Reinforcement Learning Theory

Paulo Rauber

2024

1 Asymptotic analysis

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

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 $\liminf_{n\to\infty} f(n)$ is defined by

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

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 $z < \liminf_{n\to\infty} f(n)$, then $z < f(n)$ for all sufficiently large $n \in \mathbb{N}$.

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

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

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

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 $z > \limsup_{n\to\infty} f(n)$, then $z > f(n)$ for all sufficiently large $n \in \mathbb{N}$.

Proposition 1.4. If $z < \limsup_{n\to\infty} f(n)$, 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

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

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

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

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

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

Definition 1.7. For every $f \in \mathscr{F}$ and $g \in \mathscr{F}$, let $(f/g) \in \mathscr{F}$ 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 $g \in \mathscr{F}$, then the following subsets of $\mathscr{F}$ are defined:

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

Consider a real number $a > 0$.

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

- $(n \mapsto an) \in o(n \mapsto n^2)$, 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 $\lim_{n\to\infty} n^2/an = \limsup_{n\to\infty} n^2/an = \liminf_{n\to\infty} n^2/an = \infty$:

- $(n \mapsto n^2) \notin o(n \mapsto an)$, 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 $\lim_{n\to\infty} an^2/n^2 = \limsup_{n\to\infty} an^2/n^2 = \liminf_{n\to\infty} an^2/n^2 = a$:

- $(n \mapsto an^2) \notin o(n \mapsto n^2)$, 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 $f \in \mathscr{F}$ and $g \in \mathscr{F}$, unless the product on the right side below is $0 \cdot \infty$ or $\infty \cdot 0$,

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

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

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

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

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

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

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

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

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

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

Proposition 1.10. If $f \in \mathscr{F}$, 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,

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

□

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

Proof. By Proposition 1.6,

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

□

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

Proof. By Proposition 1.7,

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

□

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

Proof. By Proposition 1.7,

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

□

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 $f \in \mathscr{F}$ and $g \in \mathscr{F}$, 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 $\limsup_{n\to\infty} f(n)/g(n) \in (0, \infty)$. In that case, $g \in \Omega(f)$, since

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

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

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

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

Proposition 1.16. If $f \in \mathscr{F}$ and $g \in \mathscr{F}$, 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. □

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

- $f \prec g$ if and only if $f \in o(g)$.
- $f \precsim g$ if and only if $f \in O(g)$.

- $f \succsim g$ 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 preoder:

- 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 $f \succ h$.

□

Proposition 1.18. The binary relations $\precsim$ and $\succsim$ are preorders.

Proof. By the definition of preorder:

- It is true that $f \precsim f$. If $f \precsim g$ and $g \precsim h$, then $f \precsim h$.
- It is true that $f \succsim f$. If $f \succsim g$ and $g \succsim h$, then $f \succsim h$.

□

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 $f \sim h$. □

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

1. If $f \prec g$, then $f \precsim g$.
2. If $f \succ g$, then $f \succsim g$.
3. If $f \precsim g$ and $g \precsim f$, then $f \sim g$.
4. If $f \succsim g$ and $g \succsim f$, then $f \sim g$.
5. If $f \prec g$, then not $f \succsim g$.
6. If $f \succ g$, then not $f \precsim g$.

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 $\liminf_{n\to\infty} f(n)/g(n) \leq \limsup_{n\to\infty} f(n)/g(n)$. □

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

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

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

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

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

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

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,

$$\limsup_{n\to\infty} \frac{f(n)}{n^2} = \limsup_{n\to\infty} \frac{n^2 + l(n)}{n^2} = 1 + \limsup_{n\to\infty} \frac{l(n)}{n^2} < \infty,$$
$$\liminf_{n\to\infty} \frac{f(n)}{n^2} = \liminf_{n\to\infty} \frac{n^2 + l(n)}{n^2} = 1 + \liminf_{n\to\infty} \frac{l(n)}{n^2} > 0,$$

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 $X : \Omega \to \mathbb{R}$ is 0-subgaussian if and only if $\mathbb{P}(X = 0) = 1$.

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

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

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

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

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 $\mathbb{E}\left(e^{\lambda X}\mathbb{I}_{\{X\geq 0\}}\right) \leq \mathbb{E}\left(e^{\lambda X}\right) \leq e^{\frac{\lambda^2\sigma^2}{2}}$ and $\mathbb{E}\left(e^{-\lambda X}\mathbb{I}_{\{X<0\}}\right) \leq \mathbb{E}\left(e^{-\lambda X}\right) \leq e^{\frac{(-\lambda)^2\sigma^2}{2}} = e^{\frac{\lambda^2\sigma^2}{2}}$. Therefore,

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

□

Proposition 2.2. If a random variable $X : \Omega \to \mathbb{R}$ is σ-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 $\phi : \mathbb{R} \to \mathbb{R}$ 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,

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

□

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

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

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

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 σ-subgaussian random variable,

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

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

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

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

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

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

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

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,

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

so that

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

□

Proposition 2.4. If a random variable $X : \Omega \to \mathbb{R}$ is σ-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 $X_1 : \Omega \to \mathbb{R}$ is σ_1-subgaussian, the random variable X_2 is σ_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 $X_1 : \Omega \to \mathbb{R}$ is σ_1-subgaussian and the random variable X_2 is σ_2-subgaussian, then $X_1 + X_2$ is $(\sigma_1 + \sigma_2)$-subgaussian.

Proof. Note that $\mathbb{E}\left(|e^{\lambda X_1}|^p\right) = \mathbb{E}\left(e^{\lambda p X_1}\right) < \infty$ and $\mathbb{E}\left(|e^{\lambda X_2}|^q\right) = \mathbb{E}\left(e^{\lambda q X_2}\right) < \infty$ 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

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

By the definition of subgaussian random variables,

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

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 $\lambda \in \mathbb{R}$,

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

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

Proposition 2.7. If a random variable $X : \Omega \to \mathbb{R}$ 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,

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

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 λ and variance 1. □

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

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

Lemma 2.1 (Hoeffding's lemma)**.** If $X : \Omega \to \mathbb{R}$ 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 $X : \Omega \to \mathbb{R}$ is a σ-subgaussian random variable, then, for every $\epsilon \geq 0$,

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

Proof. Recall that the function $g : \mathbb{R} \to [0, \infty]$ 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,

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

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

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

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

$$\mathbb{P}(X \leq -\epsilon) \leq e^{\frac{\epsilon^2}{\sigma^4}\frac{\sigma^2}{2} - \frac{\epsilon^2}{\sigma^2}} = e^{\frac{\epsilon^2}{2\sigma^2} - \frac{\epsilon^2}{\sigma^2}} = e^{\frac{\epsilon^2}{\sigma^2}\left(\frac{1}{2} - 1\right)} = e^{-\frac{\epsilon^2}{2\sigma^2}},$$
$$\mathbb{P}(X \geq \epsilon) \leq e^{\frac{\epsilon^2}{\sigma^4}\frac{\sigma^2}{2} - \frac{\epsilon^2}{\sigma^2}} = e^{\frac{\epsilon^2}{2\sigma^2} - \frac{\epsilon^2}{\sigma^2}} = e^{\frac{\epsilon^2}{\sigma^2}\left(\frac{1}{2} - 1\right)} = e^{-\frac{\epsilon^2}{2\sigma^2}}.$$

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 $X : \Omega \to \mathbb{R}$ is a σ-subgaussian random variable, then, for every $\delta \in (0, 1]$,

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

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 $X : \Omega \to \mathbb{R}$ is a σ-subgaussian random variable, then, for every $\delta \in (0, 1]$,

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

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 $(X_k : \Omega \to \mathbb{R} \mid k \in \mathbb{N}^+)$, 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 $M_t : \Omega \to \mathbb{R}$ 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 $\mathrm{Var}(M_t) = \mathrm{Var}(X)/t$.

