac{1}{\limsup_{n \rightarrow \infty} f(n)}.$$

□

Consider the functions  $f \in \mathcal{F}$ ,  $g \in \mathcal{F}$ , and  $h \in \mathcal{F}$ .

**Proposition 1.10.** If  $f \in \mathcal{F}$ , then  $f \in O(f)$ ,  $f \in \Omega(f)$ , and  $f \in \Theta(f)$ . Furthermore,  $o(f) \subseteq O(f)$  and  $\omega(f) \subseteq \Omega(f)$ .

**Proposition 1.11.** If  $f \in o(g)$  and  $g \in o(h)$ , then  $f \in o(h)$ .

*Proof.* By Proposition 1.6,

$$0 \leq \limsup_{n \rightarrow \infty} \frac{f(n)}{h(n)} = \limsup_{n \rightarrow \infty} \frac{f(n)g(n)}{g(n)h(n)} \leq \left( \limsup_{n \rightarrow \infty} \frac{f(n)}{g(n)} \right) \left( \limsup_{n \rightarrow \infty} \frac{g(n)}{h(n)} \right) = 0.$$

□

**Proposition 1.12.** If  $f \in O(g)$  and  $g \in O(h)$ , then  $f \in O(h)$ .

*Proof.* By Proposition 1.6,

$$\limsup_{n \rightarrow \infty} \frac{f(n)}{h(n)} = \limsup_{n \rightarrow \infty} \frac{f(n)g(n)}{g(n)h(n)} \leq \left( \limsup_{n \rightarrow \infty} \frac{f(n)}{g(n)} \right) \left( \limsup_{n \rightarrow \infty} \frac{g(n)}{h(n)} \right) < \infty.$$

□

**Proposition 1.13.** If  $f \in \Omega(g)$  and  $g \in \Omega(h)$ , then  $f \in \Omega(h)$ .

*Proof.* By Proposition 1.7,

$$\liminf_{n \rightarrow \infty} \frac{f(n)}{h(n)} = \liminf_{n \rightarrow \infty} \frac{f(n)g(n)}{g(n)h(n)} \geq \left( \liminf_{n \rightarrow \infty} \frac{f(n)}{g(n)} \right) \left( \liminf_{n \rightarrow \infty} \frac{g(n)}{h(n)} \right) > 0.$$

□

**Proposition 1.14.** If  $f \in \omega(g)$  and  $g \in \omega(h)$ , then  $f \in \omega(h)$ .

*Proof.* By Proposition 1.7,

$$\infty \geq \liminf_{n \rightarrow \infty} \frac{f(n)}{h(n)} = \liminf_{n \rightarrow \infty} \frac{f(n)g(n)}{g(n)h(n)} \geq \left( \liminf_{n \rightarrow \infty} \frac{f(n)}{g(n)} \right) \left( \liminf_{n \rightarrow \infty} \frac{g(n)}{h(n)} \right) = \infty.$$

□

**Proposition 1.15.** If  $f \in \Theta(g)$  and  $g \in \Theta(h)$ , then  $f \in \Theta(h)$ .

*Proof.* Since  $f \in O(g)$  and  $g \in O(h)$ , we have  $f \in O(h)$ . Since  $f \in \Omega(g)$  and  $g \in \Omega(h)$ , we have  $f \in \Omega(h)$ . □

**Theorem 1.1.** If  $f \in \mathcal{F}$  and  $g \in \mathcal{F}$ , then

- $f \in O(g)$  if and only if  $g \in \Omega(f)$ .
- $f \in o(g)$  if and only if  $g \in \omega(f)$ .

*Proof.* If  $f \in O(g)$  and  $f \notin o(g)$ , then  $\limsup_{n \rightarrow \infty} f(n)/g(n) \in (0, \infty)$ . In that case,  $g \in \Omega(f)$ , since

$$\liminf_{n \rightarrow \infty} \frac{g(n)}{f(n)} = \frac{1}{\limsup_{n \rightarrow \infty} f(n)/g(n)} > 0.$$

If  $f \in O(g)$  and  $f \in o(g)$ , then  $\limsup_{n \rightarrow \infty} f(n)/g(n) = 0$  and  $\liminf_{n \rightarrow \infty} g(n)/f(n) = \infty$ , so that  $g \in \omega(f)$ .

If  $g \in \Omega(f)$  and  $g \notin \omega(f)$ , then  $\liminf_{n \rightarrow \infty} g(n)/f(n) \in (0, \infty)$ . In that case,  $f \in O(g)$ , since

$$\limsup_{n \rightarrow \infty} \frac{f(n)}{g(n)} = \frac{1}{\liminf_{n \rightarrow \infty} g(n)/f(n)} < \infty.$$

If  $g \in \Omega(f)$  and  $g \in \omega(f)$ , then  $\liminf_{n \rightarrow \infty} g(n)/f(n) = \infty$  and  $\limsup_{n \rightarrow \infty} f(n)/g(n) = 0$ , so that  $f \in o(g)$ .  $\square$

**Proposition 1.16.** If  $f \in \mathcal{F}$  and  $g \in \mathcal{F}$ , then  $f \in \Theta(g)$  if and only if  $g \in \Theta(f)$ .

*Proof.* If  $f \in \Theta(g)$ , then  $f \in O(g)$  implies  $g \in \Omega(f)$  and  $f \in \Omega(g)$  implies  $g \in O(f)$ ; and vice versa.  $\square$

**Definition 1.9.** The following binary relations are defined on the set  $\mathcal{F}$ :

- $f \prec g$  if and only if  $f \in o(g)$ .
- $f \lesssim g$  if and only if  $f \in O(g)$ .
- $f \gtrsim g$  if and only if  $f \in \Omega(g)$ .
- $f \succ g$  if and only if  $f \in \omega(g)$ .
- $f \sim g$  if and only if  $f \in \Theta(g)$ .

**Proposition 1.17.** The binary relations  $\prec$  and  $\succ$  are strict preorders.

*Proof.* By the definition of strict preorder:

- It is false that  $f \prec f$ . If  $f \prec g$  and  $g \prec h$ , then  $f \prec h$ .
- It is false that  $f \succ g$ . If  $f \succ g$  and  $g \succ h$ , then  $f \succ h$ .

$\square$

**Proposition 1.18.** The binary relations  $\lesssim$  and  $\gtrsim$  are preorders.

*Proof.* By the definition of preorder:

- It is true that  $f \lesssim f$ . If  $f \lesssim g$  and  $g \lesssim h$ , then  $f \lesssim h$ .
- It is true that  $f \gtrsim f$ . If  $f \gtrsim g$  and  $g \gtrsim h$ , then  $f \gtrsim h$ .

$\square$

**Proposition 1.19.** The binary relation  $\sim$  is an equivalence relation.

*Proof.* It is true that  $f \sim f$ . If  $f \sim g$ , then  $g \sim f$ ; if  $g \sim f$ , then  $f \sim g$ . If  $f \sim g$  and  $g \sim h$ , then  $f \sim h$ .  $\square$

**Proposition 1.20.** The binary relations defined on the set  $\mathcal{F}$  are related by the following:

1. If  $f \prec g$ , then  $f \lesssim g$ .
2. If  $f \succ g$ , then  $f \gtrsim g$ .
3. If  $f \lesssim g$  and  $g \lesssim f$ , then  $f \sim g$ .
4. If  $f \gtrsim g$  and  $g \gtrsim f$ , then  $f \sim g$ .

5. If  $f \prec g$ , then not  $f \succsim g$ .

6. If  $f \succ g$ , then not  $f \lesssim g$ .

*Proof.* The first two claims follow from Proposition 1.10; the next two follow from Theorem 1.1; and the last two follow from the fact that  $\liminf_{n \rightarrow \infty} f(n)/g(n) \leq \limsup_{n \rightarrow \infty} f(n)/g(n)$ .  $\square$

**Definition 1.10.** Let  $A \in \{o, O, \Omega, \omega, \Theta\}$ . For any functions  $f : \mathbb{N} \rightarrow \mathbb{R}$ ,  $g : \mathbb{N} \rightarrow \mathbb{R}$ , and  $h \in \mathcal{F}$ ,

$$f(n) = g(n) + A(h(n))$$

denotes that there is a function  $l \in A(h)$  such that  $f = g + l$ .

Consider a function  $f \in \mathcal{F}$ .

**Example 1.4.** If  $a > 0$ , then  $f(n) = \Theta(af(n))$ . In order to see this, note that  $f = 0 + f$  and  $f \in \Theta(af)$ , since

$$0 < \liminf_{n \rightarrow \infty} \frac{f(n)}{af(n)} = \limsup_{n \rightarrow \infty} \frac{f(n)}{af(n)} = \frac{1}{a} < \infty.$$

**Example 1.5.** If  $f(n) = n^2 + O(n^2)$ , then  $f(n) = \Theta(n^2)$ . Suppose that there is an  $l \in O(n \mapsto n^2)$  such that  $f(n) = n^2 + l(n)$  for every  $n \in \mathbb{N}$ . In that case,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{f(n)}{n^2} &= \limsup_{n \rightarrow \infty} \frac{n^2 + l(n)}{n^2} = 1 + \limsup_{n \rightarrow \infty} \frac{l(n)}{n^2} < \infty, \\ \liminf_{n \rightarrow \infty} \frac{f(n)}{n^2} &= \liminf_{n \rightarrow \infty} \frac{n^2 + l(n)}{n^2} = 1 + \liminf_{n \rightarrow \infty} \frac{l(n)}{n^2} > 0, \end{aligned}$$

so that  $f \in \Theta(n \mapsto n^2)$ . Since  $f = 0 + f$  and  $f \in \Theta(n \mapsto n^2)$ , we have  $f(n) = \Theta(n^2)$ .

## 2 Subgaussian random variables

For details about the notation employed below, see the measure-theoretic probability notes by the same author.

Consider a probability triple  $(\Omega, \mathcal{F}, \mathbb{P})$  and a constant  $\sigma > 0$ .

**Definition 2.1.** A random variable  $X : \Omega \rightarrow \mathbb{R}$  is 0-subgaussian if and only if  $\mathbb{P}(X = 0) = 1$ .

**Definition 2.2.** A random variable  $X : \Omega \rightarrow \mathbb{R}$  is  $\sigma$ -subgaussian if and only if, for every  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}(e^{\lambda X}) \leq e^{\frac{\lambda^2 \sigma^2}{2}}.$$

**Proposition 2.1.** If a random variable  $X : \Omega \rightarrow \mathbb{R}$  is  $\sigma$ -subgaussian, then, for every  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}(e^{\lambda|X|}) \leq 2e^{\frac{\lambda^2 \sigma^2}{2}}.$$

*Proof.* For every  $\lambda \in \mathbb{R}$ , note that  $e^{\lambda|X|} = e^{\lambda X} \mathbb{I}_{\{X \geq 0\}} + e^{-\lambda X} \mathbb{I}_{\{X < 0\}}$ . Since  $e^x > 0$  for every  $x \in \mathbb{R}$ , note that  $\mathbb{E}(e^{\lambda X} \mathbb{I}_{\{X \geq 0\}}) \leq \mathbb{E}(e^{\lambda X}) \leq e^{\frac{\lambda^2 \sigma^2}{2}}$  and  $\mathbb{E}(e^{-\lambda X} \mathbb{I}_{\{X < 0\}}) \leq \mathbb{E}(e^{-\lambda X}) \leq e^{\frac{(-\lambda)^2 \sigma^2}{2}} = e^{\frac{\lambda^2 \sigma^2}{2}}$ . Therefore,

$$\mathbb{E}(e^{\lambda|X|}) = \mathbb{E}(e^{\lambda X} \mathbb{I}_{\{X \geq 0\}}) + \mathbb{E}(e^{-\lambda X} \mathbb{I}_{\{X < 0\}}) \leq 2e^{\frac{\lambda^2 \sigma^2}{2}}.$$

□

**Proposition 2.2.** If a random variable  $X : \Omega \rightarrow \mathbb{R}$  is  $\sigma$ -subgaussian, then  $\mathbb{E}(X) = 0$ .

*Proof.* Recall that  $e^x \geq x + 1$  for every  $x \in \mathbb{R}$ . Therefore,  $\mathbb{E}(e^{|X|}) \geq \mathbb{E}(|X|) + 1$  and  $\mathbb{E}(|X|) \leq 2e^{\frac{\sigma^2}{2}} - 1$ .

For every  $\lambda \in \mathbb{R}$ , recall that the function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  given by  $\phi(x) = e^{\lambda x}$  is convex. By Jensen's inequality,

$$e^{\lambda \mathbb{E}(X)} = \phi(\mathbb{E}(X)) \leq \mathbb{E}(\phi(X)) = \mathbb{E}(e^{\lambda X}) \leq e^{\frac{\lambda^2 \sigma^2}{2}},$$

so that  $\lambda \mathbb{E}(X) \leq \lambda^2 \sigma^2 / 2$  for every  $\lambda \in \mathbb{R}$ . If  $\lambda < 0$ , then  $\mathbb{E}(X) \geq \lambda \sigma^2 / 2$ . If  $\lambda > 0$ , then  $\mathbb{E}(X) \leq \lambda \sigma^2 / 2$ . Therefore,

$$0 = \lim_{\lambda \rightarrow 0^-} \frac{\lambda \sigma^2}{2} \leq \mathbb{E}(X) \leq \lim_{\lambda \rightarrow 0^+} \frac{\lambda \sigma^2}{2} = 0.$$

□

**Proposition 2.3.** If a random variable  $X : \Omega \rightarrow \mathbb{R}$  is  $\sigma$ -subgaussian, then  $\text{Var}(X) \leq \sigma^2$ .

*Proof.* Recall that  $e^x = \sum_{n=0}^{\infty} x^n / n!$  for every  $x \in \mathbb{R}$ . Therefore, for every  $\lambda \geq 0$  and  $k \in \mathbb{N}$ ,

$$e^{\lambda|X|} = \sum_{n=0}^{\infty} \frac{\lambda^n |X|^n}{n!} \geq \sum_{n=0}^k \frac{\lambda^n |X|^n}{n!} = \sum_{n=0}^k \left| \frac{\lambda^n X^n}{n!} \right| \geq \left| \sum_{n=0}^k \frac{\lambda^n X^n}{n!} \right|.$$

Since  $\mathbb{E}(e^{\lambda|X|}) < \infty$ , note that  $\mathbb{E}(|X|^k) < \infty$  for every  $k \in \mathbb{N}$ . By the dominated convergence theorem,

$$\mathbb{E}(e^{\lambda X}) = \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{\lambda^n X^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{\lambda^n \mathbb{E}(X^n)}{n!} = 1 + \frac{\lambda^2 \mathbb{E}(X^2)}{2} + \sum_{n=3}^{\infty} \frac{\lambda^n \mathbb{E}(X^n)}{n!},$$

where we also used the fact that  $\mathbb{E}(X) = 0$ .

For every  $\lambda \in [0, 1]$ , note that  $\lambda^{2n} \leq \lambda^4$  for every  $n \geq 2$ . Therefore, for every  $\lambda \in [0, 1]$ ,

$$e^{\frac{\lambda^2 \sigma^2}{2}} = \sum_{n=0}^{\infty} \frac{\lambda^{2n} \sigma^{2n}}{2^n n!} = 1 + \frac{\lambda^2 \sigma^2}{2} + \sum_{n=2}^{\infty} \frac{\lambda^{2n} \sigma^{2n}}{2^n n!} \leq 1 + \frac{\lambda^2 \sigma^2}{2} + \lambda^4 \sum_{n=2}^{\infty} \frac{\sigma^{2n}}{2^n n!} \leq 1 + \frac{\lambda^2 \sigma^2}{2} + \lambda^4 e^{\frac{\sigma^2}{2}}.$$

For every  $\lambda \in [0, 1]$ , by the definition of a  $\sigma$ -subgaussian random variable,

$$\frac{\lambda^2 \mathbb{E}(X^2)}{2} + \sum_{n=3}^{\infty} \frac{\lambda^n \mathbb{E}(X^n)}{n!} \leq \frac{\lambda^2 \sigma^2}{2} + \lambda^4 e^{\frac{\sigma^2}{2}}.$$

For every  $\lambda \in (0, 1]$ , by multiplying both sides by  $2/\lambda^2$ ,

$$\mathbb{E}(X^2) + 2 \sum_{n=3}^{\infty} \frac{\lambda^{n-2} \mathbb{E}(X^n)}{n!} \leq \sigma^2 + 2\lambda^2 e^{\frac{\sigma^2}{2}}.$$

By taking the limit of both sides when  $\lambda \rightarrow 0^+$ ,

$$\mathbb{E}(X^2) + 2 \lim_{\lambda \rightarrow 0^+} \sum_{n=3}^{\infty} \frac{\lambda^{n-2} \mathbb{E}(X^n)}{n!} \leq \sigma^2 + 2e^{\frac{\sigma^2}{2}} \lim_{\lambda \rightarrow 0^+} \lambda^2 = \sigma^2.$$

If the limit on the left side above is zero, then  $\mathbb{E}(X^2) \leq \sigma^2$ . In that case, considering that  $\mathbb{E}(X) = 0$ , note that  $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2) \leq \sigma^2$ , so that the proof will be complete. For every  $\lambda \in (0, 1]$ ,

$$\left| \sum_{n=3}^{\infty} \frac{\lambda^{n-2} \mathbb{E}(X^n)}{n!} \right| = \lambda \left| \sum_{n=3}^{\infty} \frac{\lambda^{n-3} \mathbb{E}(X^n)}{n!} \right| \leq \lambda \sum_{n=3}^{\infty} \frac{\lambda^{n-3} |\mathbb{E}(X^n)|}{n!}.$$

For every  $k \in \mathbb{N}$  and  $\lambda \in (0, 1]$ , note that  $\mathbb{E}(X^k) \leq \mathbb{E}(|X|^k) < \infty$  and  $\lambda^k \leq 1$ . Therefore,

$$\left| \sum_{n=3}^{\infty} \frac{\lambda^{n-2} \mathbb{E}(X^n)}{n!} \right| \leq \lambda \sum_{n=3}^{\infty} \frac{\lambda^{n-3} \mathbb{E}(|X|^n)}{n!} \leq \lambda \sum_{n=3}^{\infty} \frac{\mathbb{E}(|X|^n)}{n!} \leq \lambda \mathbb{E}(e^{|X|}) \leq 2\lambda e^{\frac{\sigma^2}{2}},$$

so that

$$0 \leq \lim_{\lambda \rightarrow 0^+} \left| \sum_{n=3}^{\infty} \frac{\lambda^{n-2} \mathbb{E}(X^n)}{n!} \right| \leq 2e^{\frac{\sigma^2}{2}} \lim_{\lambda \rightarrow 0^+} \lambda = 0.$$

□

**Proposition 2.4.** If a random variable  $X : \Omega \rightarrow \mathbb{R}$  is  $\sigma$ -subgaussian, then  $cX$  is  $|c|\sigma$ -subgaussian for every  $c \in \mathbb{R}$ .

*Proof.* This proposition is trivial if  $c = 0$ . If  $c \neq 0$ ,  $cX$  is a random variable and, for every  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}(e^{\lambda(cX)}) = \mathbb{E}(e^{(\lambda c)X}) \leq e^{\frac{(\lambda c)^2 \sigma^2}{2}} = e^{\frac{\lambda^2 c^2 \sigma^2}{2}} = e^{\frac{\lambda^2 |c|^2 \sigma^2}{2}} = e^{\frac{\lambda^2 (|c|\sigma)^2}{2}}.$$

□

Consider the constants  $\sigma_1 > 0$  and  $\sigma_2 > 0$ .

**Proposition 2.5.** If the random variable  $X_1 : \Omega \rightarrow \mathbb{R}$  is  $\sigma_1$ -subgaussian, the random variable  $X_2$  is  $\sigma_2$ -subgaussian, and  $X_1$  and  $X_2$  are independent, then  $X_1 + X_2$  is  $\sqrt{\sigma_1^2 + \sigma_2^2}$ -subgaussian.