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,

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

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

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

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

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

□

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

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

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

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

□

Proposition 3.5. If $X - \mu$ is a σ-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

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

□

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

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

Proof. Recall that $\mathbb{E}(X - \mu) = 0$ and $\mathrm{Var}(X - \mu) = \mathrm{Var}(X)$. 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 σ-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,

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

□

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

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

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 σ-subgaussian random variable, then, for every $t \in \mathbb{N}^+$ and $\delta \in (0, 1]$,

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

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 $(Y_k : \Omega \to \mathbb{R} \mid k \in \mathbb{N}^+)$ 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 $\mathbb{E}\left(Y_k - \mathbb{E}(Y_k)\right) = 0$ 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,

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

□

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

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

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 $k \in \{1, \ldots, t\}$ 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,

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

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

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

□

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 $h : \mathbb{R} \to \mathbb{R}$ 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 ν-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 ν-integrable, the mean μ_a^ν of action a is defined by $\mu_a^\nu = \int_{\mathbb{R}} x \; P_a(dx)$ and the supremum mean μ_*^ν is defined by $\mu_*^\nu = \sup_a \mu_a^\nu$. If $\mu_a^\nu = \mu_*^\nu$ for some action $a \in \mathcal{A}$, then ν 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 ν-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 $a \in \mathcal{A}$. □

Definition 4.4. For a set of actions $\mathcal{A}$, a policy π is a sequence of functions $(\pi_t : \mathbb{R}^t \to \mathcal{A} \mid t \in \mathbb{N}^+)$, where the so-called policy π_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 $(X_t : \Omega \to \mathbb{R} \mid t \in \mathbb{N})$ such that $\mathbb{E}(|X_t|) < \infty$ and

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

almost surely for every $t \in \mathbb{N}^+$ and $B \in \mathcal{B}(\mathbb{R})$, where $A_t = \pi_t(X_0, \ldots, X_{t-1})$. Additionally, if a function $h : \mathbb{R} \to \mathbb{R}$ is ν-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 $\{Z_{t,a} : \Omega \to \mathbb{R} \mid t \in \mathbb{N}^+ \text{ and } a \in \mathcal{A}\}$ 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 $A_t : \Omega \to \mathcal{A}$ and $X_t : \Omega \to \mathbb{R}$ be given by

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

where $X_0 : \Omega \to \mathbb{R}$ 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 $\sigma(\mathbb{I}_{\{A_t = a\}}) \subseteq \sigma(A_t) \subseteq \sigma(X_0, \ldots, X_{t-1}) \subseteq \mathcal{F}_{t-1}$. 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 $h : \mathbb{R} \to \mathbb{R}$ is ν-integrable, then $\mathbb{E}\left(|h(X_t)|\right) < \infty$ for every $t \in \mathbb{N}^+$, since

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

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

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

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

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,

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

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

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

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

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

□

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 $X_t : \Omega \to \mathbb{R}$ is given by $X_t(\omega) = \omega_t$ and the σ-algebra $\mathcal{F}$ on Ω is given by $\mathcal{F} = \sigma(X_0, X_1, \ldots)$.

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 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

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

almost surely for every $t \in \mathbb{N}^+$ and $B \in \mathcal{B}(\mathbb{R})$, where $A_t = \pi_t(X_0, \ldots, X_{t-1})$. Additionally, if a function $h : \mathbb{R} \to \mathbb{R}$ is ν-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 ν under the policy π.

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 $(\tilde{X}_t^{\nu,\pi} : \tilde{\Omega}^{\nu,\pi} \to \mathbb{R} \mid t \in \mathbb{N})$ such that, almost surely,

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

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

Consider the function $\tilde{X}^{\nu,\pi} : \tilde{\Omega}^{\nu,\pi} \to \Omega$ 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 $\mathbb{P}^{\nu,\pi} : \mathcal{F} \to [0,1]$ 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 $\sigma(\tilde{X}_0^{\nu,\pi}, \ldots, \tilde{X}_t^{\nu,\pi})/\sigma(X_0, \ldots, X_t)$-measurable for every $t \in \mathbb{N}^+$, let $\mathcal{I}_t$ be given by

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

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

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

which uses the fact that

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

Since $(\tilde{X}^{\nu,\pi})^{-1}(I_t) \in \sigma(\tilde{X}_0^{\nu,\pi}, \ldots, \tilde{X}_t^{\nu,\pi})$ for every $I_t \in \mathcal{I}_t$, $\tilde{X}^{\nu,\pi}$ is $\sigma(\tilde{X}_0^{\nu,\pi}, \ldots, \tilde{X}_t^{\nu,\pi})/\sigma(X_0, \ldots, X_t)$-measurable.

For every $t \in \mathbb{N}^+$ and $H_{t-1} \in \sigma(X_0, \ldots, X_{t-1})$, let $\tilde{H}_{t-1} = (\tilde{X}^{\nu,\pi})^{-1}(H_{t-1})$. For every $B \in \mathcal{B}(\mathbb{R})$,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In particular, because the identity function is ν-integrable, $\mathbb{E}^{\nu,\pi}\left(|X_t|\right) < \infty$ 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 ν under the policy π.

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

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

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

Proof. Since the function $h : \mathbb{R} \to \mathbb{R}$ is ν-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,

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

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 α_k is the k-th staircase function. Almost surely, since $h_k(X_t) \uparrow h(X_t)$,

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

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

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

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

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

By the linearity of the integral, almost surely,

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

□

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

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

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

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 $A_t = \pi_t(X_0,\ldots,X_{t-1})$, then

$$\mathbb{E}^{\nu,\pi}\left(X_t\right) = \mathbb{E}^{\nu,\pi}\left(\mathbb{E}^{\nu,\pi}\left(X_t \mid A_t\right)\right) = \mathbb{E}^{\nu,\pi}\left(\mu^\nu_{A_t}\right) = \sum_a \mu^\nu_a \mathbb{P}^{\nu,\pi}\left(A_t = a\right).$$

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 $R^{\nu,\pi}_t$ of policy π on ν after t time steps is given by

$$R^{\nu,\pi}_t = t\mu^\nu_* - \sum_{k=1}^t \mathbb{E}^{\nu,\pi}\left(X_k\right).$$

Definition 4.8. For every action $a \in \mathcal{A}$, the suboptimality gap is defined by $\Delta^\nu_a = \mu^\nu_* - \mu^\nu_a$, so that $\Delta^\nu_a \geq 0$.

Definition 4.9. The number of times $T^\pi_{t,a} : \Omega \to \{0,\ldots,t\}$ that policy π selects $a \in \mathcal{A}$ by time $t \in \mathbb{N}^+$ is given by

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

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

Definition 4.10. The average reward $M^\pi_{t,a} : \Omega \to \mathbb{R}$ that policy π observes for $a \in \mathcal{A}$ by time $t \in \mathbb{N}^+$ is given by

$$M^\pi_{t,a}(\omega) = \frac{1}{T^\pi_{t,a}(\omega)} \sum_{k=1}^t X_k(\omega)\mathbb{I}_{\{A_k=a\}}(\omega)$$

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

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

$$R^{\nu,\pi}_t = \sum_a \Delta^\nu_a \mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right).$$

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

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

By the definition of the regret $R^{\nu,\pi}_t$ of policy π on ν after t time steps,

$$R^{\nu,\pi}_t = t\mu^\nu_* - \sum_{k=1}^t \mathbb{E}^{\nu,\pi}\left(X_k\right) = \sum_{k=1}^t \sum_a \mu^\nu_* \mathbb{P}^{\nu,\pi}\left(A_k = a\right) - \sum_{k=1}^t \sum_a \mu^\nu_a \mathbb{P}^{\nu,\pi}\left(A_k = a\right).$$

By rearranging terms and the definition of suboptimality gap,

$$R^{\nu,\pi}_t = \sum_{k=1}^t \sum_a (\mu^\nu_* - \mu^\nu_a)\mathbb{P}^{\nu,\pi}\left(A_k = a\right) = \sum_a \Delta^\nu_a \sum_{k=1}^t \mathbb{P}^{\nu,\pi}\left(A_k = a\right) = \sum_a \Delta^\nu_a \mathbb{E}^{\nu,\pi}(T^\pi_{t,a}).$$

□

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

Proof. Since $\Delta_a^\nu \geq 0$ and $\mathbb{E}^{\nu,\pi}\left(T_{t,a}^\pi\right) \geq 0$ 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 $A_k = \pi_k(X_0, \ldots, X_{k-1})$ 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,

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

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

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

Proof. Since $R_\cdot^{\nu,\pi} : \mathbb{N} \to \mathbb{R}$ is asymptotically positive by assumption,

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

so that

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

$\square$

Definition 4.11. The number of times $T_{t,*}^{\nu,\pi} : \Omega \to \{0, \ldots, t\}$ that policy π selects an optimal action on the stochastic bandit ν 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 $A_k = \pi_k(X_0, \ldots, X_{k-1})$ for every $k \leq t$.

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

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

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,

$$\{\Delta^\nu_{A_k} = 0\} = \bigcup_a \{A_k = a \text{ and } \Delta^\nu_a = 0\} = \bigcup_{a | \Delta^\nu_a = 0} \{A_k = a\},$$

so that

$$T^{\nu,\pi}_{t,*}(\omega) = \sum_{k=1}^{t} \mathbb{I}_{\{\Delta^\nu_{A_k} = 0\}}(\omega) = \sum_{k=1}^{t} \sum_{a | \Delta^\nu_a = 0} \mathbb{I}_{\{A_k = a\}}(\omega) = \sum_{a | \Delta^\nu_a = 0} \sum_{k=1}^{t} \mathbb{I}_{\{A_k = a\}}(\omega) = \sum_{a | \Delta^\nu_a = 0} T^\pi_{t,a}(\omega).$$

□

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

$$\lim_{t \to \infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\nu,\pi}_{t,*}\right)}{t} = 1.$$

Proof. By Theorem 4.2,

$$0 = \lim_{t \to \infty} \frac{R^{\nu,\pi}_t}{t} = \lim_{t \to \infty} \frac{\sum_a \Delta^\nu_a \mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right)}{t} = \lim_{t \to \infty} \sum_a \Delta^\nu_a \frac{\mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right)}{t} = \sum_a \Delta^\nu_a \lim_{t \to \infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right)}{t},$$

so that $\lim_{t \to \infty} \mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right)/t = 0$ whenever $\Delta^\nu_a > 0$. Therefore,

$$0 = \sum_{a | \Delta^\nu_a > 0} \lim_{t \to \infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right)}{t} = \lim_{t \to \infty} \sum_{a | \Delta^\nu_a > 0} \frac{\mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right)}{t}.$$

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

$$t = \sum_a \mathbb{E}^{\nu,\pi}(T^\pi_{t,a}) = \sum_{a | \Delta^\nu_a = 0} \mathbb{E}^{\nu,\pi}(T^\pi_{t,a}) + \sum_{a | \Delta^\nu_a > 0} \mathbb{E}^{\nu,\pi}(T^\pi_{t,a}) = \mathbb{E}^{\nu,\pi}\left(T^{\nu,\pi}_{t,*}\right) + \sum_{a | \Delta^\nu_a > 0} \mathbb{E}^{\nu,\pi}(T^\pi_{t,a}),$$

so that

$$\sum_{a | \Delta^\nu_a > 0} \frac{\mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right)}{t} = 1 - \frac{\mathbb{E}^{\nu,\pi}\left(T^{\nu,\pi}_{t,*}\right)}{t}.$$

Therefore, considering a previous equation,

$$0 = \lim_{t \to \infty} 1 - \frac{\mathbb{E}^{\nu,\pi}\left(T^{\nu,\pi}_{t,*}\right)}{t} = 1 - \lim_{t \to \infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\nu,\pi}_{t,*}\right)}{t}.$$

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

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 $R^{\cdot,\pi}_t : \mathcal{E} \to [0, \infty]$ is $\mathcal{G}$-measurable for every policy π and time step $t \in \mathbb{N}^+$. The Bayesian regret B^π_t of policy π after $t \in \mathbb{N}^+$ time steps is given by

$$B^\pi_t = \int_{\mathcal{E}} R^{\nu,\pi}_t Q(d\nu).$$

Definition 4.14. The stochastic bandit $\nu = (P_a \mid a \in \mathcal{A})$ is σ-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 $Z_a(x) = x - \mu^\nu_a$ is σ-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 $(Y_n : \Omega \to \mathbb{R} \mid n \in \mathbb{N})$.

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

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

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

Proof. Recall that the function $\sup_n Y_n$ is $\mathcal{F}$-measurable, so that the function $Z_m : \Omega \to \mathbb{N} \cup \{\infty\}$ 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 $\mathcal{A} = \{1, \ldots, n\}$, 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 ν under the policy π.

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

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

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

Proposition 5.2. If the policy π 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 $A_t = \pi_t(X_0, \ldots, X_{t-1})$, so that $A_t = a_t$. For every $B \in \mathcal{B}(\mathbb{R})$,

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

□

Proposition 5.3. If the policy π implements explore-then-commit with $m \in \mathbb{N}^+$ exploration steps, then the random variables $X_0, X_1, \ldots, X_{mn}$ 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 $X_0, X_1, \ldots, X_t$ are independent for some $t \in \mathbb{N}^+$ such that $t < mn$. We will show that $X_0, X_1, \ldots, X_{t+1}$ are independent.