*Proof.* For every  $\lambda \in \mathbb{R}$ , because  $e^{\lambda X_1}$  and  $e^{\lambda X_2}$  are independent and  $\mathbb{P}$ -integrable,

$$\mathbb{E}(e^{\lambda(X_1+X_2)}) = \mathbb{E}(e^{\lambda X_1 + \lambda X_2}) = \mathbb{E}(e^{\lambda X_1} e^{\lambda X_2}) = \mathbb{E}(e^{\lambda X_1}) \mathbb{E}(e^{\lambda X_2}) \leq e^{\frac{\lambda^2 \sigma_1^2}{2}} e^{\frac{\lambda^2 \sigma_2^2}{2}} = e^{\frac{\lambda^2 (\sigma_1^2 + \sigma_2^2)}{2}},$$

so that the random variable  $X_1 + X_2$  is  $\sqrt{\sigma_1^2 + \sigma_2^2}$ -subgaussian. □

**Proposition 2.6.** If the random variable  $X_1 : \Omega \rightarrow \mathbb{R}$  is  $\sigma_1$ -subgaussian and the random variable  $X_2$  is  $\sigma_2$ -subgaussian, then  $X_1 + X_2$  is  $(\sigma_1 + \sigma_2)$ -subgaussian.

*Proof.* Note that  $\mathbb{E}(|e^{\lambda X_1}|^p) = \mathbb{E}(e^{\lambda p X_1}) < \infty$  and  $\mathbb{E}(|e^{\lambda X_2}|^q) = \mathbb{E}(e^{\lambda q X_2}) < \infty$  for every  $\lambda \in \mathbb{R}$ ,  $p \geq 1$ , and  $q \geq 1$ . By Hölder's inequality, if  $p > 1$  and  $p^{-1} + q^{-1} = 1$ , then

$$\mathbb{E}(e^{\lambda(X_1+X_2)}) = \mathbb{E}(e^{\lambda X_1 + \lambda X_2}) = \mathbb{E}(e^{\lambda X_1} e^{\lambda X_2}) \leq \mathbb{E}(|e^{\lambda X_1}|^p)^{\frac{1}{p}} \mathbb{E}(|e^{\lambda X_2}|^q)^{\frac{1}{q}} = \mathbb{E}(e^{\lambda p X_1})^{\frac{1}{p}} \mathbb{E}(e^{\lambda q X_2})^{\frac{1}{q}}.$$

By the definition of subgaussian random variables,

$$\mathbb{E}(e^{\lambda(X_1+X_2)}) \leq \left( e^{\frac{\lambda^2 p^2 \sigma_1^2}{2}} \right)^{\frac{1}{p}} \left( e^{\frac{\lambda^2 q^2 \sigma_2^2}{2}} \right)^{\frac{1}{q}} = e^{\frac{\lambda^2 p \sigma_1^2}{2}} e^{\frac{\lambda^2 q \sigma_2^2}{2}} = e^{\frac{\lambda^2}{2} (p\sigma_1^2 + q\sigma_2^2)}.$$

Let  $p = (\sigma_1 + \sigma_2)/\sigma_1$  and  $q = (\sigma_1 + \sigma_2)/\sigma_2$ , so that  $p > 1$  and  $p^{-1} + q^{-1} = 1$ . In that case, for every  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}(e^{\lambda(X_1+X_2)}) \leq e^{\frac{\lambda^2}{2} \left( \frac{\sigma_1 + \sigma_2}{\sigma_1} \sigma_1^2 + \frac{\sigma_1 + \sigma_2}{\sigma_2} \sigma_2^2 \right)} = e^{\frac{\lambda^2}{2} (\sigma_1^2 + 2\sigma_1 \sigma_2 + \sigma_2^2)} = e^{\frac{\lambda^2 (\sigma_1 + \sigma_2)^2}{2}},$$

so that the random variable  $X_1 + X_2$  is  $(\sigma_1 + \sigma_2)$ -subgaussian. □

**Proposition 2.7.** If a random variable  $X : \Omega \rightarrow \mathbb{R}$  has a normal distribution with mean 0 and variance 1, then  $X$  is 1-subgaussian.

*Proof.* For every  $\lambda \in \mathbb{R}$ , considering a probability density function for the random variable  $X$ ,

$$\mathbb{E}(e^{\lambda X}) = \int_{\mathbb{R}} e^{\lambda x} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \text{Leb}(dx) = \int_{\mathbb{R}} \frac{e^{\lambda x - \frac{x^2}{2}}}{\sqrt{2\pi}} \text{Leb}(dx) = e^{\frac{\lambda^2}{2}} \int_{\mathbb{R}} \frac{e^{-\frac{(x-\lambda)^2}{2}}}{\sqrt{2\pi}} \text{Leb}(dx) = e^{\frac{\lambda^2}{2}}.$$

where we used the fact that  $\lambda x - \frac{x^2}{2} = -\frac{(x-\lambda)^2}{2} + \frac{\lambda^2}{2}$  and recognized a probability density function for a random variable that has a normal distribution with mean  $\lambda$  and variance 1.  $\square$

**Proposition 2.8.** If a random variable  $X : \Omega \rightarrow \mathbb{R}$  has a normal distribution with mean 0 and variance  $\sigma^2$ , then  $X$  is  $\sigma$ -subgaussian.

*Proof.* Recall that  $X/\sigma$  has a normal distribution with mean 0 and variance  $\sigma^2/\sigma^2 = 1$ . Therefore,  $X/\sigma$  is 1-subgaussian, so that  $\sigma \frac{X}{\sigma} = X$  is  $|\sigma|$ -subgaussian.  $\square$

**Lemma 2.1** (Hoeffding's lemma). If  $X : \Omega \rightarrow \mathbb{R}$  is a random variable such that  $\mathbb{E}(X) = 0$  and  $\mathbb{P}(X \in [a, b]) = 1$  for some  $a < b$ , then  $X$  is  $(b - a)/2$ -subgaussian.



### 3 Concentration of measure

Consider a probability triple  $(\Omega, \mathcal{F}, \mathbb{P})$  and a constant  $\sigma > 0$ .

**Theorem 3.1.** If  $X : \Omega \rightarrow \mathbb{R}$  is a  $\sigma$ -subgaussian random variable, then, for every  $\epsilon \geq 0$ ,

$$\begin{aligned}\mathbb{P}(X \leq -\epsilon) &\leq e^{-\frac{\epsilon^2}{2\sigma^2}}, \\ \mathbb{P}(X \geq \epsilon) &\leq e^{-\frac{\epsilon^2}{2\sigma^2}}, \\ \mathbb{P}(|X| \geq \epsilon) &\leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}.\end{aligned}$$

*Proof.* Recall that the function  $g : \mathbb{R} \rightarrow [0, \infty]$  given by  $g(x) = e^{\lambda x}$  is non-decreasing for every  $\lambda \geq 0$ . For every  $\epsilon \in \mathbb{R}$ , by Markov's inequality,

$$\begin{aligned}\mathbb{E}(e^{-\lambda X}) &= \mathbb{E}(g(-X)) \geq g(\epsilon)\mathbb{P}(-X \geq \epsilon) = e^{\lambda\epsilon}\mathbb{P}(X \leq -\epsilon), \\ \mathbb{E}(e^{\lambda X}) &= \mathbb{E}(g(X)) \geq g(\epsilon)\mathbb{P}(X \geq \epsilon) = e^{\lambda\epsilon}\mathbb{P}(X \geq \epsilon).\end{aligned}$$

For every  $\epsilon \in \mathbb{R}$  and  $\lambda \geq 0$ , since  $X$  is a  $\sigma$ -subgaussian random variable and  $e^{\lambda\epsilon} > 0$ ,

$$\begin{aligned}\mathbb{P}(X \leq -\epsilon) &\leq \frac{\mathbb{E}(e^{-\lambda X})}{e^{\lambda\epsilon}} \leq \frac{e^{\frac{(-\lambda)^2\sigma^2}{2}}}{e^{\lambda\epsilon}} = e^{\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon}, \\ \mathbb{P}(X \geq \epsilon) &\leq \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda\epsilon}} \leq \frac{e^{\frac{\lambda^2\sigma^2}{2}}}{e^{\lambda\epsilon}} = e^{\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon}.\end{aligned}$$

For every  $\epsilon \geq 0$ , let  $\lambda = \epsilon/\sigma^2$ , so that  $\lambda \geq 0$ . In that case,

$$\begin{aligned}\mathbb{P}(X \leq -\epsilon) &\leq e^{\frac{\epsilon^2}{\sigma^4} \frac{\sigma^2}{2} - \frac{\epsilon^2}{\sigma^2}} = e^{\frac{\epsilon^2}{2\sigma^2} - \frac{\epsilon^2}{\sigma^2}} = e^{\frac{\epsilon^2}{\sigma^2}(\frac{1}{2} - 1)} = e^{-\frac{\epsilon^2}{2\sigma^2}}, \\ \mathbb{P}(X \geq \epsilon) &\leq e^{\frac{\epsilon^2}{\sigma^4} \frac{\sigma^2}{2} - \frac{\epsilon^2}{\sigma^2}} = e^{\frac{\epsilon^2}{2\sigma^2} - \frac{\epsilon^2}{\sigma^2}} = e^{\frac{\epsilon^2}{\sigma^2}(\frac{1}{2} - 1)} = e^{-\frac{\epsilon^2}{2\sigma^2}}.\end{aligned}$$

Therefore, for every  $\epsilon \geq 0$ ,

$$\mathbb{P}(|X| \geq \epsilon) = \mathbb{P}(\{X \leq -\epsilon\} \cup \{X \geq \epsilon\}) \leq \mathbb{P}(X \leq -\epsilon) + \mathbb{P}(X \geq \epsilon) \leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}.$$

□

**Proposition 3.1.** If  $X : \Omega \rightarrow \mathbb{R}$  is a  $\sigma$ -subgaussian random variable, then, for every  $\delta \in (0, 1]$ ,

$$\begin{aligned}\mathbb{P}\left(X \leq -\sqrt{2\sigma^2 \log(1/\delta)}\right) &\leq \delta, \\ \mathbb{P}\left(X \geq \sqrt{2\sigma^2 \log(1/\delta)}\right) &\leq \delta, \\ \mathbb{P}\left(|X| \geq \sqrt{2\sigma^2 \log(2/\delta)}\right) &\leq \delta.\end{aligned}$$

*Proof.* Let  $\delta \in (0, 1]$ . If  $\epsilon = \sqrt{2\sigma^2 \log(1/\delta)}$ , then  $\epsilon \geq 0$  and  $\delta = e^{-\frac{\epsilon^2}{2\sigma^2}}$ , which implies the first two inequalities. If  $\epsilon = \sqrt{2\sigma^2 \log(2/\delta)}$ , then  $\epsilon \geq 0$  and  $\delta = 2e^{-\frac{\epsilon^2}{2\sigma^2}}$ , which implies the last inequality. □

**Proposition 3.2.** If  $X : \Omega \rightarrow \mathbb{R}$  is a  $\sigma$ -subgaussian random variable, then, for every  $\delta \in (0, 1]$ ,

$$\begin{aligned}\mathbb{P}\left(X > -\sqrt{2\sigma^2 \log(1/\delta)}\right) &\geq 1 - \delta, \\ \mathbb{P}\left(X < \sqrt{2\sigma^2 \log(1/\delta)}\right) &\geq 1 - \delta, \\ \mathbb{P}\left(|X| < \sqrt{2\sigma^2 \log(2/\delta)}\right) &\geq 1 - \delta.\end{aligned}$$

*Proof.* These inequalities follow from Proposition 3.1 and the fact that  $\mathbb{P}(F^c) = 1 - \mathbb{P}(F)$  for every  $F \in \mathcal{F}$ . □

Consider a sequence of independent random variables  $(X_k : \Omega \rightarrow \mathbb{R} \mid k \in \mathbb{N}^+)$ , each of which has the same law as a random variable  $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$  and let  $\mu = \mathbb{E}(X)$ .

**Definition 3.1.** For every  $t \in \mathbb{N}^+$ , the sample mean  $M_t : \Omega \rightarrow \mathbb{R}$  after  $t$  observations is given by

$$M_t(\omega) = \frac{1}{t} \sum_{k=1}^t X_k(\omega).$$

**Proposition 3.3.** For every  $t \in \mathbb{N}^+$ ,  $\mathbb{E}(M_t) = \mu$  and  $\text{Var}(M_t) = \text{Var}(X)/t$ .

*Proof.* Recall that  $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$  is a vector space over  $\mathbb{R}$ , so that  $M_t \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ . By the linearity of expectation,

$$\mathbb{E}(M_t) = \mathbb{E}\left(\frac{1}{t} \sum_{k=1}^t X_k\right) = \frac{1}{t} \sum_{k=1}^t \mathbb{E}(X_k) = \frac{1}{t} t \mu.$$

For every  $c \in \mathbb{R}$  and  $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ , recall that

$$\text{Var}(cY) = \mathbb{E}((cY)^2) - \mathbb{E}(cY)^2 = \mathbb{E}(c^2 Y^2) - (c\mathbb{E}(Y))^2 = c^2 \mathbb{E}(Y^2) - c^2 \mathbb{E}(Y)^2 = c^2 \text{Var}(Y).$$

Therefore, because the random variables  $(X_k \mid k \in \mathbb{N}^+)$  are independent and identically distributed,

$$\text{Var}(M_t) = \text{Var}\left(\frac{1}{t} \sum_{k=1}^t X_k\right) = \frac{1}{t^2} \text{Var}\left(\sum_{k=1}^t X_k\right) = \frac{1}{t^2} \sum_{k=1}^t \text{Var}(X_k) = \frac{1}{t^2} t \text{Var}(X).$$

□

**Proposition 3.4.** For every  $t \in \mathbb{N}^+$  and  $\epsilon > 0$ ,

$$\mathbb{P}(|M_t - \mu| \geq \epsilon) \leq \frac{\text{Var}(X)}{t\epsilon^2}.$$

*Proof.* By Chebyshev's inequality, for every  $\epsilon \geq 0$ ,

$$\frac{\text{Var}(X)}{t} = \text{Var}(M_t) = \mathbb{E}(|M_t - \mu|^2) \geq \epsilon^2 \mathbb{P}(|M_t - \mu| \geq \epsilon).$$

□

**Proposition 3.5.** If  $X - \mu$  is a  $\sigma$ -subgaussian random variable, then, for every  $t \in \mathbb{N}^+$  and  $\epsilon > 0$ ,

$$\mathbb{P}(|M_t - \mu| \geq \epsilon) \leq \frac{\sigma^2}{t\epsilon^2}.$$

*Proof.* This proposition is a consequence of Proposition 2.3 and Proposition 3.4, since

$$\sigma^2 \geq \text{Var}(X - \mu) = \mathbb{E}((X - \mu)^2) - \mathbb{E}(X - \mu)^2 = \text{Var}(X) - (\mathbb{E}(X) - \mu)^2 = \text{Var}(X).$$

□

**Proposition 3.6.** If  $X - \mu$  is a  $\sigma$ -subgaussian random variable, then, for every  $t \in \mathbb{N}^+$  and  $\epsilon \geq 0$ ,

$$\begin{aligned} \mathbb{P}(M_t \leq \mu - \epsilon) &\leq e^{-\frac{t\epsilon^2}{2\sigma^2}}, \\ \mathbb{P}(M_t \geq \mu + \epsilon) &\leq e^{-\frac{t\epsilon^2}{2\sigma^2}}, \\ \mathbb{P}(|M_t - \mu| \geq \epsilon) &\leq 2e^{-\frac{t\epsilon^2}{2\sigma^2}}. \end{aligned}$$

*Proof.* Recall that  $\mathbb{E}(X - \mu) = 0$  and  $\text{Var}(X - \mu) = \text{Var}(X)$ . For every  $t \in \mathbb{N}^+$ ,

$$M_t - \mu = \left(\frac{1}{t} \sum_{k=1}^t X_k\right) - \frac{1}{t} t \mu = \frac{1}{t} \sum_{k=1}^t (X_k - \mu).$$

Because  $(X_k - \mu \mid k \in \mathbb{N}^+)$  are independent  $\sigma$ -subgaussian random variables, Proposition 2.5 guarantees that  $\sum_{k=1}^t (X_k - \mu)$  is  $(\sigma\sqrt{t})$ -subgaussian and Proposition 2.4 that  $M_t - \mu$  is  $(\sigma/\sqrt{t})$ -subgaussian. By Theorem 3.1,

$$\begin{aligned}\mathbb{P}(M_t - \mu \leq -\epsilon) &\leq e^{-\frac{\epsilon^2}{2(\sigma/\sqrt{t})^2}} = e^{-\frac{\epsilon^2}{2(\sigma^2/t)}} = e^{-\frac{t\epsilon^2}{2\sigma^2}}, \\ \mathbb{P}(M_t - \mu \geq \epsilon) &\leq e^{-\frac{\epsilon^2}{2(\sigma/\sqrt{t})^2}} = e^{-\frac{\epsilon^2}{2(\sigma^2/t)}} = e^{-\frac{t\epsilon^2}{2\sigma^2}}, \\ \mathbb{P}(|M_t - \mu| \geq \epsilon) &\leq 2e^{-\frac{\epsilon^2}{2(\sigma/\sqrt{t})^2}} = 2e^{-\frac{\epsilon^2}{2(\sigma^2/t)}} = 2e^{-\frac{t\epsilon^2}{2\sigma^2}}.\end{aligned}$$

□

**Proposition 3.7.** If  $X - \mu$  is a  $\sigma$ -subgaussian random variable, then, for every  $t \in \mathbb{N}^+$  and  $\delta \in (0, 1]$ ,

$$\begin{aligned}\mathbb{P}\left(M_t \leq \mu - \sqrt{2\sigma^2 \log(1/\delta)/t}\right) &\leq \delta, \\ \mathbb{P}\left(M_t \geq \mu + \sqrt{2\sigma^2 \log(1/\delta)/t}\right) &\leq \delta, \\ \mathbb{P}(|M_t - \mu| \geq \sqrt{2\sigma^2 \log(2/\delta)/t}) &\leq \delta.\end{aligned}$$

*Proof.* Let  $\delta \in (0, 1]$ . If  $\epsilon = \sqrt{2\sigma^2 \log(1/\delta)/t}$ , then  $\epsilon \geq 0$  and  $\delta = e^{-\frac{t\epsilon^2}{2\sigma^2}}$ , which implies the first two inequalities. If  $\epsilon = \sqrt{2\sigma^2 \log(2/\delta)/t}$ , then  $\epsilon \geq 0$  and  $\delta = 2e^{-\frac{t\epsilon^2}{2\sigma^2}}$ , which implies the last inequality. □

**Proposition 3.8.** If  $X - \mu$  is a  $\sigma$ -subgaussian random variable, then, for every  $t \in \mathbb{N}^+$  and  $\delta \in (0, 1]$ ,

$$\begin{aligned}\mathbb{P}\left(M_t > \mu - \sqrt{2\sigma^2 \log(1/\delta)/t}\right) &\geq 1 - \delta, \\ \mathbb{P}\left(M_t < \mu + \sqrt{2\sigma^2 \log(1/\delta)/t}\right) &\geq 1 - \delta, \\ \mathbb{P}(|M_t - \mu| < \sqrt{2\sigma^2 \log(2/\delta)/t}) &\geq 1 - \delta.\end{aligned}$$

*Proof.* These inequalities follow from Proposition 3.7 and the fact that  $\mathbb{P}(F^c) = 1 - \mathbb{P}(F)$  for every  $F \in \mathcal{F}$ . □

**Theorem 3.2** (Hoeffding's inequality). Consider a sequence of independent random variables  $(Y_k : \Omega \rightarrow \mathbb{R} \mid k \in \mathbb{N}^+)$  and suppose that there are constants  $a_k \in \mathbb{R}$  and  $b_k \in \mathbb{R}$  such that  $a_k < b_k$  and  $\mathbb{P}(Y_k \in [a_k, b_k]) = 1$  for every  $k \in \mathbb{N}^+$ . In that case, for every  $t \in \mathbb{N}^+$  and  $\epsilon \geq 0$ ,

$$\mathbb{P}\left(\frac{1}{t} \sum_{k=1}^t (Y_k - \mathbb{E}(Y_k)) \geq \epsilon\right) \leq e^{-\frac{2t^2\epsilon^2}{\sum_{k=1}^t (b_k - a_k)^2}}.$$

*Proof.* For every  $k \in \mathbb{N}^+$ , note that  $\mathbb{E}(Y_k - \mathbb{E}(Y_k)) = 0$  and  $\mathbb{P}((Y_k - \mathbb{E}(Y_k)) \in [a_k - \mathbb{E}(Y_k), b_k - \mathbb{E}(Y_k)]) = 1$ , so that  $Y_k - \mathbb{E}(Y_k)$  is  $(b_k - a_k)/2$ -subgaussian by Lemma 2.1. Because  $(Y_k - \mathbb{E}(Y_k) \mid k \in \mathbb{N}^+)$  are independent random variables, Proposition 2.5 guarantees that  $\sum_{k=1}^t (Y_k - \mathbb{E}(Y_k))$  is  $\sqrt{\sum_{k=1}^t (b_k - a_k)^2/4}$ -subgaussian and Proposition 2.4 that  $\sum_{k=1}^t (Y_k - \mathbb{E}(Y_k))/t$  is  $\sqrt{\sum_{k=1}^t (b_k - a_k)^2/(4t^2)}$ -subgaussian. By Theorem 3.1,

$$\mathbb{P}\left(\frac{1}{t} \sum_{k=1}^t (Y_k - \mathbb{E}(Y_k)) \geq \epsilon\right) \leq e^{-\frac{\epsilon^2}{2(\sqrt{\sum_{k=1}^t (b_k - a_k)^2/(4t^2)}}^2}} = e^{-\frac{\epsilon^2}{\frac{1}{2t^2} \sum_{k=1}^t (b_k - a_k)^2}} = e^{-\frac{2t^2\epsilon^2}{\sum_{k=1}^t (b_k - a_k)^2}}.$$

□

**Theorem 3.3** (Bretagnolle-Huber-Carol inequality). Suppose that there is an  $m \in \mathbb{N}^+$  such that  $X(\omega) \in \{1, \dots, m\}$  for every  $\omega \in \Omega$ . Consider a vector  $p \in [0, 1]^m$  such that  $p_i = \mathbb{P}(X = i)$  for every  $i \in \{1, \dots, m\}$  and a random vector  $P_t : \Omega \rightarrow [0, 1]^m$  such that  $P_{t,i} = 1/t \sum_{k=1}^t \mathbb{I}_{\{X_k = i\}}$  for every  $t \in \mathbb{N}^+$  and  $i \in \{1, \dots, m\}$ . For every  $\delta \in (0, 1]$ ,

$$\mathbb{P}\left(\|P_t - p\|_1 \geq \sqrt{2(\log(1/\delta) + m \log(2))/t}\right) \leq \delta.$$

*Proof.* Recall that  $|a| = \max(a, -a)$  for every  $a \in \mathbb{R}$ . Therefore, for every  $t \in \mathbb{N}^+$ ,

$$\|P_t - p\|_1 = \sum_{i=1}^m |P_{t,i} - p_i| = \sum_{i=1}^m \max_{\lambda_i \in \{-1,1\}} \lambda_i (P_{t,i} - p_i) = \max_{\lambda \in \{-1,1\}^m} \sum_{i=1}^m \lambda_i (P_{t,i} - p_i).$$

For every  $t \in \mathbb{N}^+$ , by expanding the previous expression and exchanging the order of the summations,

$$\|P_t - p\|_1 = \max_{\lambda \in \{-1,1\}^m} \sum_{i=1}^m \lambda_i \left( \frac{1}{t} \sum_{k=1}^t \mathbb{I}_{\{X_k=i\}} - \frac{1}{t} \sum_{k=1}^t p_i \right) = \max_{\lambda \in \{-1,1\}^m} \frac{1}{t} \sum_{k=1}^t \sum_{i=1}^m \lambda_i \mathbb{I}_{\{X_k=i\}} - \lambda_i p_i.$$

For every  $k \in \{1, \dots, t\}$  and  $\lambda \in \{-1,1\}^m$ , let  $Y_k^{(\lambda)} = \sum_{i=1}^m \lambda_i \mathbb{I}_{\{X_k=i\}} = \lambda_{X_k}$ , so that  $|Y_k^{(\lambda)}| \leq 1$  and

$$\mathbb{E} \left( Y_k^{(\lambda)} \right) = \mathbb{E} \left( \sum_{i=1}^m \lambda_i \mathbb{I}_{\{X_k=i\}} \right) = \sum_{i=1}^m \lambda_i \mathbb{P}(X_k = i) = \sum_{i=1}^m \lambda_i \mathbb{P}(X = i) = \sum_{i=1}^m \lambda_i p_i.$$

For every  $t \in \mathbb{N}^+$ , by rewriting a previous expression,

$$\|P_t - p\|_1 = \max_{\lambda \in \{-1,1\}^m} \frac{1}{t} \sum_{k=1}^t \left( Y_k^{(\lambda)} - \mathbb{E} \left( Y_k^{(\lambda)} \right) \right).$$

Therefore, for every  $t \in \mathbb{N}^+$  and  $\epsilon \geq 0$ ,

$$\{\|P_t - p\|_1 \geq \epsilon\} = \left\{ \max_{\lambda \in \{-1,1\}^m} \frac{1}{t} \sum_{k=1}^t \left( Y_k^{(\lambda)} - \mathbb{E} \left( Y_k^{(\lambda)} \right) \right) \geq \epsilon \right\} = \bigcup_{\lambda \in \{-1,1\}^m} \left\{ \frac{1}{t} \sum_{k=1}^t \left( Y_k^{(\lambda)} - \mathbb{E} \left( Y_k^{(\lambda)} \right) \right) \geq \epsilon \right\}.$$

By employing a union bound, Theorem 3.2, and the fact that the set  $\{-1,1\}^m$  has  $2^m$  elements,

$$\mathbb{P}(\|P_t - p\|_1 \geq \epsilon) \leq \sum_{\lambda \in \{-1,1\}^m} \mathbb{P} \left( \frac{1}{t} \sum_{k=1}^t \left( Y_k^{(\lambda)} - \mathbb{E} \left( Y_k^{(\lambda)} \right) \right) \geq \epsilon \right) \leq \sum_{\lambda \in \{-1,1\}^m} e^{-\frac{t\epsilon^2}{2}} = 2^m e^{-\frac{t\epsilon^2}{2}}$$

Let  $\delta \in (0, 1]$ . If  $\epsilon = \sqrt{2(\log(1/\delta) + m \log(2)) / t}$ , then  $\epsilon \geq 0$  and  $\delta = 2^m e^{-\frac{t\epsilon^2}{2}}$ . Therefore,

$$\mathbb{P} \left( \|P_t - p\|_1 \geq \sqrt{2(\log(1/\delta) + m \log(2)) / t} \right) \leq \delta.$$

□

## 4 Stochastic bandits

**Definition 4.1.** A set of actions  $\mathcal{A}$  is a non-empty subset of  $\mathbb{N}$ .

**Definition 4.2.** For a set of actions  $\mathcal{A}$ , consider a sequence of probability measures  $\nu = (P_a \mid a \in \mathcal{A})$  on the measurable space  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . If  $h : \mathbb{R} \rightarrow \mathbb{R}$  is a  $\mathcal{B}(\mathbb{R})$ -measurable function and there is a constant  $c \in [0, \infty)$  such that  $\int_{\mathbb{R}} |h(x)| P_a(dx) \leq c$  for every action  $a \in \mathcal{A}$ , then  $h$  is  $\nu$ -integrable.

**Definition 4.3.** For a set of actions  $\mathcal{A}$ , consider a sequence of probability measures  $\nu = (P_a \mid a \in \mathcal{A})$  on the measurable space  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . If the identity function is  $\nu$ -integrable, the mean  $\mu_a^\nu$  of action  $a$  is defined by  $\mu_a^\nu = \int_{\mathbb{R}} x P_a(dx)$  and the supremum mean  $\mu_*^\nu$  is defined by  $\mu_*^\nu = \sup_a \mu_a^\nu$ . If  $\mu_a^\nu = \mu_*^\nu$  for some action  $a \in \mathcal{A}$ , then  $\nu$  is a stochastic bandit for the set of actions  $\mathcal{A}$ .

**Proposition 4.1.** If  $\nu = (P_a \mid a \in \mathcal{A})$  is a stochastic bandit for the set of actions  $\mathcal{A}$ , then there is a constant  $c \in [0, \infty)$  such that  $\mu_a^\nu \in [-c, c]$  for every action  $a \in \mathcal{A}$ .

*Proof.* Since the identity function is  $\nu$ -integrable, there is a constant  $c \in [0, \infty)$  such that  $\int_{\mathbb{R}} |x| P_a(dx) \leq c$  for every action  $a \in \mathcal{A}$ . Therefore,  $|\mu_a^\nu| = \left| \int_{\mathbb{R}} x P_a(dx) \right| \leq \int_{\mathbb{R}} |x| P_a(dx) \leq c$  for every action  $a \in \mathcal{A}$ .  $\square$

**Definition 4.4.** For a set of actions  $\mathcal{A}$ , a policy  $\pi$  is a sequence of functions  $(\pi_t : \mathbb{R}^t \rightarrow \mathcal{A} \mid t \in \mathbb{N}^+)$ , where the so-called policy  $\pi_t$  for time step  $t$  is  $\mathcal{B}(\mathbb{R}^t)$ -measurable.

**Proposition 4.2.** For a set of actions  $\mathcal{A}$ , a stochastic bandit  $\nu = (P_a \mid a \in \mathcal{A})$ , and a policy  $\pi = (\pi_t \mid t \in \mathbb{N}^+)$ , there is a probability triple  $(\Omega, \mathcal{F}, \mathbb{P})$  carrying a stochastic process  $(X_t : \Omega \rightarrow \mathbb{R} \mid t \in \mathbb{N})$  such that  $\mathbb{E}(|X_t|) < \infty$  and

$$\mathbb{P}(X_t \in B \mid X_0, \dots, X_{t-1}) = P_{A_t}(B)$$

almost surely for every  $t \in \mathbb{N}^+$  and  $B \in \mathcal{B}(\mathbb{R})$ , where  $A_t = \pi_t(X_0, \dots, X_{t-1})$ . Additionally, if a function  $h : \mathbb{R} \rightarrow \mathbb{R}$  is  $\nu$ -integrable, then  $\mathbb{E}(|h(X_t)|) < \infty$  for every  $t \in \mathbb{N}^+$ .

*Proof.* By Kolmogorov's extension theorem, there is a probability triple  $(\Omega, \mathcal{F}, \mathbb{P})$  carrying a countable set of independent random variables  $\{Z_{t,a} : \Omega \rightarrow \mathbb{R} \mid t \in \mathbb{N}^+ \text{ and } a \in \mathcal{A}\}$  such that  $\mathbb{P}(Z_{t,a} \in B) = P_a(B)$  for every  $t \in \mathbb{N}^+$ ,  $a \in \mathcal{A}$ , and  $B \in \mathcal{B}(\mathbb{R})$ . For every  $t \in \mathbb{N}^+$ , let  $A_t : \Omega \rightarrow \mathcal{A}$  and  $X_t : \Omega \rightarrow \mathbb{R}$  be given by

$$\begin{aligned} A_t(\omega) &= \pi_t(X_0(\omega), \dots, X_{t-1}(\omega)), \\ X_t(\omega) &= Z_{t, A_t(\omega)}(\omega) = \sum_a \mathbb{I}_{\{A_t=a\}}(\omega) Z_{t,a}(\omega), \end{aligned}$$

where  $X_0 : \Omega \rightarrow \mathbb{R}$  is given by  $X_0(\omega) = 0$ .

For every  $t \in \mathbb{N}^+$ , let  $\mathcal{F}_{t-1} = \sigma\left(\bigcup_{k < t, a} \sigma(Z_{k,a})\right)$ . For every  $t \in \mathbb{N}^+$  and  $a \in \mathcal{A}$ , note that  $\sigma(\mathbb{I}_{\{A_t=a\}}) \subseteq \sigma(A_t) \subseteq \sigma(X_0, \dots, X_{t-1}) \subseteq \mathcal{F}_{t-1}$ . Because  $\mathcal{F}_{t-1}$  and  $\sigma(Z_{t,a})$  are independent, so are  $\mathbb{I}_{\{A_t=a\}}$  and  $Z_{t,a}$ .

Therefore, if a function  $h : \mathbb{R} \rightarrow \mathbb{R}$  is  $\nu$ -integrable, then  $\mathbb{E}(|h(X_t)|) < \infty$  for every  $t \in \mathbb{N}^+$ , since

$$\mathbb{E}(|h(X_t)|) = \sum_a \mathbb{E}(\mathbb{I}_{\{A_t=a\}} |h(Z_{t,a})|) = \sum_a \mathbb{E}(\mathbb{I}_{\{A_t=a\}}) \mathbb{E}(|h(Z_{t,a})|) = \sum_a \mathbb{P}(A_t = a) \int_{\mathbb{R}} |h(x)| P_a(dx) \leq c < \infty.$$

In particular, because the identity function is  $\nu$ -integrable,  $\mathbb{E}(|X_t|) < \infty$  for every  $t \in \mathbb{N}^+$ .

By definition, almost surely for every  $t \in \mathbb{N}^+$  and  $B \in \mathcal{B}(\mathbb{R})$ ,

$$\mathbb{P}(X_t \in B \mid X_0, \dots, X_{t-1}) = \mathbb{E}(\mathbb{I}_{\{X_t \in B\}} \mid \sigma(X_0, \dots, X_{t-1})).$$

For every  $t \in \mathbb{N}^+$  and  $B \in \mathcal{B}(\mathbb{R})$ , note that  $\{X_t \in B\} = \bigcup_a \{A_t = a\} \cap \{Z_{t,a} \in B\}$ . Therefore, almost surely,

$$\mathbb{P}(X_t \in B \mid X_0, \dots, X_{t-1}) = \sum_a \mathbb{E}(\mathbb{I}_{\{A_t=a\}} \mathbb{I}_{\{Z_{t,a} \in B\}} \mid \sigma(X_0, \dots, X_{t-1})).$$

For every  $t \in \mathbb{N}^+$  and  $a \in \mathcal{A}$ , recall that  $\mathbb{I}_{\{A_t=a\}}$  is  $\sigma(X_0, \dots, X_{t-1})$ -measurable. Therefore, almost surely,

$$\mathbb{P}(X_t \in B \mid X_0, \dots, X_{t-1}) = \sum_a \mathbb{I}_{\{A_t=a\}} \mathbb{E}(\mathbb{I}_{\{Z_{t,a} \in B\}} \mid \sigma(X_0, \dots, X_{t-1})).$$

Since  $\sigma(X_0, \dots, X_{t-1}) \subseteq \mathcal{F}_{t-1}$  and  $\sigma(\mathbb{I}_{\{Z_{t,a} \in B\}}) \subseteq \sigma(Z_{t,a})$  are independent, almost surely,

$$\mathbb{P}(X_t \in B \mid X_0, \dots, X_{t-1}) = \sum_a \mathbb{I}_{\{A_t=a\}} \mathbb{E}(\mathbb{I}_{\{Z_{t,a} \in B\}}) = \sum_a \mathbb{I}_{\{A_t=a\}} P_a(B) = P_{A_t}(B).$$

$\square$

**Definition 4.5.** The canonical space  $(\Omega, \mathcal{F})$  that carries the reward process  $X = (X_t \mid t \in \mathbb{N})$  is a measurable space such that  $\Omega = \mathbb{R}^\infty$ . Furthermore, for every  $t \in \mathbb{N}$ , the function  $X_t : \Omega \rightarrow \mathbb{R}$  is given by  $X_t(\omega) = \omega_t$  and the  $\sigma$ -algebra  $\mathcal{F}$  on  $\Omega$  is given by  $\mathcal{F} = \sigma(X_0, X_1, \dots)$ .

**Theorem 4.1.** For every set of actions  $\mathcal{A}$ , stochastic bandit  $\nu = (P_a \mid a \in \mathcal{A})$ , and policy  $\pi = (\pi_t \mid t \in \mathbb{N}^+)$ , there is a probability measure  $\mathbb{P}^{\nu, \pi}$  on the canonical space  $(\Omega, \mathcal{F})$  that carries the reward process  $X = (X_t \mid t \in \mathbb{N})$  such that  $\mathbb{E}^{\nu, \pi}(|X_t|) < \infty$  and

$$\mathbb{P}^{\nu, \pi}(X_t \in B \mid X_0, \dots, X_{t-1}) = P_{A_t}(B)$$

almost surely for every  $t \in \mathbb{N}^+$  and  $B \in \mathcal{B}(\mathbb{R})$ , where  $A_t = \pi_t(X_0, \dots, X_{t-1})$ . Additionally, if a function  $h : \mathbb{R} \rightarrow \mathbb{R}$  is  $\nu$ -integrable, then  $\mathbb{E}^{\nu, \pi}(|h(X_t)|) < \infty$  for every  $t \in \mathbb{N}^+$ . The probability triple  $(\Omega, \mathcal{F}, \mathbb{P}^{\nu, \pi})$  is called a canonical triple for the stochastic bandit  $\nu$  under the policy  $\pi$ .

*Proof.* Proposition 4.2 ensures that there is a probability triple  $(\tilde{\Omega}^{\nu, \pi}, \tilde{\mathcal{F}}^{\nu, \pi}, \tilde{\mathbb{P}}^{\nu, \pi})$  carrying a stochastic process  $(\tilde{X}_t^{\nu, \pi} : \tilde{\Omega}^{\nu, \pi} \rightarrow \mathbb{R} \mid t \in \mathbb{N})$  such that, almost surely,

$$\tilde{\mathbb{P}}^{\nu, \pi}(\tilde{X}_t^{\nu, \pi} \in B \mid \tilde{X}_0^{\nu, \pi}, \dots, \tilde{X}_{t-1}^{\nu, \pi}) = P_{\tilde{A}_t}(B)$$

for every  $t \in \mathbb{N}^+$  and  $B \in \mathcal{B}(\mathbb{R})$ , where  $\tilde{A}_t = \pi_t(\tilde{X}_0^{\nu, \pi}, \dots, \tilde{X}_{t-1}^{\nu, \pi})$ .

Consider the function  $\tilde{X}^{\nu, \pi} : \tilde{\Omega}^{\nu, \pi} \rightarrow \Omega$  given by  $\tilde{X}^{\nu, \pi}(\tilde{\omega}) = (\tilde{X}_t^{\nu, \pi}(\tilde{\omega}) \mid t \in \mathbb{N})$ . The function  $\tilde{X}^{\nu, \pi}$  is  $\tilde{\mathcal{F}}^{\nu, \pi}/\mathcal{F}$ -measurable, so that the function  $\mathbb{P}^{\nu, \pi} : \mathcal{F} \rightarrow [0, 1]$  defined by

$$\mathbb{P}^{\nu, \pi}(F) = \tilde{\mathbb{P}}^{\nu, \pi}\left(\left(\tilde{X}^{\nu, \pi}\right)^{-1}(F)\right) = \tilde{\mathbb{P}}^{\nu, \pi}\left(\{\tilde{\omega} \in \tilde{\Omega}^{\nu, \pi} \mid \tilde{X}^{\nu, \pi}(\tilde{\omega}) \in F\}\right)$$

is a probability measure on the measurable space  $(\Omega, \mathcal{F})$ .