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

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

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

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

□

Proposition 5.4. If the policy π implements explore-then-commit with $m \in \mathbb{N}^+$ exploration steps and ν is a 1-subgaussian stochastic bandit, then $X_t - \mu^\nu_{a_t}$ 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 $Z_a(x) = x - \mu^\nu_a$. By Proposition 5.2, the law of X_t is P_{a_t} for every $t \in \{1, \dots, mn\}$. For every $\lambda \in \mathbb{R}$,

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

□

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

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

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

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

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

$$R^{\nu,\pi}_{mn} = \sum_{a=1}^{n}\Delta^\nu_a \mathbb{E}^{\nu,\pi}\left(T^\pi_{mn,a}\right) = m\sum_{a=1}^{n}\Delta^\nu_a.$$

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

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

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

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

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

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

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

$$\mathbb{P}^{\nu,\pi}\left(M^\pi_{mn,a} \geq \sup_{a'} M^\pi_{mn,a'}\right) \leq \mathbb{P}^{\nu,\pi}\left(M^\pi_{mn,a} - M^\pi_{mn,a^*} \geq 0\right) = \mathbb{P}^{\nu,\pi}\left(M^\pi_{mn,a} - M^\pi_{mn,a^*} + (\mu^\nu_{a^*} - \mu^\nu_a) \geq \Delta^\nu_a\right),$$

so that

$$\mathbb{P}^{\nu,\pi}\left(M^\pi_{mn,a} \geq \sup_{a'} M^\pi_{mn,a'}\right) \leq \mathbb{P}^{\nu,\pi}\left(\left(M^\pi_{mn,a} - \mu^\nu_a\right) - \left(M^\pi_{mn,a^*} - \mu^\nu_{a^*}\right) \geq \Delta^\nu_a\right).$$

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

$$M^\pi_{mn,a}(\omega) - \mu^\nu_a = \left(\frac{1}{m} \sum_{i=0}^{m-1} X_{a+in}(\omega) \right) - \frac{1}{m} \sum_{i=0}^{m-1} \mu^\nu_a = \frac{1}{m} \sum_{i=0}^{m-1} \left(X_{a+in}(\omega) - \mu^\nu_a \right).$$

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

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

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

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

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

$$R^{\nu,\pi}_t = \sum_{a=1}^{n} \Delta^\nu_a \mathbb{E}^{\nu,\pi} \left(T^\pi_{t,a} \right) \leq \sum_{a=1}^{n} \Delta^\nu_a \left(m + (t - mn) e^{-\frac{m(\Delta^\nu_a)^2}{4}} \right) = \left(m \sum_{a=1}^{n} \Delta^\nu_a \right) + (t - mn) \sum_{a=1}^{n} \Delta^\nu_a e^{-\frac{m(\Delta^\nu_a)^2}{4}}.$$

□

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 $\Delta = \max(\Delta^\nu_1, \Delta^\nu_2)$, 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 π is a policy that implements explore-then-commit with m exploration steps, then

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

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

$$R^{\nu,\pi}_t = \sum_{a=1}^{2} \Delta^\nu_a \mathbb{E}^{\nu,\pi} \left(T^\pi_{t,a} \right) \leq \Delta \sum_{a=1}^{2} \mathbb{E}^{\nu,\pi} \left(T^\pi_{t,a} \right) = \Delta \mathbb{E}^{\nu,\pi} \left(\sum_{a=1}^{2} T^\pi_{t,a} \right) = t\Delta \leq t \frac{2}{\sqrt{t}} = 2\sqrt{t}.$$

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,

$$R^{\nu,\pi}_t = \Delta^\nu_1 \mathbb{E}^{\nu,\pi} \left(T^\pi_{t,1} \right) + \Delta^\nu_2 \mathbb{E}^{\nu,\pi} \left(T^\pi_{t,2} \right) \leq m\Delta.$$

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

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

Because the function $f : (0, \infty) \to (0, \infty)$ 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 $g : (0, \infty) \to \mathbb{R}$ given by $g(x) = x \log(4t/x^2) + x$, so that $g(4/\Delta) = (4/\Delta) \log\left(t\Delta^2/4\right) + 4/\Delta$. 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 $g(x) \leq g\left(2\sqrt{t}/\sqrt{e}\right) = 4\sqrt{t}/\sqrt{e}$ 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 ν under the policy π.

Definition 6.1. A policy π restarts to the policy π' 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 π restarts to the policy π' after $t \in \mathbb{N}$ steps, then

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

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 $B_1 \in \mathcal{B}(\mathbb{R})$,

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

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

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

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

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

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

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

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

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

By taking out what is known,

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

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 $A_{t+k+1} = \pi'_{k+1}(0, X_{t+1}, \dots, X_{t+k})$,

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

where the function $f : \mathbb{R}^k \to [0, 1]$ 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 $(X_1, \dots, X_k)$,

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

□

Proposition 6.2. If a policy π restarts to the policy π' 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 π 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 π 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

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

□

Proposition 6.4. If the policy π 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 $f : \mathbb{N}^+ \to [0, \infty)$ 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 π 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 $R^{\nu,\pi^{(k)}}_{2^{k-1}} \leq \sqrt{2^{k-1}}$ for every $k \in \mathbb{N}^+$, then, for every $t \in \mathbb{N}^+$,

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

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,

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

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

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

□

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 ν under the policy π. 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 $C^\pi_{m,a} : \Omega \to \mathbb{N}^+ \cup \{\infty\}$ until policy π selects $a \in \mathcal{A}$ exactly $m \in \mathbb{N}^+$ times is given by

$$C^\pi_{m,a}(\omega) = \inf \left(\{ t \in \mathbb{N}^+ \mid T^\pi_{t,a}(\omega) \geq m \} \right).$$

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

Proposition 7.1. The time $C^\pi_{m,a} : \Omega \to \mathbb{N}^+ \cup \{\infty\}$ until π selects $a \in \mathcal{A}$ exactly $m \in \mathbb{N}^+$ times is a stopping time.

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

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

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

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

Definition 7.3. For every $a \in \mathcal{A}$, the constant policy $\pi^{(a)} = (\pi^{(a)}_t \mid t \in \mathbb{N}^+)$ is given by $\pi^{(a)}_t = a$ 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 $\pi^{(a)}_m(X_0, \dots, X_{m-1}) = a$,

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

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,

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

□

Proposition 7.3. For every $a \in \mathcal{A}$, $m \in \mathbb{N}^+$, and $t \in \mathbb{N}^+$, if $h : \mathbb{R} \to \mathbb{R}$ is $\mathcal{B}(\mathbb{R})$-measurable, then the function $\mathbb{I}_{\{C^\pi_{m,a} = t\}} \prod_{k=1}^{m-1} h(X_{C^\pi_{k,a}})$ 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 $\{C^\pi_{k,a} = t_k\} = \{C^\pi_{k,a} \leq t_k\} \cap \{C^\pi_{k,a} \leq t_k - 1\}^c$, so that $\{C^\pi_{k,a} = t_k\} \in \mathcal{F}_{t_k - 1}$. For every $\omega \in \Omega$, $m \in \mathbb{N}^+$, and $t \in \mathbb{N}^+$, if $C^\pi_{m,a}(\omega) = t$, then $C^\pi_{1,a}(\omega) < \cdots < C^\pi_{m,a}(\omega) = t$, so

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

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

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

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

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

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

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

which uses the fact that $X_1, \ldots, X_m$ 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,

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

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

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

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

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

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

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

By returning to a previous equation,

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

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

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

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

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

□

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

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

Proof. If ν 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 $\lambda \in \mathbb{R}$,