In order to show that  $\tilde{X}^{\nu, \pi}$  is  $\sigma(\tilde{X}_0^{\nu, \pi}, \dots, \tilde{X}_t^{\nu, \pi})/\sigma(X_0, \dots, X_t)$ -measurable for every  $t \in \mathbb{N}^+$ , let  $\mathcal{I}_t$  be given by

$$\mathcal{I}_t = \left\{ \bigcap_{k=0}^t \{X_k \in B_k\} \mid B_k \in \mathcal{B}(\mathbb{R}) \text{ for every } k \in \{0, \dots, t\} \right\},$$

so that  $\mathcal{I}_t$  is a  $\pi$ -system on  $\Omega$  such that  $\sigma(\mathcal{I}_t) = \sigma(X_0, \dots, X_t)$ . For every  $t \in \mathbb{N}^+$  and  $I_t \in \mathcal{I}_t$ ,

$$(\tilde{X}^{\nu, \pi})^{-1}(I_t) = (\tilde{X}^{\nu, \pi})^{-1}\left(\bigcap_{k=0}^t \{X_k \in B_k\}\right) = \bigcap_{k=0}^t (\tilde{X}^{\nu, \pi})^{-1}(\{X_k \in B_k\}) = \bigcap_{k=0}^t \{\tilde{X}_k^{\nu, \pi} \in B_k\},$$

which uses the fact that

$$(\tilde{X}^{\nu, \pi})^{-1}(\{X_k \in B_k\}) = \left\{ \tilde{\omega} \in \tilde{\Omega}^{\nu, \pi} \mid \tilde{X}^{\nu, \pi}(\tilde{\omega}) \in \{\omega \in \Omega \mid \omega_k \in B_k\} \right\} = \{\tilde{X}_k^{\nu, \pi} \in B_k\}.$$

Since  $(\tilde{X}^{\nu, \pi})^{-1}(I_t) \in \sigma(\tilde{X}_0^{\nu, \pi}, \dots, \tilde{X}_t^{\nu, \pi})$  for every  $I_t \in \mathcal{I}_t$ ,  $\tilde{X}^{\nu, \pi}$  is  $\sigma(\tilde{X}_0^{\nu, \pi}, \dots, \tilde{X}_t^{\nu, \pi})/\sigma(X_0, \dots, X_t)$ -measurable. For every  $t \in \mathbb{N}^+$  and  $H_{t-1} \in \sigma(X_0, \dots, X_{t-1})$ , let  $\tilde{H}_{t-1} = (\tilde{X}^{\nu, \pi})^{-1}(H_{t-1})$ . For every  $B \in \mathcal{B}(\mathbb{R})$ ,

$$\mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{X_t \in B\}} \mathbb{I}_{H_{t-1}}) = \mathbb{P}^{\nu, \pi}(\{X_t \in B\} \cap H_{t-1}) = \tilde{\mathbb{P}}^{\nu, \pi}\left((\tilde{X}^{\nu, \pi})^{-1}(\{X_t \in B\}) \cap (\tilde{X}^{\nu, \pi})^{-1}(H_{t-1})\right).$$

Because  $\tilde{H}_{t-1} \in \sigma(\tilde{X}_0^{\nu, \pi}, \dots, \tilde{X}_{t-1}^{\nu, \pi})$ ,

$$\mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{X_t \in B\}} \mathbb{I}_{H_{t-1}}) = \tilde{\mathbb{P}}^{\nu, \pi}\left(\{\tilde{X}_t^{\nu, \pi} \in B\} \cap \tilde{H}_{t-1}\right) = \tilde{\mathbb{E}}^{\nu, \pi}\left(\mathbb{I}_{\{\tilde{X}_t^{\nu, \pi} \in B\}} \mathbb{I}_{\tilde{H}_{t-1}}\right) = \tilde{\mathbb{E}}^{\nu, \pi}\left(P_{\tilde{A}_t}(B) \mathbb{I}_{\tilde{H}_{t-1}}\right),$$

where  $\tilde{A}_t = \pi_t(\tilde{X}_0^{\nu, \pi}, \dots, \tilde{X}_{t-1}^{\nu, \pi})$ . Therefore,

$$\mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{X_t \in B\}} \mathbb{I}_{H_{t-1}}) = \tilde{\mathbb{E}}^{\nu, \pi}\left(\sum_a \mathbb{I}_{\{\tilde{A}_t = a\}} P_a(B) \mathbb{I}_{\tilde{H}_{t-1}}\right) = \sum_a P_a(B) \tilde{\mathbb{P}}^{\nu, \pi}\left(\{\tilde{A}_t = a\} \cap \tilde{H}_{t-1}\right).$$

For every  $a \in \mathcal{A}$ , note that  $\mathbb{P}^{\nu, \pi}(\{A_t = a\} \cap H_{t-1})$  is given by

$$\mathbb{P}^{\nu, \pi}(\{A_t = a\} \cap H_{t-1}) = \tilde{\mathbb{P}}^{\nu, \pi}\left((\tilde{X}^{\nu, \pi})^{-1}(\{A_t = a\}) \cap (\tilde{X}^{\nu, \pi})^{-1}(H_{t-1})\right) = \tilde{\mathbb{P}}^{\nu, \pi}\left(\{\tilde{A}_t = a\} \cap \tilde{H}_{t-1}\right),$$

which uses the fact that

$$(\tilde{X}^{\nu,\pi})^{-1}(\{A_t = a\}) = \{\tilde{\omega} \in \tilde{\Omega}^{\nu,\pi} \mid \tilde{X}^{\nu,\pi}(\tilde{\omega}) \in \{\omega \in \Omega \mid \pi_t(\omega_0, \dots, \omega_{t-1}) = a\}\} = \{\tilde{A}_t = a\}.$$

Finally, for every  $t \in \mathbb{N}^+$ ,  $H_{t-1} \in \sigma(X_0, \dots, X_{t-1})$ ,  $B \in \mathcal{B}(\mathbb{R})$ ,

$$\mathbb{E}^{\nu,\pi}(\mathbb{I}_{\{X_t \in B\}} \mathbb{I}_{H_{t-1}}) = \sum_a P_a(B) \mathbb{P}^{\nu,\pi}(\{A_t = a\} \cap H_{t-1}) = \mathbb{E}^{\nu,\pi}(P_{A_t}(B) \mathbb{I}_{H_{t-1}}).$$

Because  $P_{A_t}(B)$  is  $\sigma(X_0, \dots, X_{t-1})$ -measurable, almost surely,

$$\mathbb{P}^{\nu,\pi}(X_t \in B \mid X_0, \dots, X_{t-1}) = \mathbb{E}^{\nu,\pi}(\mathbb{I}_{\{X_t \in B\}} \mid \sigma(X_0, \dots, X_{t-1})) = P_{A_t}(B).$$

For every  $t \in \mathbb{N}^+$ , consider the law  $\mathcal{L}_t : \mathcal{B}(\mathbb{R}) \rightarrow [0, 1]$  given by

$$\mathcal{L}_t(B) = \mathbb{P}^{\nu,\pi}(X_t \in B) = \tilde{\mathbb{P}}^{\nu,\pi}((\tilde{X}^{\nu,\pi})^{-1}(\{X_t \in B\})) = \tilde{\mathbb{P}}^{\nu,\pi}(\tilde{X}_t^{\nu,\pi} \in B).$$

If a function  $h : \mathbb{R} \rightarrow \mathbb{R}$  is  $\nu$ -integrable, then  $\mathbb{E}^{\nu,\pi}(|h(X_t)|) < \infty$  for every  $t \in \mathbb{N}^+$ , since

$$\mathbb{E}^{\nu,\pi}(|h(X_t)|) = \int_{\mathbb{R}} |h(x)| \mathcal{L}_t(dx) = \tilde{\mathbb{E}}^{\nu,\pi}(|h(\tilde{X}_t^{\nu,\pi})|) < \infty.$$

In particular, because the identity function is  $\nu$ -integrable,  $\mathbb{E}^{\nu,\pi}(|X_t|) < \infty$  for every  $t \in \mathbb{N}^+$ . □

For the remaining, consider a set of actions  $\mathcal{A}$ , a stochastic bandit  $\nu = (P_a \mid a \in \mathcal{A})$ , a policy  $\pi = (\pi_t \mid t \in \mathbb{N}^+)$ , and let  $(\Omega, \mathcal{F}, \mathbb{P}^{\nu,\pi})$  be a canonical triple for the stochastic bandit  $\nu$  under the policy  $\pi$ .

**Proposition 4.3.** For every  $t \in \mathbb{N}^+$ , if a function  $h : \mathbb{R} \rightarrow \mathbb{R}$  is  $\nu$ -integrable, then

$$\mathbb{E}^{\nu,\pi}(h(X_t) \mid X_0, \dots, X_{t-1}) = \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} h(x) P_a(dx)$$

almost surely, where  $A_t = \pi_t(X_0, \dots, X_{t-1})$ .

*Proof.* Since the function  $h : \mathbb{R} \rightarrow \mathbb{R}$  is  $\nu$ -integrable, recall that  $\mathbb{E}^{\nu,\pi}(|h(X_t)|) < \infty$  for every  $t \in \mathbb{N}^+$ .

First, suppose that  $h = \mathbb{I}_B$  for some  $B \in \mathcal{B}(\mathbb{R})$ . Because  $\mathbb{I}_B(X_t) = \mathbb{I}_{\{X_t \in B\}}$ , almost surely,

$$\mathbb{E}^{\nu,\pi}(\mathbb{I}_B(X_t) \mid X_0, \dots, X_{t-1}) = P_{A_t}(B) = \sum_a \mathbb{I}_{\{A_t=a\}} P_a(B) = \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} \mathbb{I}_B(x) P_a(dx).$$

Next, suppose that  $h$  is a simple function that can be written as  $h = \sum_{k=1}^m b_k \mathbb{I}_{B_k}$  for some fixed  $b_1, b_2, \dots, b_m \in [0, \infty]$  and  $B_1, B_2, \dots, B_m \in \mathcal{B}(\mathbb{R})$ . Almost surely,

$$\mathbb{E}^{\nu,\pi}\left(\sum_{k=1}^m b_k \mathbb{I}_{B_k}(X_t) \mid X_0, \dots, X_{t-1}\right) = \sum_{k=1}^m b_k \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} \mathbb{I}_{B_k}(x) P_a(dx) = \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} \sum_{k=1}^m b_k \mathbb{I}_{B_k}(x) P_a(dx).$$

Next, suppose that  $h$  is a non-negative  $\mathcal{B}(\mathbb{R})$ -measurable function. For any  $k \in \mathbb{N}$ , consider the simple function  $h_k = \alpha_k \circ h$ , where  $\alpha_k$  is the  $k$ -th staircase function. Almost surely, since  $h_k(X_t) \uparrow h(X_t)$ ,

$$\mathbb{E}^{\nu,\pi}(h(X_t) \mid X_0, \dots, X_{t-1}) = \mathbb{E}^{\nu,\pi}\left(\lim_{k \rightarrow \infty} h_k(X_t) \mid X_0, \dots, X_{t-1}\right) = \lim_{k \rightarrow \infty} \mathbb{E}^{\nu,\pi}(h_k(X_t) \mid X_0, \dots, X_{t-1}).$$

Since  $h_k \uparrow h$ , by the monotone-convergence theorem, almost surely,

$$\mathbb{E}^{\nu,\pi}(h(X_t) \mid X_0, \dots, X_{t-1}) = \lim_{k \rightarrow \infty} \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} h_k(x) P_a(dx) = \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} \lim_{k \rightarrow \infty} h_k(x) P_a(dx).$$

Finally, suppose that  $h = h^+ - h^-$  is a  $\mathcal{B}(\mathbb{R})$ -measurable function. Almost surely,

$$\mathbb{E}^{\nu,\pi}(h(X_t) \mid X_0, \dots, X_{t-1}) = \left(\sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} h^+(x) P_a(dx)\right) - \left(\sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} h^-(x) P_a(dx)\right).$$

By the linearity of the integral, almost surely,

$$\mathbb{E}^{\nu, \pi} (h(X_t) \mid X_0, \dots, X_{t-1}) = \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} (h^+(x) - h^-(x)) P_a(dx) = \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} h(x) P_a(dx).$$

□

**Proposition 4.4.** If  $t \in \mathbb{N}^+$  and  $A_t = \pi_t(X_0, \dots, X_{t-1})$ , then  $\mathbb{E}^{\nu, \pi} (X_t \mid A_t) = \mu_{A_t}^{\nu}$  almost surely.

*Proof.* For every  $t \in \mathbb{N}^+$ ,  $\mathbb{E}^{\nu, \pi} (|X_t|) < \infty$  and  $A_t$  is  $\sigma(X_0, \dots, X_{t-1})$ -measurable. Therefore, almost surely,

$$\mathbb{E}^{\nu, \pi} (X_t \mid A_t) = \mathbb{E}^{\nu, \pi} (\mathbb{E}^{\nu, \pi} (X_t \mid X_0, \dots, X_{t-1}) \mid A_t) = \sum_a \mathbb{I}_{\{A_t=a\}} \int_{\mathbb{R}} x P_a(dx) = \sum_a \mathbb{I}_{\{A_t=a\}} \mu_a^{\nu} = \mu_{A_t}^{\nu},$$

by the tower property, Proposition 4.3 applied to the identity function, and taking out what is known. □

**Proposition 4.5.** If  $t \in \mathbb{N}^+$  and  $A_t = \pi_t(X_0, \dots, X_{t-1})$ , then

$$\mathbb{E}^{\nu, \pi} (X_t) = \mathbb{E}^{\nu, \pi} (\mathbb{E}^{\nu, \pi} (X_t \mid A_t)) = \mathbb{E}^{\nu, \pi} (\mu_{A_t}^{\nu}) = \sum_a \mu_a^{\nu} \mathbb{P}^{\nu, \pi} (A_t = a).$$

**Definition 4.6.** For every  $t \in \mathbb{N}^+$ , the total reward  $S_t$  after  $t$  time steps is given by  $S_t = \sum_{k=1}^t X_k$ .

**Definition 4.7.** For every  $t \in \mathbb{N}^+$ , the regret  $R_t^{\nu, \pi}$  of policy  $\pi$  on  $\nu$  after  $t$  time steps is given by

$$R_t^{\nu, \pi} = t\mu_*^{\nu} - \sum_{k=1}^t \mathbb{E}^{\nu, \pi} (X_k).$$

**Definition 4.8.** For every action  $a \in \mathcal{A}$ , the suboptimality gap is defined by  $\Delta_a^{\nu} = \mu_*^{\nu} - \mu_a^{\nu}$ , so that  $\Delta_a^{\nu} \geq 0$ .

**Definition 4.9.** The number of times  $T_{t,a}^{\pi} : \Omega \rightarrow \{0, \dots, t\}$  that policy  $\pi$  selects  $a \in \mathcal{A}$  by time  $t \in \mathbb{N}^+$  is given by

$$T_{t,a}^{\pi}(\omega) = \sum_{k=1}^t \mathbb{I}_{\{A_k=a\}}(\omega),$$

where  $A_k = \pi_k(X_0, \dots, X_{k-1})$  for every  $k \leq t$ . Note that  $\sum_a T_{t,a}^{\pi}(\omega) = t$  for every  $\omega \in \Omega$ .

**Definition 4.10.** The average reward  $M_{t,a}^{\pi} : \Omega \rightarrow \mathbb{R}$  that policy  $\pi$  observes for  $a \in \mathcal{A}$  by time  $t \in \mathbb{N}^+$  is given by

$$M_{t,a}^{\pi}(\omega) = \frac{1}{T_{t,a}^{\pi}(\omega)} \sum_{k=1}^t X_k(\omega) \mathbb{I}_{\{A_k=a\}}(\omega)$$

whenever  $T_{t,a}^{\pi}(\omega) > 0$ , where  $A_k = \pi_k(X_0, \dots, X_{k-1})$  for every  $k \leq t$ .

**Theorem 4.2.** For every  $t \in \mathbb{N}^+$ , the regret  $R_t^{\nu, \pi}$  of policy  $\pi$  on  $\nu$  after  $t$  time steps is given by

$$R_t^{\nu, \pi} = \sum_a \Delta_a^{\nu} \mathbb{E}^{\nu, \pi} (T_{t,a}^{\pi}).$$

*Proof.* For every  $t \in \mathbb{N}^+$ , let  $A_k = \pi_k(X_0, \dots, X_{k-1})$  for every  $k \leq t$ , so that  $\mathbb{E}^{\nu, \pi} (T_{t,a}^{\pi}) = \sum_{k=1}^t \mathbb{P}^{\nu, \pi} (A_k = a)$  and

$$\sum_a \mathbb{E}^{\nu, \pi} (T_{t,a}^{\pi}) = \sum_a \sum_{k=1}^t \mathbb{P}^{\nu, \pi} (A_k = a) = \sum_{k=1}^t \sum_a \mathbb{P}^{\nu, \pi} (A_k = a) = t.$$

By the definition of the regret  $R_t^{\nu, \pi}$  of policy  $\pi$  on  $\nu$  after  $t$  time steps,

$$R_t^{\nu, \pi} = t\mu_*^{\nu} - \sum_{k=1}^t \mathbb{E}^{\nu, \pi} (X_k) = \sum_{k=1}^t \sum_a \mu_a^{\nu} \mathbb{P}^{\nu, \pi} (A_k = a) - \sum_{k=1}^t \sum_a \mu_a^{\nu} \mathbb{P}^{\nu, \pi} (A_k = a).$$

By rearranging terms and the definition of suboptimality gap,

$$R_t^{\nu, \pi} = \sum_{k=1}^t \sum_a (\mu_*^{\nu} - \mu_a^{\nu}) \mathbb{P}^{\nu, \pi} (A_k = a) = \sum_a \Delta_a^{\nu} \sum_{k=1}^t \mathbb{P}^{\nu, \pi} (A_k = a) = \sum_a \Delta_a^{\nu} \mathbb{E}^{\nu, \pi} (T_{t,a}^{\pi}).$$

□



**Proposition 4.6.** If  $t \in \mathbb{N}^+$ , then  $R_t^{\nu, \pi} \geq 0$ .

*Proof.* Since  $\Delta_a^\nu \geq 0$  and  $\mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \geq 0$  for every  $a \in \mathcal{A}$  and  $t \in \mathbb{N}^+$ , the claim is a consequence of Theorem 4.2.  $\square$

**Proposition 4.7.** Consider an action  $a^* \in \mathcal{A}$  such that  $\mu_{a^*}^\nu = \mu_*^\nu$ . If  $\pi_t = a^*$  for every  $t \in \mathbb{N}^+$ , then  $R_t^{\nu, \pi} = 0$ .

*Proof.* For every  $t \in \mathbb{N}^+$ , note that  $T_{t,a}^\pi = 0$  for every  $a \neq a^*$ . Therefore,

$$R_t^{\nu, \pi} = \sum_a \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) = \Delta_{a^*}^\nu \mathbb{E}^{\nu, \pi}(T_{t,a^*}^\pi) = (\mu_*^\nu - \mu_{a^*}^\nu) \mathbb{E}^{\nu, \pi}(T_{t,a^*}^\pi) = 0.$$

$\square$

**Proposition 4.8.** For every  $t \in \mathbb{N}^+$ , let  $A_k = \pi_k(X_0, \dots, X_{k-1})$  for every  $k \leq t$ . If  $R_t^{\nu, \pi} = 0$ , then  $\mu_{A_k}^\nu = \mu_*^\nu$  almost surely for every  $k \leq t$ .

*Proof.* For every  $t \in \mathbb{N}^+$ , by Theorem 4.2,

$$R_t^{\nu, \pi} = \sum_a \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) = \sum_a \Delta_a^\nu \sum_{k=1}^t \mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{A_k=a\}}) = \sum_{k=1}^t \mathbb{E}^{\nu, \pi} \left( \sum_a \mathbb{I}_{\{A_k=a\}} \Delta_a^\nu \right) = \sum_{k=1}^t \mathbb{E}^{\nu, \pi}(\Delta_{A_k}^\nu).$$

Suppose that  $\mathbb{P}^{\nu, \pi}(\mu_{A_k}^\nu = \mu_*^\nu) < 1$  for some  $k \leq t$ , so that  $\mathbb{P}^{\nu, \pi}(\mu_{A_k}^\nu < \mu_*^\nu) > 0$  and  $\mathbb{P}^{\nu, \pi}(\Delta_{A_k}^\nu > 0) > 0$ . In that case,  $\mathbb{E}^{\nu, \pi}(\Delta_{A_k}^\nu) > 0$ , so that  $R_t^{\nu, \pi} > 0$ .  $\square$

For convenience, let  $R_0^{\nu, \pi} = 0$ .

**Proposition 4.9.** If  $R_t^{\nu, \pi} = o(t)$ , then

$$\mu_*^\nu = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{k=1}^t \mathbb{E}^{\nu, \pi}(X_k).$$

*Proof.* Since  $R_t^{\nu, \pi} : \mathbb{N} \rightarrow \mathbb{R}$  is asymptotically positive by assumption,

$$0 = \limsup_{t \rightarrow \infty} \frac{R_t^{\nu, \pi}}{t} \geq \liminf_{t \rightarrow \infty} \frac{R_t^{\nu, \pi}}{t} \geq 0,$$

so that

$$0 = \lim_{t \rightarrow \infty} \frac{R_t^{\nu, \pi}}{t} = \lim_{t \rightarrow \infty} \mu_*^\nu - \frac{1}{t} \sum_{k=1}^t \mathbb{E}^{\nu, \pi}(X_k) = \mu_*^\nu - \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{k=1}^t \mathbb{E}^{\nu, \pi}(X_k).$$

$\square$

**Definition 4.11.** The number of times  $T_{t,*}^{\nu, \pi} : \Omega \rightarrow \{0, \dots, t\}$  that policy  $\pi$  selects an optimal action on the stochastic bandit  $\nu$  by time step  $t \in \mathbb{N}^+$  is given by

$$T_{t,*}^{\nu, \pi}(\omega) = \sum_{k=1}^t \mathbb{I}_{\{\mu_{A_k}^\nu = \mu_*^\nu\}}(\omega) = \sum_{k=1}^t \mathbb{I}_{\{\Delta_{A_k}^\nu = 0\}}(\omega),$$

where  $A_k = \pi_k(X_0, \dots, X_{k-1})$  for every  $k \leq t$ .

**Proposition 4.10.** The number of times  $T_{t,*}^{\nu, \pi} : \Omega \rightarrow \{0, \dots, t\}$  that policy  $\pi$  selects an optimal action on the stochastic bandit  $\nu$  by time step  $t \in \mathbb{N}^+$  is given by

$$T_{t,*}^{\nu, \pi}(\omega) = \sum_{a | \Delta_a^\nu = 0} T_{t,a}^\pi(\omega).$$

*Proof.* For every  $t \in \mathbb{N}^+$ , let  $A_k = \pi_k(X_0, \dots, X_{k-1})$  for every  $k \leq t$ . In that case,

$$\{\Delta_{A_k}^\nu = 0\} = \bigcup_a \{A_k = a \text{ and } \Delta_a^\nu = 0\} = \bigcup_{a|\Delta_a^\nu=0} \{A_k = a\},$$

so that

$$T_{t,*}^{\nu,\pi}(\omega) = \sum_{k=1}^t \mathbb{I}_{\{\Delta_{A_k}^\nu=0\}}(\omega) = \sum_{k=1}^t \sum_{a|\Delta_a^\nu=0} \mathbb{I}_{\{A_k=a\}}(\omega) = \sum_{a|\Delta_a^\nu=0} \sum_{k=1}^t \mathbb{I}_{\{A_k=a\}}(\omega) = \sum_{a|\Delta_a^\nu=0} T_{t,a}^\pi(\omega).$$

□

**Proposition 4.11.** If the set of actions  $\mathcal{A}$  is finite and  $R_t^{\nu,\pi} = o(t)$ , then

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi})}{t} = 1.$$

*Proof.* By Theorem 4.2,

$$0 = \lim_{t \rightarrow \infty} \frac{R_t^{\nu,\pi}}{t} = \lim_{t \rightarrow \infty} \frac{\sum_a \Delta_a^\nu \mathbb{E}^{\nu,\pi}(T_{t,a}^\pi)}{t} = \lim_{t \rightarrow \infty} \sum_a \Delta_a^\nu \frac{\mathbb{E}^{\nu,\pi}(T_{t,a}^\pi)}{t} = \sum_a \Delta_a^\nu \lim_{t \rightarrow \infty} \frac{\mathbb{E}^{\nu,\pi}(T_{t,a}^\pi)}{t},$$

so that  $\lim_{t \rightarrow \infty} \mathbb{E}^{\nu,\pi}(T_{t,a}^\pi)/t = 0$  whenever  $\Delta_a^\nu > 0$ . Therefore,

$$0 = \sum_{a|\Delta_a^\nu>0} \lim_{t \rightarrow \infty} \frac{\mathbb{E}^{\nu,\pi}(T_{t,a}^\pi)}{t} = \lim_{t \rightarrow \infty} \sum_{a|\Delta_a^\nu>0} \frac{\mathbb{E}^{\nu,\pi}(T_{t,a}^\pi)}{t}.$$

For every  $t \in \mathbb{N}^+$ , recall that  $\sum_a T_{t,a}^\pi = t$ . By Proposition 4.10,

$$t = \sum_a \mathbb{E}^{\nu,\pi}(T_{t,a}^\pi) = \sum_{a|\Delta_a^\nu=0} \mathbb{E}^{\nu,\pi}(T_{t,a}^\pi) + \sum_{a|\Delta_a^\nu>0} \mathbb{E}^{\nu,\pi}(T_{t,a}^\pi) = \mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi}) + \sum_{a|\Delta_a^\nu>0} \mathbb{E}^{\nu,\pi}(T_{t,a}^\pi),$$

so that

$$\sum_{a|\Delta_a^\nu>0} \frac{\mathbb{E}^{\nu,\pi}(T_{t,a}^\pi)}{t} = 1 - \frac{\mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi})}{t}.$$

Therefore, considering a previous equation,

$$0 = \lim_{t \rightarrow \infty} 1 - \frac{\mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi})}{t} = 1 - \lim_{t \rightarrow \infty} \frac{\mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi})}{t}.$$

Since  $\mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi}) > 0$  for some  $t \in \mathbb{N}^+$  and  $\mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi}) \leq \mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi})$ , note that  $\mathbb{E}^{\nu,\pi}(T_{t,*}^{\nu,\pi}) = \Theta(t)$ . □

**Definition 4.12.** For a set of actions  $\mathcal{A}$ , an environment class  $\mathcal{E}$  is a set of stochastic bandits for  $\mathcal{A}$ .

**Definition 4.13.** For a set of actions  $\mathcal{A}$  and an environment class  $\mathcal{E}$ , consider a probability triple  $(\mathcal{E}, \mathcal{G}, \mathbb{Q})$  such that  $R_t^{\nu,\pi} : \mathcal{E} \rightarrow [0, \infty]$  is  $\mathcal{G}$ -measurable for every policy  $\pi$  and time step  $t \in \mathbb{N}^+$ . The Bayesian regret  $B_t^\pi$  of policy  $\pi$  after  $t \in \mathbb{N}^+$  time steps is given by

$$B_t^\pi = \int_{\mathcal{E}} R_t^{\nu,\pi} Q(d\nu).$$

**Definition 4.14.** The stochastic bandit  $\nu = (P_a \mid a \in \mathcal{A})$  is  $\sigma$ -subgaussian if, for every  $a \in \mathcal{A}$ , the random variable  $Z_a$  on the probability triple  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_a)$  given by  $Z_a(x) = x - \mu_a^\nu$  is  $\sigma$ -subgaussian. Note that  $\mathbb{E}_a(Z_a) = 0$ .

## 5 Explore-then-commit

**Definition 5.1.** If  $(x_n \in \mathbb{R} \mid n \in \mathbb{N})$  is a sequence of real numbers, then  $\arg \max_n x_n$  is given by

$$\arg \max_n x_n = \inf(\{m \in \mathbb{N} \mid x_m = \sup_n x_n\}).$$

Note that  $\arg \max_n x_n \in \mathbb{N} \cup \{\infty\}$ , since  $\inf(\emptyset) = \infty$ .

Consider a measurable space  $(\Omega, \mathcal{F})$  and a stochastic process  $(Y_n : \Omega \rightarrow \mathbb{R} \mid n \in \mathbb{N})$ .

**Definition 5.2.** The function  $\arg \max_n Y_n : \Omega \rightarrow \mathbb{N} \cup \{\infty\}$  is given by

$$\left( \arg \max_n Y_n \right) (\omega) = \arg \max_n Y_n(\omega).$$

**Proposition 5.1.** The function  $\arg \max_n Y_n : \Omega \rightarrow \mathbb{N} \cup \{\infty\}$  is  $\mathcal{F}$ -measurable.

*Proof.* Recall that the function  $\sup_n Y_n$  is  $\mathcal{F}$ -measurable, so that the function  $Z_m : \Omega \rightarrow \mathbb{N} \cup \{\infty\}$  given by

$$Z_m(\omega) = m \mathbb{I}_{\{Y_m = \sup_n Y_n\}}(\omega) + \infty \mathbb{I}_{\{Y_m \neq \sup_n Y_n\}}(\omega) = \begin{cases} m, & \text{if } Y_m(\omega) = \sup_n Y_n(\omega), \\ \infty, & \text{if } Y_m(\omega) \neq \sup_n Y_n(\omega) \end{cases}$$

is  $\mathcal{F}$ -measurable for every  $m \in \mathbb{N}$ . Furthermore, recall that the function  $\inf_m Z_m$  is  $\mathcal{F}$ -measurable and note that

$$\inf_m Z_m(\omega) = \inf \left( \left\{ m \in \mathbb{N} \mid Y_m(\omega) = \sup_n Y_n(\omega) \right\} \right) = \arg \max_n Y_n(\omega) = \left( \arg \max_n Y_n \right) (\omega).$$

□

Consider a number of actions  $n \in \mathbb{N}^+$ , a set of actions  $\mathcal{A} = \{1, \dots, n\}$ , a stochastic bandit  $\nu = (P_a \mid a \in \mathcal{A})$ , a policy  $\pi = (\pi_t \mid t \in \mathbb{N}^+)$ , and let  $(\Omega, \mathcal{F}, \mathbb{P}^{\nu, \pi})$  be a canonical triple for the stochastic bandit  $\nu$  under the policy  $\pi$ .

**Definition 5.3.** A policy  $\pi$  implements explore-then-commit with  $m \in \mathbb{N}^+$  exploration steps if, for every  $t \in \mathbb{N}^+$ ,

$$\pi_t(X_0, \dots, X_{t-1}) = \begin{cases} ((t-1) \bmod n) + 1, & \text{if } t \leq mn, \\ \arg \max_a M_{mn, a}^\pi, & \text{if } t > mn. \end{cases}$$

Note that  $M_{t, a}^\pi$  is well-defined for every  $t \geq n$  and  $a \in \mathcal{A}$ .

**Proposition 5.2.** If the policy  $\pi$  implements explore-then-commit with  $m \in \mathbb{N}^+$  exploration steps and  $t \leq mn$ , then  $\mathbb{P}^{\nu, \pi}(X_t \in B) = P_{a_t}(B)$  for every  $B \in \mathcal{B}(\mathbb{R})$ , where  $a_t = ((t-1) \bmod n) + 1$ .

*Proof.* For every  $t \in \mathbb{N}^+$  such that  $t \leq mn$ , let  $A_t = \pi_t(X_0, \dots, X_{t-1})$ , so that  $A_t = a_t$ . For every  $B \in \mathcal{B}(\mathbb{R})$ ,

$$\mathbb{P}^{\nu, \pi}(X_t \in B) = \mathbb{E}^{\nu, \pi}(\mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{X_t \in B\}} \mid X_0, \dots, X_{t-1})) = \mathbb{E}^{\nu, \pi}(P_{A_t}(B)) = \mathbb{E}^{\nu, \pi}(P_{a_t}(B)) = P_{a_t}(B).$$

□

**Proposition 5.3.** If the policy  $\pi$  implements explore-then-commit with  $m \in \mathbb{N}^+$  exploration steps, then the random variables  $X_0, X_1, \dots, X_{mn}$  are independent in  $(\Omega, \mathcal{F}, \mathbb{P}^{\nu, \pi})$ .

*Proof.* Note that  $X_0$  and  $X_1$  are independent because  $\sigma(X_0) = \{\emptyset, \Omega\}$ . Suppose that  $X_0, X_1, \dots, X_t$  are independent for some  $t \in \mathbb{N}^+$  such that  $t < mn$ . We will show that  $X_0, X_1, \dots, X_{t+1}$  are independent.

For every  $B_0, B_1, \dots, B_{t+1} \in \mathcal{B}(\mathbb{R})$ , by taking out what is known,

$$\mathbb{P}^{\nu, \pi} \left( \bigcap_{k=0}^{t+1} \{X_k \in B_k\} \right) = \mathbb{E}^{\nu, \pi} \left( \prod_{k=0}^{t+1} \mathbb{I}_{\{X_k \in B_k\}} \right) = \mathbb{E}^{\nu, \pi} \left( \prod_{k=0}^t \mathbb{I}_{\{X_k \in B_k\}} \mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{X_{t+1} \in B_{t+1}\}} \mid X_0, \dots, X_t) \right).$$

Let  $a_{t+1} = (t \bmod n) + 1$ , so that  $\pi_{t+1}(X_0, \dots, X_t) = a_{t+1}$ . In that case,

$$\mathbb{P}^{\nu, \pi} \left( \bigcap_{k=0}^{t+1} \{X_k \in B_k\} \right) = \mathbb{E}^{\nu, \pi} \left( \left( \prod_{k=0}^t \mathbb{I}_{\{X_k \in B_k\}} \right) P_{a_{t+1}}(B_{t+1}) \right) = \mathbb{E}^{\nu, \pi} \left( \prod_{k=0}^t \mathbb{I}_{\{X_k \in B_k\}} \right) P_{a_{t+1}}(B_{t+1}).$$

By Proposition 5.2 and because  $X_0, X_1, \dots, X_t$  are independent by assumption,

$$\mathbb{P}^{\nu, \pi} \left( \bigcap_{k=0}^{t+1} \{X_k \in B_k\} \right) = \mathbb{P}^{\nu, \pi} \left( \bigcap_{k=0}^t \{X_k \in B_k\} \right) \mathbb{P}^{\nu, \pi} (X_{t+1} \in B_{t+1}) = \prod_{k=0}^{t+1} \mathbb{P}^{\nu, \pi} (X_k \in B_k).$$

□

**Proposition 5.4.** If the policy  $\pi$  implements explore-then-commit with  $m \in \mathbb{N}^+$  exploration steps and  $\nu$  is a 1-subgaussian stochastic bandit, then  $X_t - \mu_{a_t}^\nu$  is 1-subgaussian for every  $t \leq mn$ , where  $a_t = ((t-1) \bmod n) + 1$ .

*Proof.* For every  $a \in \mathcal{A}$ , recall that the random variable  $Z_a$  on the probability triple  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_a)$  is 1-subgaussian, where  $Z_a(x) = x - \mu_a^\nu$ . By Proposition 5.2, the law of  $X_t$  is  $P_{a_t}$  for every  $t \in \{1, \dots, mn\}$ . For every  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}^{\nu, \pi} \left( e^{\lambda(X_t - \mu_{a_t}^\nu)} \right) = \int_{\mathbb{R}} e^{\lambda(x - \mu_{a_t}^\nu)} P_{a_t}(dx_t) = \int_{\mathbb{R}} e^{\lambda Z_{a_t}(x_t)} P_{a_t}(dx_t) = \mathbb{E}_{a_t} (e^{\lambda Z_{a_t}}) \leq e^{\frac{\lambda^2}{2}}.$$

□

**Theorem 5.1.** If the policy  $\pi$  implements explore-then-commit with  $m \in \mathbb{N}^+$  exploration steps and  $\nu$  is a 1-subgaussian stochastic bandit, for every  $t \in \mathbb{N}^+$  such that  $t \geq mn$ ,

$$R_t^{\nu, \pi} \leq \left( m \sum_{a=1}^n \Delta_a^\nu \right) + (t - mn) \sum_{a=1}^n \Delta_a^\nu e^{-\frac{m(\Delta_a^\nu)^2}{4}}.$$

*Proof.* For every  $k \in \mathbb{N}^+$ , let  $A_k = \pi_k(X_0, \dots, X_{k-1})$ . For every  $a \in \mathcal{A}$ ,

$$T_{mn,a}^\pi(\omega) = \sum_{k=1}^{mn} \mathbb{I}_{\{A_k=a\}}(\omega) = \sum_{k=1}^{mn} \mathbb{I}_{\{((k-1) \bmod n)+1=a\}}(\omega) = m.$$

Theorem 4.2 completes the proof for the case where  $t = mn$ , since  $(t - mn) = 0$  and

$$R_{mn}^{\nu, \pi} = \sum_{a=1}^n \Delta_a^\nu \mathbb{E}^{\nu, \pi} (T_{mn,a}^\pi) = m \sum_{a=1}^n \Delta_a^\nu.$$

Consider a time step  $t \in \mathbb{N}^+$  such that  $t > mn$ . In that case,

$$T_{t,a}^\pi(\omega) = \sum_{k=1}^{mn} \mathbb{I}_{\{A_k=a\}}(\omega) + \sum_{k=mn+1}^t \mathbb{I}_{\{A_k=a\}}(\omega) = m + (t - mn) \mathbb{I}_{\{a = \arg \max_{a'} M_{mn,a'}^\pi\}}(\omega).$$

Because ties are possible, for every  $a \in \mathcal{A}$  and  $t > mn$ ,

$$\mathbb{E}^{\nu, \pi} (T_{t,a}^\pi) = m + (t - mn) \mathbb{P}^{\nu, \pi} \left( a = \arg \max_{a'} M_{mn,a'}^\pi \right) \leq m + (t - mn) \mathbb{P}^{\nu, \pi} \left( M_{mn,a}^\pi \geq \sup_{a'} M_{mn,a'}^\pi \right).$$

Let  $a^*$  denote an action such that  $\mu_{a^*}^\nu = \mu_{*}^\nu$ . For every  $a \in \mathcal{A}$  and  $t > mn$ ,

$$\mathbb{P}^{\nu, \pi} \left( M_{mn,a}^\pi \geq \sup_{a'} M_{mn,a'}^\pi \right) = \mathbb{P}^{\nu, \pi} \left( \bigcap_{a'} \{M_{mn,a}^\pi \geq M_{mn,a'}^\pi\} \right) \leq \mathbb{P}^{\nu, \pi} (M_{mn,a}^\pi \geq M_{mn,a^*}^\pi).$$

For every  $a \in \mathcal{A}$  and  $t > mn$ , by adding  $\Delta_a^\nu$  to both sides of the inequality that defines an event,

$$\mathbb{P}^{\nu, \pi} \left( M_{mn,a}^\pi \geq \sup_{a'} M_{mn,a'}^\pi \right) \leq \mathbb{P}^{\nu, \pi} (M_{mn,a}^\pi - M_{mn,a^*}^\pi \geq 0) = \mathbb{P}^{\nu, \pi} (M_{mn,a}^\pi - M_{mn,a^*}^\pi + (\mu_{a^*}^\nu - \mu_a^\nu) \geq \Delta_a^\nu),$$

so that

$$\mathbb{P}^{\nu, \pi} \left( M_{mn,a}^\pi \geq \sup_{a'} M_{mn,a'}^\pi \right) \leq \mathbb{P}^{\nu, \pi} ((M_{mn,a}^\pi - \mu_a^\nu) - (M_{mn,a^*}^\pi - \mu_{a^*}^\nu) \geq \Delta_a^\nu).$$

For every  $a \in \mathcal{A}$ , by the definition of the average reward  $M_{mn,a}^\pi$  that policy  $\pi$  observes for  $a$  by time  $mn$ ,

$$M_{mn,a}^\pi(\omega) - \mu_a^\nu = \left( \frac{1}{m} \sum_{i=0}^{m-1} X_{a+in}(\omega) \right) - \frac{1}{m} \sum_{i=0}^{m-1} \mu_a^\nu = \frac{1}{m} \sum_{i=0}^{m-1} (X_{a+in}(\omega) - \mu_a^\nu).$$

Proposition 5.4 guarantees that  $X_{a+in} - \mu_a^\nu$  is 1-subgaussian for every  $a \in \{1, \dots, n\}$  and  $i \in \{0, \dots, m-1\}$ , since  $((a+in-1) \bmod n) + 1 = a$ . Proposition 5.3 guarantees that  $X_a, X_{a+n}, \dots, X_{a+(m-1)n}$  are independent. Therefore,  $\sum_{i=0}^{m-1} (X_{a+in} - \mu_a^\nu)$  is  $\sqrt{m}$ -subgaussian, which implies that  $M_{mn,a}^\pi - \mu_a^\nu$  is  $1/\sqrt{m}$ -subgaussian. Since this applies for every  $a \in \mathcal{A}$ , we also conclude that  $M_{mn,a^*}^\pi - \mu_{a^*}^\nu$  is  $1/\sqrt{m}$ -subgaussian. For every  $a \in \mathcal{A}$ , note that  $M_{mn,a}^\pi - \mu_a^\nu$  is  $\sigma(X_a, X_{a+n}, \dots, X_{a+(m-1)n})$ -measurable. By Proposition 5.3, if  $a \neq a^*$ , then  $(M_{mn,a}^\pi - \mu_a^\nu)$  and  $-(M_{mn,a^*}^\pi - \mu_{a^*}^\nu)$  are independent, which further implies that  $(M_{mn,a}^\pi - \mu_a^\nu) - (M_{mn,a^*}^\pi - \mu_{a^*}^\nu)$  is  $\sqrt{2/m}$ -subgaussian. If  $a = a^*$ , then  $(M_{mn,a}^\pi - \mu_a^\nu) - (M_{mn,a^*}^\pi - \mu_{a^*}^\nu) = 0$ , and therefore also  $\sqrt{2/m}$ -subgaussian. By Theorem 3.1, since  $\Delta_a^\nu \geq 0$ ,

$$\mathbb{P}^{\nu, \pi} \left( M_{mn,a}^\pi \geq \sup_{a'} M_{mn,a'}^\pi \right) \leq e^{-\frac{(\Delta_a^\nu)^2}{2(\sqrt{2/m})^2}} = e^{-\frac{m(\Delta_a^\nu)^2}{4}}.$$

By returning to a previous inequality, for every  $a \in \mathcal{A}$  and  $t > mn$ ,

$$\mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq m + (t - mn)e^{-\frac{m(\Delta_a^\nu)^2}{4}}.$$

For every  $t > mn$ , Theorem 4.2 once again completes the proof, since

$$R_t^{\nu, \pi} = \sum_{a=1}^n \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq \sum_{a=1}^n \Delta_a^\nu \left( m + (t - mn)e^{-\frac{m(\Delta_a^\nu)^2}{4}} \right) = \left( m \sum_{a=1}^n \Delta_a^\nu \right) + (t - mn) \sum_{a=1}^n \Delta_a^\nu e^{-\frac{m(\Delta_a^\nu)^2}{4}}.$$

□

In order to minimize the regret, the previous result suggests that the exploration factor  $m$  should balance between the first term (non-decreasing with respect to  $m$ ) and the second term (non-increasing with respect to  $m$ ). This is a specific instance of the so-called exploration-exploitation trade-off.