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

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 $h : \mathbb{R} \to [0,\infty]$ given by $h(x) = e^{\lambda x}$ is ν-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 $\lambda \in \mathbb{R}$,

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

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

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

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

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

For every $\lambda \in \mathbb{R}$, since $\mathbb{I}_{\{C^\pi_{m,a} \le t\}} = \mathbb{I}_{\{C^\pi_{m-1,a} \le t\}}\mathbb{I}_{\{C^\pi_{m,a} \le t\}}$,

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

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

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

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

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

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

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

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

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

□

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

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

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

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

For every $a \in \mathcal{A}$, consider the function $h_a : \mathbb{R} \to [0, \infty]$ 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 ν-integrable, h_a is also ν-integrable. Furthermore, for every $t \in \mathbb{N}^+$,

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

By Proposition 7.4,

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

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

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

Since $X_1 - \mu_a^\nu, \ldots, X_m - \mu_a^\nu$ 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,

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

□

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

$$\mathbb{P}^{\nu,\pi}\left(C^\pi_{m,a} < \infty, \frac{1}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu) \leq -\epsilon\right) \leq e^{-\frac{m\epsilon^2}{2}},$$

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

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

$$\mathbb{I}_{\{C^\pi_{m,a}<\infty\}} e^{-\frac{\lambda}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu)} \geq \mathbb{I}_{\{C^\pi_{m,a}<\infty\}} e^{-\frac{\lambda}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu)} \mathbb{I}_{\{-\frac{1}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu) \geq \epsilon\}},$$

$$\mathbb{I}_{\{C^\pi_{m,a}<\infty\}} e^{\frac{\lambda}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu)} \geq \mathbb{I}_{\{C^\pi_{m,a}<\infty\}} e^{\frac{\lambda}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu)} \mathbb{I}_{\{\frac{1}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu) \geq \epsilon\}}.$$

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

$$\mathbb{I}_{\{C^\pi_{m,a}<\infty\}} e^{-\frac{\lambda}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu)} \geq \mathbb{I}_{\{C^\pi_{m,a}<\infty\}} e^{\lambda\epsilon} \mathbb{I}_{\{-\frac{1}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu) \geq \epsilon\}},$$

$$\mathbb{I}_{\{C^\pi_{m,a}<\infty\}} e^{\frac{\lambda}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu)} \geq \mathbb{I}_{\{C^\pi_{m,a}<\infty\}} e^{\lambda\epsilon} \mathbb{I}_{\{\frac{1}{m}\sum_{k=1}^m (X_{C^\pi_{k,a}} - \mu_a^\nu) \geq \epsilon\}}.$$

By taking expectations of both sides of the inequalities above,

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

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

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

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

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

By rewriting the previous inequalities,

$$\mathbb{P}^{\nu,\pi}\left(C^{\pi}_{m,a} < \infty, \frac{1}{m}\sum_{k=1}^{m}(X_{C^{\pi}_{k,a}} - \mu^{\nu}_a) \leq -\epsilon\right) \leq e^{\frac{\lambda^2}{2m} - \lambda\epsilon},$$
$$\mathbb{P}^{\nu,\pi}\left(C^{\pi}_{m,a} < \infty, \frac{1}{m}\sum_{k=1}^{m}(X_{C^{\pi}_{k,a}} - \mu^{\nu}_a) \geq \epsilon\right) \leq e^{\frac{\lambda^2}{2m} - \lambda\epsilon}.$$

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

$$\mathbb{P}^{\nu,\pi}\left(C^{\pi}_{m,a} < \infty, \frac{1}{m}\sum_{k=1}^{m}(X_{C^{\pi}_{k,a}} - \mu^{\nu}_a) \leq -\epsilon\right) \leq e^{-\frac{m\epsilon^2}{2}},$$
$$\mathbb{P}^{\nu,\pi}\left(C^{\pi}_{m,a} < \infty, \frac{1}{m}\sum_{k=1}^{m}(X_{C^{\pi}_{k,a}} - \mu^{\nu}_a) \geq \epsilon\right) \leq e^{-\frac{m\epsilon^2}{2}}.$$

□

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

$$\mathbb{P}^{\nu,\pi}\left(C^{\pi}_{m,a} < \infty, \frac{1}{m}\sum_{k=1}^{m}(X_{C^{\pi}_{k,a}} - \mu^{\nu}_a) \leq -\sqrt{\frac{2\log(1/\delta)}{m}}\right) \leq \delta,$$
$$\mathbb{P}^{\nu,\pi}\left(C^{\pi}_{m,a} < \infty, \frac{1}{m}\sum_{k=1}^{m}(X_{C^{\pi}_{k,a}} - \mu^{\nu}_a) \geq \sqrt{\frac{2\log(1/\delta)}{m}}\right) \leq \delta.$$

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 $\mathcal{A} = \{1, \ldots, n\}$, 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 ν under the policy π.

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

$$U^{\pi,\delta}_{t,a}(\omega) = M^{\pi}_{t,a}(\omega) + \sqrt{\frac{2\log(1/\delta)}{T^{\pi}_{t,a}(\omega)}}$$

whenever $T^{\pi}_{t,a}(\omega) > 0$. Intuitively, the role of $U^{\pi,\delta}_{t,a}$ is to overestimate μ^{ν}_a with high probability when δ is small.

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

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

whenever $T^{\pi}_{t,a}(\omega) = m$ 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 $T^{\pi}_{t,a}(\omega) = m$, then $C^{\pi}_{k,a}(\omega) \leq t$ for every $k \leq m$, so that

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

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

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

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

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

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

□

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

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

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

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

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

Proof. If $t \leq n$, then $T^{\pi}_{t,a} \leq 1$ for every $a \in \mathcal{A}$, so that $R^{\nu,\pi}_t = \sum_a \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right) \leq \sum_a \Delta^{\nu}_a$.

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

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

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,

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

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,

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

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,

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

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,

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

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

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

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

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

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,

$$\mathbb{P}^{\nu,\pi}\left(\left\{T^{\pi}_{t,a} > m\right\} \cap E^c\right) \leq \mathbb{P}^{\nu,\pi}\left(C^{\pi}_{m,a} < \infty, \frac{1}{m}\sum_{k=1}^{m}\left(X_{C^{\pi}_{k,a}} - \mu^{\nu}_a\right) \geq \frac{\Delta^{\nu}_a}{2}\right) \leq e^{-\frac{m(\Delta^{\nu}_a)^2}{8}}.$$

By returning to a previous equation,

$$\mathbb{P}^{\nu,\pi}\left(T^{\pi}_{t,a} > m\right) = \mathbb{P}^{\nu,\pi}\left(\left\{T^{\pi}_{t,a} > m\right\} \cap E\right) + \mathbb{P}^{\nu,\pi}\left(\left\{T^{\pi}_{t,a} > m\right\} \cap E^c\right) \leq t\delta + e^{-\frac{m(\Delta^{\nu}_a)^2}{8}}.$$

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

$$\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right) \leq t\mathbb{P}^{\nu,\pi}\left(T^{\pi}_{t,a} > m\right) + m \leq te^{-\frac{m(\Delta^{\nu}_a)^2}{8}} + m + 1.$$