**Proposition 5.5.** Consider a 1-subgaussian stochastic bandit  $\nu = (P_1, P_2)$ . Let  $\Delta = \max(\Delta_1^\nu, \Delta_2^\nu)$ , and suppose that  $\Delta > 0$ . For some  $t \in \mathbb{N}^+$ , let  $m = 1$  if  $t \leq 4/\Delta^2$  and let  $m = \left\lceil \frac{4}{\Delta^2} \log \left( \frac{t\Delta^2}{4} \right) \right\rceil$  if  $t > 4/\Delta^2$ . If  $\pi$  is a policy that implements explore-then-commit with  $m$  exploration steps, then

$$R_t^{\nu, \pi} \leq \Delta + \frac{4}{\sqrt{e}} \sqrt{t}.$$

*Proof.* First, consider some  $t \in \mathbb{N}^+$  such that  $t \leq 4/\Delta^2$ , so that  $m = 1$ . By Theorem 4.2, since  $\Delta \leq 2/\sqrt{t}$ ,

$$R_t^{\nu, \pi} = \sum_{a=1}^2 \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq \Delta \sum_{a=1}^2 \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) = \Delta \mathbb{E}^{\nu, \pi} \left( \sum_{a=1}^2 T_{t,a}^\pi \right) = t\Delta \leq t \frac{2}{\sqrt{t}} = 2\sqrt{t}.$$

Second, consider some  $t \in \mathbb{N}^+$  such that  $t > 4/\Delta^2$ , so that  $m = \left\lceil \frac{4}{\Delta^2} \log \left( \frac{t\Delta^2}{4} \right) \right\rceil$ . Note that  $m \geq 1$  and

$$m\Delta = \Delta \left\lceil \frac{4}{\Delta^2} \log \left( \frac{t\Delta^2}{4} \right) \right\rceil \leq \Delta \left( 1 + \frac{4}{\Delta^2} \log \left( \frac{t\Delta^2}{4} \right) \right) = \Delta + \frac{4}{\Delta} \log \left( \frac{t\Delta^2}{4} \right).$$

Consider the case where  $t < 2m$ . By Theorem 4.2,

$$R_t^{\nu, \pi} = \Delta_1^\nu \mathbb{E}^{\nu, \pi}(T_{t,1}^\pi) + \Delta_2^\nu \mathbb{E}^{\nu, \pi}(T_{t,2}^\pi) \leq m\Delta.$$

Now consider the case where  $t \geq 2m$ . By Theorem 5.1,

$$R_t^{\nu, \pi} \leq m\Delta + (t - 2m)\Delta e^{-\frac{m\Delta^2}{4}} \leq m\Delta + t\Delta e^{-\frac{m\Delta^2}{4}}.$$

Because the function  $f : (0, \infty) \rightarrow (0, \infty)$  given by  $f(x) = t\Delta e^{-\frac{x\Delta^2}{4}}$  is decreasing,

$$t\Delta e^{-\frac{m\Delta^2}{4}} = f(m) = f\left(\left\lceil \frac{4}{\Delta^2} \log\left(\frac{t\Delta^2}{4}\right) \right\rceil\right) \leq f\left(\frac{4}{\Delta^2} \log\left(\frac{t\Delta^2}{4}\right)\right) = t\Delta e^{-\log\left(\frac{t\Delta^2}{4}\right)} = \frac{4}{\Delta}.$$

Therefore, for every  $t \in \mathbb{N}^+$  such that  $t > 4/\Delta^2$ ,

$$R_t^{\nu, \pi} \leq m\Delta + t\Delta e^{-\frac{m\Delta^2}{4}} \leq \Delta + \frac{4}{\Delta} \log\left(\frac{t\Delta^2}{4}\right) + \frac{4}{\Delta}.$$

Consider the function  $g : (0, \infty) \rightarrow \mathbb{R}$  given by  $g(x) = x \log(4t/x^2) + x$ , so that  $g(4/\Delta) = (4/\Delta) \log(t\Delta^2/4) + 4/\Delta$ . Note that  $g(x) = x \log(4t) - 2x \log(x) + x$ ,  $g'(x) = \log(4t) - 2 \log(x) - 1$ , and  $g''(x) = -2/x$ . The second derivative test guarantees that  $g(x) \leq g(2\sqrt{t}/\sqrt{e}) = 4\sqrt{t}/\sqrt{e}$  for every  $x \in (0, \infty)$ . Therefore, for every  $t \in \mathbb{N}^+$ ,

$$R_t^{\nu, \pi} \leq \Delta + \frac{4}{\sqrt{e}} \sqrt{t}.$$

□

The previous result suggests a specific number of exploration steps for a policy that implements explore-then-commit. However, this policy is only suitable for a fixed horizon and a fixed suboptimality gap.

## 6 Restarts

Consider a number of actions  $n \in \mathbb{N}^+$ , a set of actions  $\mathcal{A} = \{1, \dots, n\}$ , a stochastic bandit  $\nu = (P_a \mid a \in \mathcal{A})$ , a policy  $\pi = (\pi_t \mid t \in \mathbb{N}^+)$ , and let  $(\Omega, \mathcal{F}, \mathbb{P}^{\nu, \pi})$  be a canonical triple for the stochastic bandit  $\nu$  under the policy  $\pi$ .

**Definition 6.1.** A policy  $\pi$  restarts to the policy  $\pi'$  after  $t \in \mathbb{N}$  steps if, for all  $k \in \mathbb{N}^+$  and  $(x_0, \dots, x_{t+k-1}) \in \mathbb{R}^{t+k}$ ,

$$\pi_{t+k}(x_0, \dots, x_{t+k-1}) = \pi'_k(0, x_{t+1}, \dots, x_{t+k-1}).$$

**Proposition 6.1.** If a policy  $\pi$  restarts to the policy  $\pi'$  after  $t \in \mathbb{N}$  steps, then

$$\mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1, \dots, X_{t+k} \in B_k) = \mathbb{P}^{\nu, \pi'}(X_1 \in B_1, \dots, X_k \in B_k)$$

for every  $k \in \mathbb{N}^+$  and  $B_1, \dots, B_k \in \mathcal{B}(\mathbb{R})$ .

*Proof.* Consider the case where  $k = 1$ . For every  $B_1 \in \mathcal{B}(\mathbb{R})$ ,

$$\mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1) = \mathbb{E}^{\nu, \pi}(\mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{X_{t+1} \in B_1\}} \mid X_0, \dots, X_t)) = \mathbb{E}^{\nu, \pi}(P_{A_{t+1}}(B_1)),$$

where  $A_{t+1} = \pi_{t+1}(X_0, \dots, X_t) = \pi'_1(0)$ . Because  $A_{t+1}$  is a constant function,

$$\mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1) = P_{\pi'_1(0)}(B_1) = \mathbb{E}^{\nu, \pi'}(P_{\pi'_1(0)}(B_1)) = \mathbb{E}^{\nu, \pi'}(P_{\pi'_1(X_0)}(B_1)) = \mathbb{P}^{\nu, \pi'}(X_1 \in B_1).$$

In order to employ induction, suppose that there is a  $k \in \mathbb{N}^+$  such that, for every  $B_1, \dots, B_k \in \mathcal{B}(\mathbb{R})$ ,

$$\mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1, \dots, X_{t+k} \in B_k) = \mathbb{P}^{\nu, \pi'}(X_1 \in B_1, \dots, X_k \in B_k).$$

In that case, there is a probability measure  $\mathcal{L} : \mathcal{B}(\mathbb{R}^k) \rightarrow [0, 1]$  on the measurable space  $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$  such that

$$\mathcal{L}(B_1 \times \dots \times B_k) = \mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1, \dots, X_{t+k} \in B_k) = \mathbb{P}^{\nu, \pi'}(X_1 \in B_1, \dots, X_k \in B_k)$$

for every  $B_1, \dots, B_k \in \mathcal{B}(\mathbb{R})$ , so that  $\mathcal{L}$  is the joint law of  $(X_{t+1}, \dots, X_{t+k})$  and the joint law of  $(X_1, \dots, X_k)$ .

For every  $B_1, \dots, B_{k+1} \in \mathcal{B}(\mathbb{R})$ ,

$$\begin{aligned} \mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1, \dots, X_{t+k+1} \in B_{k+1}) &= \mathbb{E}^{\nu, \pi}(\mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{X_{t+1} \in B_1, \dots, X_{t+k} \in B_k\}} \mathbb{I}_{\{X_{t+k+1} \in B_{k+1}\}} \mid X_0, \dots, X_{t+k})), \\ \mathbb{P}^{\nu, \pi'}(X_1 \in B_1, \dots, X_{k+1} \in B_{k+1}) &= \mathbb{E}^{\nu, \pi'}(\mathbb{E}^{\nu, \pi'}(\mathbb{I}_{\{X_1 \in B_1, \dots, X_k \in B_k\}} \mathbb{I}_{\{X_{k+1} \in B_{k+1}\}} \mid X_0, \dots, X_k)). \end{aligned}$$

By taking out what is known,

$$\begin{aligned} \mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1, \dots, X_{t+k+1} \in B_{k+1}) &= \mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{X_{t+1} \in B_1, \dots, X_{t+k} \in B_k\}} P_{A_{t+k+1}}(B_{k+1})), \\ \mathbb{P}^{\nu, \pi'}(X_1 \in B_1, \dots, X_{k+1} \in B_{k+1}) &= \mathbb{E}^{\nu, \pi'}(\mathbb{I}_{\{X_1 \in B_1, \dots, X_k \in B_k\}} P_{A'_{k+1}}(B_{k+1})), \end{aligned}$$

where  $A_{t+k+1} = \pi_{t+k+1}(X_0, \dots, X_{t+k})$  and  $A'_{k+1} = \pi'_{k+1}(0, X_1, \dots, X_k)$ . Since  $A_{t+k+1} = \pi'_{k+1}(0, X_{t+1}, \dots, X_{t+k})$ ,

$$\begin{aligned} \mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1, \dots, X_{t+k+1} \in B_{k+1}) &= \mathbb{E}^{\nu, \pi}(f(X_{t+1}, \dots, X_{t+k})), \\ \mathbb{P}^{\nu, \pi'}(X_1 \in B_1, \dots, X_{k+1} \in B_{k+1}) &= \mathbb{E}^{\nu, \pi'}(f(X_1, \dots, X_k)), \end{aligned}$$

where the function  $f : \mathbb{R}^k \rightarrow [0, 1]$  is given by

$$f(x) = \left( \prod_{i=1}^k \mathbb{I}_{B_i}(x_i) \right) P_{\pi'_{k+1}(0, x_1, \dots, x_k)}(B_{k+1}).$$

Since  $\mathcal{L}$  is the joint law of  $(X_{t+1}, \dots, X_{t+k})$  and the joint law of  $(X_1, \dots, X_k)$ ,

$$\mathbb{P}^{\nu, \pi}(X_{t+1} \in B_1, \dots, X_{t+k+1} \in B_{k+1}) = \int_{\mathbb{R}^k} f(x) \mathcal{L}(dx) = \mathbb{P}^{\nu, \pi'}(X_1 \in B_1, \dots, X_{k+1} \in B_{k+1}).$$

□

**Proposition 6.2.** If a policy  $\pi$  restarts to the policy  $\pi'$  after  $t \in \mathbb{N}^+$  steps, for every  $h \in \mathbb{N}^+$ ,

$$R_{t+h}^{\nu, \pi} = R_t^{\nu, \pi} + R_h^{\nu, \pi'}.$$

*Proof.* For every  $h \in \mathbb{N}^+$ , by definition of the regret  $R_{t+h}^{\nu, \pi}$ ,

$$R_{t+h}^{\nu, \pi} = (t+h)\mu_*^\nu - \sum_{k=1}^{t+h} \mathbb{E}^{\nu, \pi}(X_k) = \left( t\mu_*^\nu - \sum_{k=1}^t \mathbb{E}^{\nu, \pi}(X_k) \right) + \left( h\mu_*^\nu - \sum_{k=t+1}^{t+h} \mathbb{E}^{\nu, \pi}(X_k) \right).$$

By definition of the regret  $R_t^{\nu, \pi}$  and changing the indices of the second summation,

$$R_{t+h}^{\nu, \pi} = R_t^{\nu, \pi} + \left( h\mu_*^\nu - \sum_{k=1}^h \mathbb{E}^{\nu, \pi}(X_{t+k}) \right).$$

By Proposition 6.1, we know that  $\mathbb{P}^{\nu, \pi}(X_{t+k} \in B) = \mathbb{P}^{\nu, \pi'}(X_k \in B)$  for every  $k \in \mathbb{N}^+$  and  $B \in \mathcal{B}(\mathbb{R})$ . Therefore,  $\mathbb{E}^{\nu, \pi}(X_{t+k}) = \mathbb{E}^{\nu, \pi'}(X_k)$  for every  $k \in \mathbb{N}^+$  and

$$R_{t+h}^{\nu, \pi} = R_t^{\nu, \pi} + \left( h\mu_*^\nu - \sum_{k=1}^h \mathbb{E}^{\nu, \pi'}(X_k) \right) = R_t^{\nu, \pi} + R_h^{\nu, \pi'}.$$

□

**Definition 6.2.** Consider a sequence of policies  $(\pi^{(k)} \mid k \in \mathbb{N}^+)$  and a sequence of positive natural numbers  $(h_k \in \mathbb{N}^+ \mid k \in \mathbb{N}^+)$ . For every  $k \in \mathbb{N}^+$ , suppose that the policy  $\pi^{(k)}$  restarts to the policy  $\pi^{(k+1)}$  after  $h_k$  steps. If  $\pi = \pi^{(1)}$ , we say that policy  $\pi$  restarts to the sequence of policies  $(\pi^{(k)} \mid k \in \mathbb{N}^+)$  given the sequence of relative steps  $(h_k \mid k \in \mathbb{N}^+)$ .

**Proposition 6.3.** If the policy  $\pi$  restarts to the sequence of policies  $(\pi^{(k)} \mid k \in \mathbb{N}^+)$  given the sequence of relative steps  $(h_k \in \mathbb{N}^+ \mid k \in \mathbb{N}^+)$ , for every  $l \in \mathbb{N}^+$ ,

$$R_{\sum_{k=1}^l h_k}^{\nu, \pi} = \sum_{k=1}^l R_{h_k}^{\nu, \pi^{(k)}}.$$

*Proof.* If  $l = 1$ , then  $R_{h_1}^{\nu, \pi} = R_{h_1}^{\nu, \pi^{(1)}}$ . By Proposition 6.2, if  $l > 1$ , then

$$R_{\sum_{k=1}^l h_k}^{\nu, \pi} = R_{\sum_{k=1}^l h_k}^{\nu, \pi^{(1)}} = R_{h_1}^{\nu, \pi^{(1)}} + R_{\sum_{k=2}^l h_k}^{\nu, \pi^{(2)}} = \dots = \sum_{k=1}^l R_{h_k}^{\nu, \pi^{(k)}}.$$

□

**Proposition 6.4.** If the policy  $\pi$  restarts to the sequence of policies  $(\pi^{(k)} \mid k \in \mathbb{N}^+)$  given the sequence of relative steps  $(h_k \in \mathbb{N}^+ \mid k \in \mathbb{N}^+)$  and there is a function  $f : \mathbb{N}^+ \rightarrow [0, \infty)$  such that  $R_{h_k}^{\nu, \pi^{(k)}} \leq f(h_k)$  for every  $k \in \mathbb{N}^+$ , then

$$R_t^{\nu, \pi} \leq \sum_{k=1}^{p_t} f(h_k)$$

for every  $t \in \mathbb{N}^+$ , where  $p_t = \min\{l \in \mathbb{N}^+ \mid \sum_{k=1}^l h_k \geq t\}$  is the number of restarts by time step  $t$ .

*Proof.* For every  $t \in \mathbb{N}^+$ , let  $p_t = \min\{l \in \mathbb{N}^+ \mid \sum_{k=1}^l h_k \geq t\}$ , so that  $\sum_{k=1}^{p_t} h_k \geq t$ . By Proposition 6.3,

$$R_t^{\nu, \pi} \leq R_{\sum_{k=1}^{p_t} h_k}^{\nu, \pi} = \sum_{k=1}^{p_t} R_{h_k}^{\nu, \pi^{(k)}} \leq \sum_{k=1}^{p_t} f(h_k).$$

□

The previous result can be used to provide a regret upper bound based on the regret upper bounds of policies suitable for fixed horizons. This is exemplified by the so-called doubling trick, which is presented below.



**Proposition 6.5.** If the policy  $\pi$  restarts to the sequence of policies  $(\pi^{(k)} \mid k \in \mathbb{N}^+)$  given the sequence of relative steps  $(2^{k-1} \mid k \in \mathbb{N}^+)$  and  $R_{2^{k-1}}^{\nu, \pi^{(k)}} \leq \sqrt{2^{k-1}}$  for every  $k \in \mathbb{N}^+$ , then, for every  $t \in \mathbb{N}^+$ ,

$$R_t^{\nu, \pi} \leq 2(1 + \sqrt{2})\sqrt{t}.$$

*Proof.* For every  $t \in \mathbb{N}^+$ , let  $p_t = \min\{l \in \mathbb{N}^+ \mid \sum_{k=1}^l 2^{k-1} \geq t\}$ , so that  $p_t = \lceil \log_2(t+1) \rceil$ . By Proposition 6.4,

$$R_t^{\nu, \pi} \leq \sum_{k=1}^{p_t} \sqrt{2^{k-1}} = \sum_{k=1}^{p_t} (\sqrt{2})^{k-1} = \frac{(\sqrt{2})^{p_t} - 1}{\sqrt{2} - 1} \leq \frac{(\sqrt{2})^{p_t}}{\sqrt{2} - 1}.$$

Since  $p_t \leq \log_2(t+1) + 1 = \log_2(t+1) + \log_2(2) = \log_2 2(t+1)$  and  $1 + 1/t \leq 2$ ,

$$R_t^{\nu, \pi} \leq \frac{(\sqrt{2})^{\log_2 2(t+1)}}{\sqrt{2} - 1} = \frac{\sqrt{2(t+1)}}{\sqrt{2} - 1} = \frac{1}{\sqrt{2} - 1} \sqrt{2t \left(1 + \frac{1}{t}\right)} \leq \frac{\sqrt{4t}}{\sqrt{2} - 1} = \frac{2\sqrt{t}}{\sqrt{2} - 1}.$$

□

Note that doubling the horizon after each restart is not generally appropriate.