Since $m \geq 8\log(1/\delta)/(\Delta^{\nu}_a)^2$ implies $-m(\Delta^{\nu}_a)^2/8 \leq \log\delta$,

$$\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right) \leq t\delta + m + 1 = \frac{1}{t} + m + 1 \leq 2 + m \leq 3 + \frac{8\log(1/\delta)}{(\Delta^{\nu}_a)^2} = 3 + \frac{16\log(t)}{(\Delta^{\nu}_a)^2}.$$

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

$$R^{\nu,\pi}_t = \sum_{a \mid \Delta^{\nu}_a > 0} \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right) \leq \sum_{a \mid \Delta^{\nu}_a > 0} \Delta^{\nu}_a\left(3 + \frac{16\log(t)}{(\Delta^{\nu}_a)^2}\right) = \left(3\sum_{a=1}^{n}\Delta^{\nu}_a\right) + \sum_{a \mid \Delta^{\nu}_a > 0}\frac{16\log(t)}{\Delta^{\nu}_a}.$$

□

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

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

Proof. If $t \leq n$, then $T^{\pi}_{t,a} \leq 1$ for every $a \in \mathcal{A}$, so that $R^{\nu,\pi}_t = \sum_a \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right) \leq \sum_a \Delta^{\nu}_a$.

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

$$R^{\nu,\pi}_t = \left(\sum_{a \mid \Delta^{\nu}_a < \Delta} \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right)\right) + \left(\sum_{a \mid \Delta^{\nu}_a \geq \Delta} \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right)\right) \leq t\Delta + \sum_{a \mid \Delta^{\nu}_a \geq \Delta} \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right).$$

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

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

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

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

□

9 Relative entropy

Consider probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $(\Omega, \mathcal{F})$.

Proposition 9.1. If λ_1 and λ_2 are σ-finite measures on $(\Omega, \mathcal{F})$, then $\lambda = \lambda_1 + \lambda_2$ is a σ-finite measure on $(\Omega, \mathcal{F})$.

Proof. Clearly, $\lambda(\emptyset) = \lambda_1(\emptyset) + \lambda_2(\emptyset) = 0$. For any sequence $(F_n \in \mathcal{F} \mid n \in \mathbb{N})$ such that $F_n \cap F_m = \emptyset$ for $n \neq m$,

$$\lambda\left(\bigcup_n F_n\right) = \lambda_1\left(\bigcup_n F_n\right) + \lambda_2\left(\bigcup_n F_n\right) = \sum_n \lambda_1(F_n) + \lambda_2(F_n) = \sum_n \lambda(F_n).$$

Consider a sequence $(F_n^1 \in \mathcal{F} \mid n \in \mathbb{N})$ such that $\bigcup_n F_n^1 = \Omega$ and $\lambda_1(F_n^1) < \infty$ for every $n \in \mathbb{N}$. Analogously, consider a sequence $(F_n^2 \in \mathcal{F} \mid n \in \mathbb{N})$ such that $\bigcup_n F_n^2 = \Omega$ and $\lambda_2(F_n^2) < \infty$ for every $n \in \mathbb{N}$.

Let $F_{i,j} = F_i^1 \cap F_j^2$, so that $\bigcup_{i,j} F_{i,j} = \Omega$ and $\lambda(F_{i,j}) = \lambda_1(F_i^1 \cap F_j^2) + \lambda_2(F_i^1 \cap F_j^2) \leq \lambda_1(F_i^1) + \lambda_2(F_j^2) < \infty$. Because the set $\{F_{i,j} \mid i \in \mathbb{N} \text{ and } j \in \mathbb{N}\}$ is countable, λ is a σ-finite measure on $(\Omega, \mathcal{F})$. □

Proposition 9.2. There is a σ-finite measure λ on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$.

Proof. Let $\lambda : \mathcal{F} \to [0, \infty]$ be given by $\lambda(F) = \mathbb{P}(F) + \mathbb{Q}(F)$. Because $\mathbb{P}$ and $\mathbb{Q}$ are σ-finite measures on $(\Omega, \mathcal{F})$, λ is a σ-finite measure on $(\Omega, \mathcal{F})$. If $\lambda(F) = 0$, then $\mathbb{P}(F) = 0$ and $\mathbb{Q}(F) = 0$. Therefore, $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$. □

Proposition 9.3. For every σ-finite measure λ on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$, there is an $\mathcal{F}$-measurable function $p : \Omega \to [0, \infty)$ such that $p = d\mathbb{P}/d\lambda$ almost everywhere and an $\mathcal{F}$-measurable function $q : \Omega \to [0, \infty)$ such that $q = d\mathbb{Q}/d\lambda$ almost everywhere.

Proof. This is a direct consequence of the Radon-Nikodym theorem. □

Definition 9.1. Consider an $\mathcal{F}$-measurable function $p : \Omega \to [0, \infty)$ and an $\mathcal{F}$-measurable function $q : \Omega \to [0, \infty)$. The $\mathcal{F}$-measurable function $p \log(p/q) : \Omega \to \mathbb{R}$ is defined by

$$\left(p \log\left(\frac{p}{q}\right)\right)(\omega) = \begin{cases} p(\omega)\log(p(\omega)/q(\omega)), & \text{if } p(\omega)q(\omega) > 0, \\ 0, & \text{if } p(\omega)q(\omega) = 0. \end{cases}$$

Definition 9.2. Consider a σ-finite measure λ on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$. Let $p = d\mathbb{P}/d\lambda$ almost everywhere and $q = d\mathbb{Q}/d\lambda$ almost everywhere. The relative entropy $D(\mathbb{P}, \mathbb{Q})$ between $\mathbb{P}$ and $\mathbb{Q}$ under λ is given by

$$D(\mathbb{P}, \mathbb{Q}) = \int_\Omega p \log\left(\frac{p}{q}\right) \, d\lambda$$

whenever $p \log(p/q)$ is λ-integrable and $\mathbb{P}(q = 0) = 0$. Otherwise, $D(\mathbb{P}, \mathbb{Q}) = \infty$.

The relative entropy is also called Kullback-Leibler divergence.

Proposition 9.4. If λ_1 is a σ-finite measure on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda_1$ and $\mathbb{Q} \ll \lambda_1$ and λ_2 is a σ-finite measure on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda_2$ and $\mathbb{Q} \ll \lambda_2$, then the relative entropy $D(\mathbb{P}, \mathbb{Q})$ between $\mathbb{P}$ and $\mathbb{Q}$ under λ_1 is equal to the relative entropy $D(\mathbb{P}, \mathbb{Q})$ between $\mathbb{P}$ and $\mathbb{Q}$ under λ_2.

Proof. Let $p_1 = d\mathbb{P}/d\lambda_1$ almost everywhere, $q_1 = d\mathbb{Q}/d\lambda_1$ almost everywhere, $p_2 = d\mathbb{P}/d\lambda_2$ almost everywhere, and $q_2 = d\mathbb{Q}/d\lambda_2$ almost everywhere. Recall that $\lambda = \lambda_1 + \lambda_2$ is a σ-finite measure on $(\Omega, \mathcal{F})$. Since $\lambda_1 \ll \lambda$ and $\lambda_2 \ll \lambda$, let $l_1 = d\lambda_1/d\lambda$ almost everywhere and $l_2 = d\lambda_2/d\lambda$ almost everywhere. Since $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$, let $p = d\mathbb{P}/d\lambda$ almost everywhere and $q = d\mathbb{Q}/d\lambda$ almost everywhere. By the Radon-Nikodym chain rule, $p = p_1 l_1 = p_2 l_2$ almost everywhere and $q = q_1 l_1 = q_2 l_2$ almost everywhere.

We will first show that $p_1 \log(p_1/q_1)$ is λ_1-integrable if and only if $p_2 \log(p_2/q_2)$ is λ_2-integrable.

If $p_1 \log(p_1/q_1)$ is λ_1-integrable or $p \log(p/q)$ is λ-integrable,

$$\int_\Omega p_1 \log\left(\frac{p_1}{q_1}\right) \, d\lambda_1 = \int_\Omega l_1 \left(p_1 \log\left(\frac{p_1}{q_1}\right)\right) \, d\lambda = \int_\Omega p_1 l_1 \log\left(\frac{p_1 l_1}{q_1 l_1}\right) \, d\lambda = \int_\Omega p \log\left(\frac{p}{q}\right) \, d\lambda < \infty.$$

If $p_2 \log(p_2/q_2)$ is λ_2-integrable or $p \log(p/q)$ is λ-integrable,

$$\int_\Omega p_2 \log\left(\frac{p_2}{q_2}\right) \, d\lambda_2 = \int_\Omega l_2 \left(p_2 \log\left(\frac{p_2}{q_2}\right)\right) \, d\lambda = \int_\Omega p_2 l_2 \log\left(\frac{p_2 l_2}{q_2 l_2}\right) \, d\lambda = \int_\Omega p \log\left(\frac{p}{q}\right) \, d\lambda < \infty.$$

Therefore, $p_1 \log(p_1/q_1)$ is λ_1-integrable if and only if $p_2 \log(p_2/q_2)$ is λ_2-integrable, In that case,

$$\int_\Omega p_1 \log\left(\frac{p_1}{q_1}\right)\, d\lambda_1 = \int_\Omega p \log\left(\frac{p}{q}\right)\, d\lambda = \int_\Omega p_2 \log\left(\frac{p_2}{q_2}\right)\, d\lambda_2.$$

It remains to show that $\mathbb{P}(q_1 = 0) = 0$ if and only if $\mathbb{P}(q_2 = 0) = 0$, which follows from the fact that

$$\mathbb{P}(q = 0) = \int_{\{q_1 l_1 = 0\}} p_1 l_1 \, d\lambda = \int_{\{q_1 l_1 = 0, p_1 l_1 > 0\}} p_1 l_1 \, d\lambda = \int_{\{q_1 = 0, p_1 l_1 > 0\}} p_1 l_1 \, d\lambda = \int_{\{q_1 = 0\}} p_1 \, d\lambda_1 = \mathbb{P}(q_1 = 0),$$
$$\mathbb{P}(q = 0) = \int_{\{q_2 l_2 = 0\}} p_2 l_2 \, d\lambda = \int_{\{q_2 l_2 = 0, p_2 l_2 > 0\}} p_2 l_2 \, d\lambda = \int_{\{q_2 = 0, p_2 l_2 > 0\}} p_2 l_2 \, d\lambda = \int_{\{q_2 = 0\}} p_2 \, d\lambda_2 = \mathbb{P}(q_2 = 0).$$

□

Proposition 9.5. Consider a σ-finite measure λ on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$. Let $p = d\mathbb{P}/d\lambda$ almost everywhere and $q = d\mathbb{Q}/d\lambda$ almost everywhere. If $D(\mathbb{P}, \mathbb{Q}) < \infty$, then $\lambda(p > 0, q = 0) = 0$.

Proof. If $D(\mathbb{P}, \mathbb{Q}) < \infty$, then $\mathbb{P}(q = 0) = 0$. Since $p = d\mathbb{P}/d\lambda$ almost everywhere,

$$0 = \mathbb{P}(q = 0) = \int_{\{q=0\}} p \, d\lambda = \int_\Omega \mathbb{I}_{\{p>0, q=0\}} p \, d\lambda,$$

so that $\lambda(\mathbb{I}_{\{p>0,q=0\}} p > 0) = 0$. Since $\{\mathbb{I}_{\{p>0,q=0\}} p > 0\} = \{p > 0, q = 0\}$, we have $\lambda(p > 0, q = 0) = 0$. □

Proposition 9.6. Consider a σ-finite measure λ on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$. Let $p = d\mathbb{P}/d\lambda$ almost everywhere and $q = d\mathbb{Q}/d\lambda$ almost everywhere. If $D(\mathbb{P}, \mathbb{Q}) < \infty$, then $\int_\Omega pq \, d\lambda > 0$ and $\int_{\{pq>0\}} q \, d\lambda > 0$.

Proof. If $D(\mathbb{P}, \mathbb{Q}) < \infty$, then $\mathbb{P}(q = 0) = \int_{\{q=0\}} p \, d\lambda = 0$. Therefore,

$$1 = \mathbb{P}(\Omega) = \int_\Omega p \, d\lambda = \int_{\{q=0\}} p \, d\lambda + \int_{\{q>0\}} p \, d\lambda = \int_{\{pq>0\}} p \, d\lambda,$$

so that $\lambda(pq > 0) > 0$. Consequently, $\int_\Omega pq \, d\lambda > 0$ and $\int_{\{pq>0\}} q \, d\lambda > 0$. □

Proposition 9.7. The relative entropy $D(\mathbb{P}, \mathbb{Q})$ between $\mathbb{P}$ and $\mathbb{Q}$ is non-negative.

Proof. Consider a σ-finite measure λ on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$. Let $p = d\mathbb{P}/d\lambda$ almost everywhere and $q = d\mathbb{Q}/d\lambda$ almost everywhere. It is sufficient to show that the relative entropy $D(\mathbb{P}, \mathbb{Q})$ between $\mathbb{P}$ and $\mathbb{Q}$ under λ is non-negative when $D(\mathbb{P}, \mathbb{Q}) < \infty$. In that case, because $p = d\mathbb{P}/d\lambda$ almost everywhere,

$$D(\mathbb{P}, \mathbb{Q}) = \int_\Omega p \log\left(\frac{p}{q}\right)\, d\lambda = \int_{\{pq>0\}} p \log\left(\frac{p}{q}\right)\, d\lambda = \int_{\{pq>0\}} -\log\left(\frac{q}{p}\right)\, d\mathbb{P}.$$

Consider the measure space $(A, \mathcal{F}_A, \mathbb{P}_A)$ restricted to $A = \{pq > 0\}$ and recall that

$$D(\mathbb{P}, \mathbb{Q}) = \int_{\{pq>0\}} -\log\left(\frac{q}{p}\right)\, d\mathbb{P} = \int_A -\log\left(\frac{q_{|A}}{p_{|A}}\right)\, d\mathbb{P}_A.$$

Note that the restricted function $q_{|A}/p_{|A} : A \to (0, \infty)$ is $\mathbb{P}_A$-integrable, since

$$\int_A \frac{q_{|A}}{p_{|A}}\, d\mathbb{P}_A = \int_{\{pq>0\}} \frac{q}{p}\, d\mathbb{P} = \int_{\{pq>0\}} p\frac{q}{p}\, d\lambda = \int_{\{pq>0\}} q \, d\lambda \le \int_\Omega q \, d\lambda = \mathbb{Q}(\Omega) = 1.$$

By Jensen's inequality, because the function $\phi : (0, \infty) \to \mathbb{R}$ given by $\phi(x) = -\log(x)$ is convex,

$$D(\mathbb{P}, \mathbb{Q}) \ge -\log\left(\int_A \frac{q_{|A}}{p_{|A}}\, d\mathbb{P}_A\right) \ge -\log(1) = 0.$$

□

Theorem 9.1 (Bretagnolle-Huber inequality)**.** If $F \in \mathcal{F}$, then $\mathbb{P}(F) + \mathbb{Q}(F^c) \ge e^{-D(\mathbb{P},\mathbb{Q})}/2$.