## 7 Action times

Consider a number of actions  $n \in \mathbb{N}^+$ , a set of actions  $\mathcal{A} = \{1, \dots, n\}$ , a stochastic bandit  $\nu = (P_a \mid a \in \mathcal{A})$ , a policy  $\pi = (\pi_t \mid t \in \mathbb{N}^+)$ , and a canonical triple  $(\Omega, \mathcal{F}, \mathbb{P}^{\nu, \pi})$  for the stochastic bandit  $\nu$  under the policy  $\pi$ . Furthermore, let  $(\mathcal{F}_t)_t$  denote the natural filtration of the reward process  $(X_t \mid t \in \mathbb{N})$ , so that  $\mathcal{F}_t = \sigma(X_0, \dots, X_t)$  for every  $t \in \mathbb{N}$ .

**Definition 7.1.** The time  $C_{m,a}^\pi : \Omega \rightarrow \mathbb{N}^+ \cup \{\infty\}$  until policy  $\pi$  selects  $a \in \mathcal{A}$  exactly  $m \in \mathbb{N}^+$  times is given by

$$C_{m,a}^\pi(\omega) = \inf \left( \{t \in \mathbb{N}^+ \mid T_{t,a}^\pi(\omega) \geq m\} \right).$$

If  $t \in \mathbb{N}^+$  and  $C_{m,a}^\pi(\omega) = t$ , then  $\pi_t(X_0(\omega), \dots, X_{t-1}(\omega)) = a$  and  $C_{m+1,a}^\pi(\omega) > t$ .

**Proposition 7.1.** The time  $C_{m,a}^\pi : \Omega \rightarrow \mathbb{N}^+ \cup \{\infty\}$  until  $\pi$  selects  $a \in \mathcal{A}$  exactly  $m \in \mathbb{N}^+$  times is a stopping time.

*Proof.* Recall that  $C_{m,a}^\pi$  is a stopping time if  $\{C_{m,a}^\pi \leq t\} \in \mathcal{F}_t$  for every  $t \in \mathbb{N} \cup \{\infty\}$ . If  $t = 0$ , then  $\{C_{m,a}^\pi \leq 0\} = \emptyset$ . If  $t \in \mathbb{N}^+$ , then  $\{C_{m,a}^\pi \leq t\} = \{T_{t,a}^\pi \geq m\}$  and  $\{T_{t,a}^\pi \geq m\} \in \mathcal{F}_{t-1}$ . If  $t = \infty$ , then  $\{C_{m,a}^\pi \leq \infty\} = \Omega$ .  $\square$

**Definition 7.2.** For every  $a \in \mathcal{A}$  and  $m \in \mathbb{N}^+$ , the function  $X_{C_{m,a}^\pi} : \Omega \rightarrow \mathbb{R}$  is given by

$$X_{C_{m,a}^\pi}(\omega) = \begin{cases} X_{C_{m,a}^\pi(\omega)}(\omega), & \text{if } C_{m,a}^\pi(\omega) < \infty, \\ 0, & \text{if } C_{m,a}^\pi(\omega) = \infty. \end{cases}$$

Recall that  $X_{C_{m,a}^\pi}$  is  $\mathcal{F}$ -measurable because  $(X_t \mid t \in \mathbb{N})$  is adapted to  $(\mathcal{F}_t)_t$  and  $C_{m,a}^\pi$  is a stopping time.

**Definition 7.3.** For every  $a \in \mathcal{A}$ , the constant policy  $\pi^{(a)} = (\pi_t^{(a)} \mid t \in \mathbb{N}^+)$  is given by  $\pi_t^{(a)} = a$  for every  $t \in \mathbb{N}^+$ .

**Proposition 7.2.** For every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $B_1, \dots, B_m \in \mathcal{B}(\mathbb{R})$ ,

$$\mathbb{P}^{\nu, \pi^{(a)}}(X_1 \in B_1, \dots, X_m \in B_m) = \prod_{k=1}^m P_a(B_k).$$

*Proof.* For every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $B_1, \dots, B_m \in \mathcal{B}(\mathbb{R})$ , if the empty product denotes one,

$$\mathbb{P}^{\nu, \pi^{(a)}}(X_1 \in B_1, \dots, X_m \in B_m) = \mathbb{E}^{\nu, \pi^{(a)}} \left( \mathbb{E}^{\nu, \pi^{(a)}} \left( \left( \prod_{k=1}^{m-1} \mathbb{I}_{\{X_k \in B_k\}} \right) \mathbb{I}_{\{X_m \in B_m\}} \mid X_0, \dots, X_{m-1} \right) \right).$$

By taking out what is known and using the fact that  $\pi_m^{(a)}(X_0, \dots, X_{m-1}) = a$ ,

$$\mathbb{P}^{\nu, \pi^{(a)}}(X_1 \in B_1, \dots, X_m \in B_m) = P_a(B_m) \mathbb{E}^{\nu, \pi^{(a)}} \left( \prod_{k=1}^{m-1} \mathbb{I}_{\{X_k \in B_k\}} \right).$$

Therefore,  $\mathbb{P}^{\nu, \pi^{(a)}}(X_1 \in B_1) = P_a(B_1)$ . Suppose that the proposition is true for some  $m-1 \in \mathbb{N}^+$ . In that case,

$$\mathbb{P}^{\nu, \pi^{(a)}}(X_1 \in B_1, \dots, X_m \in B_m) = P_a(B_m) \mathbb{P}^{\nu, \pi^{(a)}}(X_1 \in B_1, \dots, X_{m-1} \in B_{m-1}) = \prod_{k=1}^m P_a(B_k).$$

$\square$

**Proposition 7.3.** For every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $t \in \mathbb{N}^+$ , if  $h : \mathbb{R} \rightarrow \mathbb{R}$  is  $\mathcal{B}(\mathbb{R})$ -measurable, then the function  $\mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^{m-1} h(X_{C_{k,a}^\pi})$  is  $\mathcal{F}_{t-1}$ -measurable.

*Proof.* For every  $a \in \mathcal{A}$ ,  $k \in \mathbb{N}^+$ , and  $t_k \in \mathbb{N}^+$ , note that  $\{C_{k,a}^\pi = t_k\} = \{C_{k,a}^\pi \leq t_k\} \cap \{C_{k,a}^\pi \leq t_k - 1\}^c$ , so that  $\{C_{k,a}^\pi = t_k\} \in \mathcal{F}_{t_k-1}$ . For every  $\omega \in \Omega$ ,  $m \in \mathbb{N}^+$ , and  $t \in \mathbb{N}^+$ , if  $C_{m,a}^\pi(\omega) = t$ , then  $C_{1,a}^\pi(\omega) < \dots < C_{m,a}^\pi(\omega) = t$ , so

$$\mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^{m-1} h(X_{C_{k,a}^\pi}) = \mathbb{I}_{\{C_{m,a}^\pi = t\}} \left( \prod_{k=1}^{m-1} \sum_{t_k < t} \mathbb{I}_{\{C_{k,a}^\pi = t_k\}} h(X_{t_k}) \right).$$

If  $k \in \mathbb{N}^+$  and  $t_k \leq t$ , then  $\mathbb{I}_{\{C_{k,a}^\pi = t_k\}}$  is  $\mathcal{F}_{t-1}$ -measurable. If  $t_k < t$ , then  $h(X_{t_k})$  is also  $\mathcal{F}_{t-1}$ -measurable.  $\square$

**Proposition 7.4.** For every  $a \in \mathcal{A}$  and  $m \in \mathbb{N}^+$ , if a function  $h : \mathbb{R} \rightarrow [0, \infty]$  is  $\nu$ -integrable, then

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) \leq \mathbb{E}^{\nu, \pi^{(a)}} \left( \prod_{k=1}^m h(X_k) \right)$$

whenever  $\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) < \infty$  for every  $t \in \mathbb{N}^+$ .

*Proof.* For every  $a \in \mathcal{A}$  and  $t \in \mathbb{N}^+$ , if  $h$  is  $\nu$ -integrable, then  $\mathbb{E}^{\nu, \pi^{(a)}}(h(X_t)) < \infty$ . Therefore, for every  $m \in \mathbb{N}^+$ ,

$$\mathbb{E}^{\nu, \pi^{(a)}} \left( \prod_{k=1}^m h(X_k) \right) = \prod_{k=1}^m \mathbb{E}^{\nu, \pi^{(a)}}(h(X_k)) = \prod_{k=1}^m \int_{\mathbb{R}} h(x) P_a(dx) = \left( \int_{\mathbb{R}} h(x) P_a(dx) \right)^m,$$

which uses the fact that  $X_1, \dots, X_m$  are independent and identically distributed with respect to  $\mathbb{P}^{\nu, \pi^{(a)}}$ .

For every  $a \in \mathcal{A}$  and  $m \in \mathbb{N}^+$ , if the empty product denotes one,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) = \sum_{t \in \mathbb{N}^+} \mathbb{E}^{\nu, \pi} \left( \left( \mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^{m-1} h(X_{C_{k,a}^\pi}) \right) h(X_t) \right).$$

Since each expectation on the right side above is finite by assumption, by taking out what is known,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) = \sum_{t \in \mathbb{N}^+} \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^{m-1} h(X_{C_{k,a}^\pi}) \mathbb{E}^{\nu, \pi}(h(X_t) \mid X_0, \dots, X_{t-1}) \right).$$

By Proposition 4.3, if  $A_t = \pi_t(X_0, \dots, X_{t-1})$ , then almost surely

$$\mathbb{E}^{\nu, \pi}(h(X_t) \mid X_0, \dots, X_{t-1}) = \sum_{a'} \mathbb{I}_{\{A_t = a'\}} \int_{\mathbb{R}} h(x) P_{a'}(dx).$$

For every  $\omega \in \Omega$ , recall that  $C_{m,a}^\pi(\omega) = t$  implies  $A_t(\omega) = a$ . Therefore, almost surely,

$$\mathbb{I}_{\{C_{m,a}^\pi = t\}} \mathbb{E}^{\nu, \pi}(h(X_t) \mid X_0, \dots, X_{t-1}) = \mathbb{I}_{\{C_{m,a}^\pi = t\}} \int_{\mathbb{R}} h(x) P_a(dx).$$

By returning to a previous equation,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) = \left( \int_{\mathbb{R}} h(x) P_a(dx) \right) \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} \prod_{k=1}^{m-1} h(X_{C_{k,a}^\pi}) \right).$$

The proposition is true for  $m = 1$ , since

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{1,a}^\pi < \infty\}} h(X_{C_{1,a}^\pi}) \right) = \left( \int_{\mathbb{R}} h(x) P_a(dx) \right) \mathbb{P}^{\nu, \pi}(C_{1,a}^\pi < \infty) \leq \int_{\mathbb{R}} h(x) P_a(dx).$$

If the proposition is true for some  $m - 1 \in \mathbb{N}^+$ , because  $C_{m,a}^\pi(\omega) < \infty$  implies  $C_{m-1,a}^\pi(\omega) < \infty$  for every  $\omega \in \Omega$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) \leq \left( \int_{\mathbb{R}} h(x) P_a(dx) \right) \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m-1,a}^\pi < \infty\}} \prod_{k=1}^{m-1} h(X_{C_{k,a}^\pi}) \right) \leq \left( \int_{\mathbb{R}} h(x) P_a(dx) \right)^m.$$

□

**Proposition 7.5.** If  $\nu$  is a 1-subgaussian stochastic bandit and  $\lambda \in \mathbb{R}$ , then the function  $h : \mathbb{R} \rightarrow [0, \infty]$  given by  $h(x) = e^{\lambda x}$  is  $\nu$ -integrable. Furthermore, for every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $t \in \mathbb{N}^+$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) < \infty.$$

*Proof.* If  $\nu$  is a 1-subgaussian stochastic bandit, recall that the random variable  $Z_a$  on the probability triple  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_a)$  given by  $Z_a(x) = x - \mu_a^\nu$  is 1-subgaussian for every  $a \in \mathcal{A}$ . For every  $\lambda \in \mathbb{R}$ ,

$$\int_{\mathbb{R}} e^{\lambda x} P_a(dx) = \int_{\mathbb{R}} e^{\lambda(Z_a(x) + \mu_a^\nu)} P_a(dx) = e^{\lambda \mu_a^\nu} \int_{\mathbb{R}} e^{\lambda Z_a(x)} P_a(dx) \leq e^{\lambda \mu_a^\nu} e^{\frac{\lambda^2}{2}}.$$

By Proposition 4.1, there is a constant  $c \in [0, \infty)$  such that  $\mu_a^\nu \in [-c, c]$  for every  $a \in \mathcal{A}$ . Therefore, the function  $h : \mathbb{R} \rightarrow [0, \infty]$  given by  $h(x) = e^{\lambda x}$  is  $\nu$ -integrable.

Let  $a \in \mathcal{A}$  and  $t \in \mathbb{N}^+$ . We will use induction to show that, for every  $m \in \mathbb{N}^+$  and  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi \leq t\}} e^{\lambda \sum_{k=1}^m X_{C_{k,a}^\pi}} \right) < \infty.$$

Consider the case where  $m = 1$ . For every  $\lambda \in \mathbb{R}$ , since  $\mathbb{E}^{\nu, \pi}(e^{\lambda X_{t'}}) < \infty$  for every  $t' \in \mathbb{N}^+$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{1,a}^\pi \leq t\}} e^{\lambda X_{C_{1,a}^\pi}} \right) = \sum_{t' \leq t} \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{1,a}^\pi = t'\}} e^{\lambda X_{t'}} \right) \leq \sum_{t' \leq t} \mathbb{E}^{\nu, \pi} (e^{\lambda X_{t'}}) < \infty.$$

Suppose that there is an  $m-1 \in \mathbb{N}^+$  such that, for every  $\lambda' \in \mathbb{R}$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m-1,a}^\pi \leq t\}} e^{\lambda' \sum_{k=1}^{m-1} X_{C_{k,a}^\pi}} \right) < \infty.$$

For every  $\lambda \in \mathbb{R}$ , since  $\mathbb{I}_{\{C_{m,a}^\pi \leq t\}} = \mathbb{I}_{\{C_{m-1,a}^\pi \leq t\}} \mathbb{I}_{\{C_{m,a}^\pi \leq t\}}$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi \leq t\}} e^{\lambda \sum_{k=1}^m X_{C_{k,a}^\pi}} \right) = \mathbb{E}^{\nu, \pi} \left( \left( \mathbb{I}_{\{C_{m-1,a}^\pi \leq t\}} e^{\lambda \sum_{k=1}^{m-1} X_{C_{k,a}^\pi}} \right) \left( \mathbb{I}_{\{C_{m,a}^\pi \leq t\}} e^{\lambda X_{C_{m,a}^\pi}} \right) \right).$$

If  $\lambda' = 2\lambda$ , by the inductive hypothesis,

$$\mathbb{E}^{\nu, \pi} \left( \left( \mathbb{I}_{\{C_{m-1,a}^\pi \leq t\}} e^{\lambda \sum_{k=1}^{m-1} X_{C_{k,a}^\pi}} \right)^2 \right) = \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m-1,a}^\pi \leq t\}} e^{\lambda' \sum_{k=1}^{m-1} X_{C_{k,a}^\pi}} \right) < \infty.$$

Since  $\mathbb{E}^{\nu, \pi}(e^{\lambda' X_{t'}}) < \infty$  for every  $t' \in \mathbb{N}^+$ ,

$$\mathbb{E}^{\nu, \pi} \left( \left( \mathbb{I}_{\{C_{m,a}^\pi \leq t\}} e^{\lambda X_{C_{m,a}^\pi}} \right)^2 \right) = \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi \leq t\}} e^{\lambda' X_{C_{m,a}^\pi}} \right) = \sum_{t' \leq t} \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi = t'\}} e^{\lambda' X_{t'}} \right) \leq \sum_{t' \leq t} \mathbb{E}^{\nu, \pi} (e^{\lambda' X_{t'}}) < \infty.$$

By the Schwarz inequality, for every  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi \leq t\}} e^{\lambda \sum_{k=1}^m X_{C_{k,a}^\pi}} \right) < \infty.$$

Therefore, for every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ ,  $t \in \mathbb{N}^+$ , and  $\lambda \in \mathbb{R}$ , if  $h : \mathbb{R} \rightarrow [0, \infty]$  is given by  $h(x) = e^{\lambda x}$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) \leq \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi \leq t\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) = \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi \leq t\}} e^{\lambda \sum_{k=1}^m X_{C_{k,a}^\pi}} \right) < \infty.$$

□

**Proposition 7.6.** If  $\nu$  is a 1-subgaussian stochastic bandit, then, for every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \right) \leq e^{\frac{\lambda^2}{2m}}.$$

*Proof.* For some  $m \in \mathbb{N}^+$  and  $\lambda \in \mathbb{R}$ , consider the function  $h : \mathbb{R} \rightarrow [0, \infty]$  given by  $h(x) = e^{\frac{\lambda}{m} x}$ , which is  $\nu$ -integrable by Proposition 7.5. Recall that, for every  $a \in \mathcal{A}$  and  $t \in \mathbb{N}^+$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) < \infty.$$

For every  $a \in \mathcal{A}$ , consider the function  $h_a : \mathbb{R} \rightarrow [0, \infty]$  given by  $h_a(x) = e^{\frac{\lambda}{m}(x - \mu_a^\nu)} = h(x)e^{-\frac{\lambda}{m}\mu_a^\nu}$ . Since  $h$  is  $\nu$ -integrable,  $h_a$  is also  $\nu$ -integrable. Furthermore, for every  $t \in \mathbb{N}^+$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^m h_a(X_{C_{k,a}^\pi}) \right) = \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi = t\}} \prod_{k=1}^m h(X_{C_{k,a}^\pi}) \right) e^{-\lambda \mu_a^\nu} < \infty.$$

By Proposition 7.4,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} \prod_{k=1}^m h_a(X_{C_{k,a}^\pi}) \right) \leq \mathbb{E}^{\nu, \pi^{(a)}} \left( \prod_{k=1}^m h_a(X_k) \right).$$

By rewriting the previous inequality, for every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \right) \leq \mathbb{E}^{\nu, \pi^{(a)}} \left( e^{\frac{\lambda}{m} \sum_{k=1}^m (X_k - \mu_a^\nu)} \right).$$

Since  $X_1 - \mu_a^\nu, \dots, X_m - \mu_a^\nu$  are independent 1-subgaussian random variables with respect to  $\mathbb{P}^{\nu, \pi^{(a)}}$ , the random variable  $\sum_{k=1}^m (X_k - \mu_a^\nu)$  is  $\sqrt{m}$ -subgaussian, which implies that  $(1/m) \sum_{k=1}^m (X_k - \mu_a^\nu)$  is  $1/\sqrt{m}$ -subgaussian. Therefore, by the definition of a  $1/\sqrt{m}$ -subgaussian random variable,

$$\mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \right) \leq \mathbb{E}^{\nu, \pi^{(a)}} \left( e^{\lambda \frac{1}{m} \sum_{k=1}^m (X_k - \mu_a^\nu)} \right) \leq e^{\frac{\lambda^2}{2m}}.$$

□

**Proposition 7.7.** If  $\nu$  is a 1-subgaussian stochastic bandit, then, for every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $\epsilon \geq 0$ ,

$$\begin{aligned} \mathbb{P}^{\nu, \pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \leq -\epsilon \right) &\leq e^{-\frac{m\epsilon^2}{2}}, \\ \mathbb{P}^{\nu, \pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon \right) &\leq e^{-\frac{m\epsilon^2}{2}}. \end{aligned}$$

*Proof.* For every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ ,  $\epsilon \in \mathbb{R}$ , and  $\lambda \geq 0$ ,

$$\begin{aligned} \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{-\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} &\geq \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{-\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \mathbb{I}_{\{-\frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon\}}, \\ \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} &\geq \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \mathbb{I}_{\{\frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon\}}. \end{aligned}$$

Since the function  $g : \mathbb{R} \rightarrow [0, \infty]$  given by  $g(x) = e^{\lambda x}$  is non-decreasing for  $\lambda \geq 0$ ,

$$\begin{aligned} \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{-\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} &\geq \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\lambda \epsilon} \mathbb{I}_{\{-\frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon\}}, \\ \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} &\geq \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\lambda \epsilon} \mathbb{I}_{\{\frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon\}}. \end{aligned}$$

By taking expectations of both sides of the inequalities above,

$$\begin{aligned} \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{-\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \right) &\geq e^{\lambda \epsilon} \mathbb{P}^{\nu, \pi} \left( C_{m,a}^\pi < \infty, -\frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon \right), \\ \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \right) &\geq e^{\lambda \epsilon} \mathbb{P}^{\nu, \pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon \right). \end{aligned}$$

By Proposition 7.6, for every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $\lambda \geq 0$ ,

$$\begin{aligned} \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{-\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \right) &\leq e^{\frac{(-\lambda)^2}{2m}}, \\ \mathbb{E}^{\nu, \pi} \left( \mathbb{I}_{\{C_{m,a}^\pi < \infty\}} e^{\frac{\lambda}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu)} \right) &\leq e^{\frac{\lambda^2}{2m}}. \end{aligned}$$

By rewriting the previous inequalities,

$$\begin{aligned}\mathbb{P}^{\nu,\pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \leq -\epsilon \right) &\leq e^{\frac{\lambda^2}{2m} - \lambda\epsilon}, \\ \mathbb{P}^{\nu,\pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon \right) &\leq e^{\frac{\lambda^2}{2m} - \lambda\epsilon}.\end{aligned}$$

For every  $\epsilon \geq 0$ , let  $\lambda = \epsilon m$ , so that  $\lambda \geq 0$ . In that case,

$$\begin{aligned}\mathbb{P}^{\nu,\pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \leq -\epsilon \right) &\leq e^{-\frac{m\epsilon^2}{2}}, \\ \mathbb{P}^{\nu,\pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \epsilon \right) &\leq e^{-\frac{m\epsilon^2}{2}}.\end{aligned}$$

□

**Proposition 7.8.** If  $\nu$  is a 1-subgaussian stochastic bandit, then, for every  $a \in \mathcal{A}$ ,  $m \in \mathbb{N}^+$ , and  $\delta \in (0, 1]$ ,

$$\begin{aligned}\mathbb{P}^{\nu,\pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \leq -\sqrt{\frac{2 \log(1/\delta)}{m}} \right) &\leq \delta, \\ \mathbb{P}^{\nu,\pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \sqrt{\frac{2 \log(1/\delta)}{m}} \right) &\leq \delta.\end{aligned}$$

*Proof.* Let  $\delta \in (0, 1]$ . If  $\epsilon = \sqrt{2 \log(1/\delta)/m}$ , then  $\epsilon \geq 0$  and  $\delta = e^{-\frac{m\epsilon^2}{2}}$ , which implies the two inequalities. □

## 8 Upper confidence bounds

Consider a number of actions  $n \in \mathbb{N}^+$ , a set of actions  $\mathcal{A} = \{1, \dots, n\}$ , a stochastic bandit  $\nu = (P_a \mid a \in \mathcal{A})$ , a policy  $\pi = (\pi_t \mid t \in \mathbb{N}^+)$ , and a canonical triple  $(\Omega, \mathcal{F}, \mathbb{P}^{\nu, \pi})$  for the stochastic bandit  $\nu$  under the policy  $\pi$ .

**Definition 8.1.** The upper confidence bound  $U_{t,a}^{\pi, \delta} : \Omega \rightarrow \mathbb{R}$  that policy  $\pi$  induces for action  $a \in \mathcal{A}$  by time step  $t \in \mathbb{N}^+$  with error  $\delta \in (0, 1)$  is given by

$$U_{t,a}^{\pi, \delta}(\omega) = M_{t,a}^{\pi}(\omega) + \sqrt{\frac{2 \log(1/\delta)}{T_{t,a}^{\pi}(\omega)}}$$

whenever  $T_{t,a}^{\pi}(\omega) > 0$ . Intuitively, the role of  $U_{t,a}^{\pi, \delta}$  is to overestimate  $\mu_a^{\nu}$  with high probability when  $\delta$  is small.

**Proposition 8.1.** The upper confidence bound  $U_{t,a}^{\pi, \delta} : \Omega \rightarrow \mathbb{R}$  that policy  $\pi$  induces for action  $a \in \mathcal{A}$  by time step  $t \in \mathbb{N}^+$  with error  $\delta \in (0, 1)$  is given by

$$U_{t,a}^{\pi, \delta}(\omega) = \frac{1}{m} \sum_{k=1}^m X_{C_{k,a}^{\pi}}(\omega) + \sqrt{\frac{2 \log(1/\delta)}{m}}$$

whenever  $T_{t,a}^{\pi}(\omega) = m$  for some  $m \in \mathbb{N}^+$ .

*Proof.* Let  $\omega \in \Omega$ ,  $a \in \mathcal{A}$ ,  $t \in \mathbb{N}^+$ , and  $m \in \mathbb{N}^+$ . If  $T_{t,a}^{\pi}(\omega) = m$ , then  $C_{k,a}^{\pi}(\omega) \leq t$  for every  $k \leq m$ , so that

$$\sum_{k=1}^m X_{C_{k,a}^{\pi}}(\omega) = \sum_{k=1}^m X_{C_{k,a}^{\pi}}(\omega) \mathbb{I}_{\{C_{k,a}^{\pi} \leq t\}}(\omega) = \sum_{k=1}^m X_{C_{k,a}^{\pi}}(\omega) \sum_{t'=1}^t \mathbb{I}_{\{C_{k,a}^{\pi} = t'\}}(\omega) = \sum_{t'=1}^t X_{t'}(\omega) \sum_{k=1}^m \mathbb{I}_{\{C_{k,a}^{\pi} = t'\}}(\omega).$$

Note that  $\{C_{k,a}^{\pi} = t'\} \cap \{C_{k',a}^{\pi} = t'\} = \emptyset$  if  $k \neq k'$  and  $t' \in \mathbb{N}^+$ .

Let  $t' \leq t$  and  $A_{t'} = \pi_{t'}(X_0, \dots, X_{t'-1})$ . Since  $A_{t'}(\omega) = a$  if and only if  $C_{k,a}^{\pi}(\omega) = t'$  for some  $k \leq m$ ,

$$\sum_{k=1}^m X_{C_{k,a}^{\pi}}(\omega) = \sum_{t'=1}^t X_{t'}(\omega) \mathbb{I}_{\bigcup_{k=1}^m \{C_{k,a}^{\pi} = t'\}}(\omega) = \sum_{t'=1}^t X_{t'}(\omega) \mathbb{I}_{\{A_{t'} = a\}}(\omega).$$

Therefore, for every  $\delta \in (0, 1)$ ,

$$U_{t,a}^{\pi, \delta}(\omega) = \frac{1}{T_{t,a}^{\pi}(\omega)} \sum_{k=1}^t X_k(\omega) \mathbb{I}_{\{A_k = a\}}(\omega) + \sqrt{\frac{2 \log(1/\delta)}{T_{t,a}^{\pi}(\omega)}} = \frac{1}{m} \sum_{k=1}^m X_{C_{k,a}^{\pi}}(\omega) + \sqrt{\frac{2 \log(1/\delta)}{m}}.$$

□

**Definition 8.2.** A policy  $\pi$  implements upper confidence bounds with error  $\delta \in (0, 1)$  if, for every  $t \in \mathbb{N}^+$ ,

$$\pi_t(X_0, \dots, X_{t-1}) = \begin{cases} t, & \text{if } t \leq n, \\ \arg \max_a U_{t-1,a}^{\pi, \delta}, & \text{if } t > n. \end{cases}$$

Note that  $U_{t-1,a}^{\pi, \delta}$  is well-defined for every time step  $t > n$  and action  $a \in \mathcal{A}$ .

**Theorem 8.1.** If  $\nu$  is a 1-subgaussian stochastic bandit and the policy  $\pi$  implements upper confidence bounds with error  $\delta = 1/t^2$  for some  $t \in \mathbb{N}^+$ , then

$$R_t^{\nu, \pi} \leq \left( 3 \sum_{a=1}^n \Delta_a^{\nu} \right) + \sum_{a \mid \Delta_a^{\nu} > 0} \frac{16 \log(t)}{\Delta_a^{\nu}}.$$

*Proof.* If  $t \leq n$ , then  $T_{t,a}^{\pi} \leq 1$  for every  $a \in \mathcal{A}$ , so that  $R_t^{\nu, \pi} = \sum_a \Delta_a^{\nu} \mathbb{E}^{\nu, \pi}(T_{t,a}^{\pi}) \leq \sum_a \Delta_a^{\nu}$ .

Let  $t > n$  and consider an action  $a \in \mathcal{A}$  such that  $\Delta_a^{\nu} > 0$ . For every  $m \in \mathbb{N}^+$ , since  $T_{t,a}^{\pi} \leq t$ ,

$$\mathbb{E}^{\nu, \pi}(T_{t,a}^{\pi}) = \mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{T_{t,a}^{\pi} > m\}} T_{t,a}^{\pi}) + \mathbb{E}^{\nu, \pi}(\mathbb{I}_{\{T_{t,a}^{\pi} \leq m\}} T_{t,a}^{\pi}) \leq t \mathbb{P}^{\nu, \pi}(T_{t,a}^{\pi} > m) + m.$$

Let  $\delta = 1/t^2$  and  $m = \lceil 8 \log(1/\delta)/(\Delta_a^\nu)^2 \rceil$ , so that  $m \in \mathbb{N}^+$ . Furthermore, consider the event  $E$  given by

$$E = \left\{ \frac{1}{m} \sum_{k=1}^m X_{C_{k,a}^\pi} + \sqrt{\frac{2 \log(1/\delta)}{m}} < \mu_*^\nu \right\}.$$

Because the events  $E$  and  $E^c$  are disjoint,

$$\mathbb{P}^{\nu,\pi}(T_{t,a}^\pi > m) = \mathbb{P}^{\nu,\pi}(\{T_{t,a}^\pi > m\} \cap E) + \mathbb{P}^{\nu,\pi}(\{T_{t,a}^\pi > m\} \cap E^c).$$

We will consider the two terms on the right side of the equation above separately.

First, consider an action  $a^* \in \mathcal{A}$  such that  $\mu_{a^*}^\nu = \mu_*^\nu$ , so that  $a^* \neq a$ . Furthermore, consider an  $\omega \in E$  such that  $T_{t,a}^\pi(\omega) > m$ . In order to find a contradiction, suppose that  $\mu_*^\nu < U_{t'-1,a^*}^{\pi,\delta}(\omega)$  for every  $t' \in \mathbb{N}^+$  such that  $n < t' \leq t$ . Since  $T_{t,a}^\pi(\omega) > m$ , there is a  $t' \in \mathbb{N}^+$  such that  $C_{m+1,a}^\pi(\omega) = t'$  and  $n < t' \leq t$ . Therefore,

$$\pi_{t'}(X_0(\omega), \dots, X_{t'-1}(\omega)) = \arg \max_{a'} U_{t'-1,a'}^{\pi,\delta}(\omega) = a.$$

By Proposition 8.1, since  $T_{t'-1,a}^\pi(\omega) = m$  and  $\omega \in E$ ,

$$U_{t'-1,a}^{\pi,\delta}(\omega) = \frac{1}{m} \sum_{k=1}^m X_{C_{k,a}^\pi}(\omega) + \sqrt{\frac{2 \log(1/\delta)}{m}} < \mu_*^\nu < U_{t'-1,a^*}^{\pi,\delta}(\omega),$$

which is a contradiction because  $U_{t'-1,a}^{\pi,\delta}(\omega) = \sup_{a'} U_{t'-1,a'}^{\pi,\delta}(\omega)$ .

Therefore, if  $\omega \in E$  and  $T_{t,a}^\pi(\omega) > m$ , then  $\mu_*^\nu \geq U_{t'-1,a^*}^{\pi,\delta}(\omega)$  for some  $t' \in \mathbb{N}^+$  such that  $n < t' \leq t$ . Consequently, there is an  $m' \in \mathbb{N}^+$  such that  $m' \leq t$  and  $T_{t,a^*}^\pi(\omega) \geq m'$  and

$$\mu_*^\nu \geq \frac{1}{m'} \sum_{k=1}^{m'} X_{C_{k,a^*}^\pi}(\omega) + \sqrt{\frac{2 \log(1/\delta)}{m'}}.$$

From the previous statement,

$$\mathbb{P}^{\nu,\pi}(\{T_{t,a}^\pi > m\} \cap E) \leq \mathbb{P}^{\nu,\pi} \left( \bigcup_{m' \leq t} \left\{ T_{t,a^*}^\pi \geq m', \mu_*^\nu \geq \frac{1}{m'} \sum_{k=1}^{m'} X_{C_{k,a^*}^\pi} + \sqrt{\frac{2 \log(1/\delta)}{m'}} \right\} \right).$$

By the union bound, the fact that  $T_{t,a^*}^\pi(\omega) \geq m'$  implies  $C_{m',a^*}^\pi(\omega) < \infty$ , and Proposition 7.8,

$$\mathbb{P}^{\nu,\pi}(\{T_{t,a}^\pi > m\} \cap E) \leq \sum_{m' \leq t} \mathbb{P}^{\nu,\pi} \left( C_{m',a^*}^\pi < \infty, \mu_*^\nu \geq \frac{1}{m'} \sum_{k=1}^{m'} X_{C_{k,a^*}^\pi} + \sqrt{\frac{2 \log(1/\delta)}{m'}} \right) \leq t\delta.$$

Second, consider an  $\omega \in E^c$  such that  $T_{t,a}^\pi(\omega) > m$ . Since  $C_{m,a}^\pi(\omega) < \infty$ ,

$$\mathbb{P}^{\nu,\pi}(\{T_{t,a}^\pi > m\} \cap E^c) \leq \mathbb{P}^{\nu,\pi}(\{C_{m,a}^\pi < \infty\} \cap E^c) = \mathbb{P}^{\nu,\pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m X_{C_{k,a}^\pi} + \sqrt{\frac{2 \log(1/\delta)}{m}} \geq \mu_*^\nu \right).$$

By subtracting  $\mu_a^\nu + \sqrt{2 \log(1/\delta)/m}$  from both sides of an inequality above and the definition of  $\Delta_a^\nu$ ,

$$\mathbb{P}^{\nu,\pi}(\{T_{t,a}^\pi > m\} \cap E^c) \leq \mathbb{P}^{\nu,\pi} \left( C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \Delta_a^\nu - \sqrt{\frac{2 \log(1/\delta)}{m}} \right).$$

Since  $m \geq 8 \log(1/\delta)/(\Delta_a^\nu)^2$ , note that  $\sqrt{2 \log(1/\delta)/m} \leq \Delta_a^\nu/2 = \Delta_a^\nu - \Delta_a^\nu/2$  and

$$\Delta_a^\nu - \sqrt{\frac{2 \log(1/\delta)}{m}} \geq \frac{\Delta_a^\nu}{2}.$$



Therefore, by the previous inequality and Proposition 7.7,

$$\mathbb{P}^{\nu, \pi}(\{T_{t,a}^\pi > m\} \cap E^c) \leq \mathbb{P}^{\nu, \pi}\left(C_{m,a}^\pi < \infty, \frac{1}{m} \sum_{k=1}^m (X_{C_{k,a}^\pi} - \mu_a^\nu) \geq \frac{\Delta_a^\nu}{2}\right) \leq e^{-\frac{m(\Delta_a^\nu)^2}{8}}.$$

By returning to a previous equation,

$$\mathbb{P}^{\nu, \pi}(T_{t,a}^\pi > m) = \mathbb{P}^{\nu, \pi}(\{T_{t,a}^\pi > m\} \cap E) + \mathbb{P}^{\nu, \pi}(\{T_{t,a}^\pi > m\} \cap E^c) \leq t\delta + e^{-\frac{m(\Delta_a^\nu)^2}{8}}.$$

By returning to a previous inequality, since  $\delta = 1/t^2$ ,

$$\mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq t\mathbb{P}^{\nu, \pi}(T_{t,a}^\pi > m) + m \leq te^{-\frac{m(\Delta_a^\nu)^2}{8}} + m + 1.$$

Since  $m \geq 8 \log(1/\delta)/(\Delta_a^\nu)^2$  implies  $-m(\Delta_a^\nu)^2/8 \leq \log \delta$ ,

$$\mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq t\delta + m + 1 = \frac{1}{t} + m + 1 \leq 2 + m \leq 3 + \frac{8 \log(1/\delta)}{(\Delta_a^\nu)^2} = 3 + \frac{16 \log(t)}{(\Delta_a^\nu)^2}.$$

For every  $t > n$ , since  $\mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq 3 + 16 \log(t)/(\Delta_a^\nu)^2$  for every  $a \in \mathcal{A}$  such that  $\Delta_a^\nu > 0$ ,

$$R_t^{\nu, \pi} = \sum_{a|\Delta_a^\nu > 0} \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq \sum_{a|\Delta_a^\nu > 0} \Delta_a^\nu \left(3 + \frac{16 \log(t)}{(\Delta_a^\nu)^2}\right) = \left(3 \sum_{a=1}^n \Delta_a^\nu\right) + \sum_{a|\Delta_a^\nu > 0} \frac{16 \log(t)}{\Delta_a^\nu}.$$

□

**Theorem 8.2.** If  $\nu$  is a 1-subgaussian stochastic bandit and the policy  $\pi$  implements upper confidence bounds with error  $\delta = 1/t^2$  for some  $t \in \mathbb{N}^+$ , then

$$R_t^{\nu, \pi} \leq 8\sqrt{tn \log(t)} + 3 \sum_{a=1}^n \Delta_a^\nu.$$

*Proof.* If  $t \leq n$ , then  $T_{t,a}^\pi \leq 1$  for every  $a \in \mathcal{A}$ , so that  $R_t^{\nu, \pi} = \sum_a \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq \sum_a \Delta_a^\nu$ .

Let  $t > n$ . For every  $\Delta > 0$ , since  $\sum_a \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) = t$ ,

$$R_t^{\nu, \pi} = \left(\sum_{a|\Delta_a^\nu < \Delta} \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi)\right) + \left(\sum_{a|\Delta_a^\nu \geq \Delta} \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi)\right) \leq t\Delta + \sum_{a|\Delta_a^\nu \geq \Delta} \Delta_a^\nu \mathbb{E}^{\nu, \pi}(T_{t,a}^\pi).$$

From the proof of Theorem 8.1, recall that  $\mathbb{E}^{\nu, \pi}(T_{t,a}^\pi) \leq 3 + 16 \log(t)/(\Delta_a^\nu)^2$  if  $\Delta_a^\nu > 0$ . Therefore,

$$R_t^{\nu, \pi} \leq t\Delta + \sum_{a|\Delta_a^\nu \geq \Delta} \Delta_a^\nu \left(3 + \frac{16 \log(t)}{(\Delta_a^\nu)^2}\right) \leq t\Delta + \left(\sum_{a|\Delta_a^\nu \geq \Delta} \frac{16 \log(t)}{\Delta_a^\nu}\right) + 3 \sum_{a=1}^n \Delta_a^\nu.$$

Let  $\Delta = \sqrt{16n \log(t)}/t$ , so that  $\Delta > 0$ . Since  $\Delta_a^\nu \geq \Delta$  implies  $16 \log(t)/\Delta_a^\nu \leq 16 \log(t)/\Delta$ ,


$$R_t^{\nu, \pi} \leq t\Delta + \frac{16n \log(t)}{\Delta} + 3 \sum_{a=1}^n \Delta_a^\nu = \sqrt{t} \sqrt{16n \log(t)} + \sqrt{t} \sqrt{16n \log(t)} + 3 \sum_{a=1}^n \Delta_a^\nu = 8\sqrt{tn \log(t)} + 3 \sum_{a=1}^n \Delta_a^\nu.$$

□

## Acknowledgements

I would like to thank Daniel Valesin for the ideas behind some of the proofs found in these notes.

## License

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License .

## References

- [1] Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein, C. *Introduction to algorithms*. MIT press, 2022.
- [2] Kaczor, W.J., Nowak, M.T. *Problems in Mathematical Analysis I*. American Mathematical Society, 2000.
- [3] Lattimore, T., and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.
- [4] Rivasplata, O. *Subgaussian random variables: An expository note*. 2012.
- [5] Wainwright, M. J. *High-Dimensional Statistics - A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.