Proof. It is sufficient to show that if $F \in \mathcal{F}$, then $\mathbb{P}(F) + \mathbb{Q}(F^c) \geq e^{-D(\mathbb{P},\mathbb{Q})}/2$ when $D(\mathbb{P}, \mathbb{Q}) < \infty$.

Consider a σ-finite measure λ on $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \lambda$ and $\mathbb{Q} \ll \lambda$. Let $p = d\mathbb{P}/d\lambda$ almost everywhere and $q = d\mathbb{Q}/d\lambda$ almost everywhere. Since $p + q = \min(p, q) + \max(p, q)$,

$$1 = \frac{1}{2}\left(\mathbb{P}(\Omega) + \mathbb{Q}(\Omega)\right) = \frac{1}{2}\int_\Omega (p+q)\ d\lambda = \frac{1}{2}\int_\Omega (\min(p,q) + \max(p,q))\ d\lambda \geq \frac{1}{2}\int_\Omega \max(p,q)\ d\lambda.$$

Since $\min(p, q)\max(p, q) = pq$ and $\min(p, q)$ and $\max(p, q)$ are λ-integrable, by the Schwarz inequality,

$$\left(\int_\Omega \sqrt{pq}\ d\lambda\right)^2 = \left(\int_\Omega \sqrt{\min(p,q)}\sqrt{\max(p,q)}\ d\lambda\right)^2 \leq \left(\int_\Omega \min(p,q)\ d\lambda\right)\left(\int_\Omega \max(p,q)\ d\lambda\right).$$

Considering a previous inequality,

$$\frac{1}{2}\left(\int_\Omega \sqrt{pq}\ d\lambda\right)^2 \leq \frac{1}{2}\left(\int_\Omega \min(p,q)\ d\lambda\right)\left(\int_\Omega \max(p,q)\ d\lambda\right) \leq \int_\Omega \min(p,q)\ d\lambda.$$

Note that, for every $F \in \mathcal{F}$,

$$\mathbb{P}(F) + \mathbb{Q}(F^c) = \int_F p\ d\lambda + \int_{F^c} q\ d\lambda \geq \int_F \min(p,q)\ d\lambda + \int_{F^c} \min(p,q)\ d\lambda = \int_\Omega \min(p,q)\ d\lambda.$$

Considering a previous inequality, for every $F \in \mathcal{F}$,

$$\mathbb{P}(F) + \mathbb{Q}(F^c) \geq \frac{1}{2}\left(\int_\Omega \sqrt{pq}\ d\lambda\right)^2.$$

Note that $\int_\Omega pq\ d\lambda > 0$ implies $\int_\Omega \sqrt{pq}\ d\lambda > 0$. Since $x^2 = e^{2\log(x)}$ for every $x \in (0, \infty)$,

$$\mathbb{P}(F) + \mathbb{Q}(F^c) \geq \frac{1}{2}e^{2\log\left(\int_\Omega \sqrt{pq}\ d\lambda\right)}.$$

Consider the measure space $(A, \mathcal{F}_A, \mathbb{P}_A)$ restricted to $A = \{pq > 0\}$.
Note that the restricted function $\sqrt{q_{|A}/p_{|A}} : A \to (0, \infty)$ is $\mathbb{P}_A$-integrable, since

$$\int_A \sqrt{\frac{q_{|A}}{p_{|A}}}\ d\mathbb{P}_A = \int_{\{pq>0\}} \sqrt{\frac{q}{p}}\ d\mathbb{P} = \int_{\{pq>0\}} p\sqrt{\frac{q}{p}}\ d\lambda = \int_{\{pq>0\}} \sqrt{pq}\ d\lambda \leq \int_\Omega \sqrt{pq}\ d\lambda.$$

By Jensen's inequality, because the function $\phi : (0, \infty) \to \mathbb{R}$ given by $\phi(x) = -\log(x)$ is convex,

$$-\log\left(\int_\Omega \sqrt{pq}\ d\lambda\right) = -\log\left(\int_{\{pq>0\}} \sqrt{pq}\ d\lambda\right) = -\log\left(\int_A \sqrt{\frac{q_{|A}}{p_{|A}}}\ d\mathbb{P}_A\right) \leq \int_A -\log\sqrt{\frac{q_{|A}}{p_{|A}}}\ d\mathbb{P}_A.$$

Therefore,

$$\log\left(\int_\Omega \sqrt{pq}\ d\lambda\right) \geq \int_{\{pq>0\}} \log\sqrt{\frac{q}{p}}\ d\mathbb{P} = -\frac{1}{2}\int_{\{pq>0\}} p\log\left(\frac{p}{q}\right)\ d\lambda = -\frac{1}{2}D(\mathbb{P}, \mathbb{Q}).$$

Considering a previous inequality,

$$\mathbb{P}(F) + \mathbb{Q}(F^c) \geq \frac{1}{2}e^{2\log\left(\int_\Omega \sqrt{pq}\ d\lambda\right)} \geq \frac{1}{2}e^{-D(\mathbb{P},\mathbb{Q})}.$$

□

10 Divergence decomposition

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 ν under the policy π.

Definition 10.1. For every $t \in \mathbb{N}^+$, the joint law $\mathcal{L}^{\nu,\pi}_{1:t} : \mathcal{B}(\mathbb{R}^t) \to [0,1]$ is the measure on $(\mathbb{R}^t, \mathcal{B}(\mathbb{R}^t))$ given by

$$\mathcal{L}^{\nu,\pi}_{1:t}(\Gamma) = \mathbb{P}^{\nu,\pi}\left((X_1, \dots, X_t) \in \Gamma\right).$$

Proposition 10.1. There is a σ-finite measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $P_a \ll \lambda$ for every $a \in \mathcal{A}$.

Proof. Let $\lambda : \mathcal{B}(\mathbb{R}) \to [0, \infty]$ be given by $\lambda(B) = \sum_a P_a(B)$. Because P_a is a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for every $a \in \mathcal{A}$, λ is a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If $\lambda(B) = 0$, then $P_a(B) = 0$ for every $a \in \mathcal{A}$, so that $P_a \ll \lambda$. □

Proposition 10.2. Consider a σ-finite measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $P_a \ll \lambda$ for every $a \in \mathcal{A}$. Let $p_a = dP_a/d\lambda$ almost everywhere for every $a \in \mathcal{A}$. For every $t \in \mathbb{N}^+$, consider the function $p^{\nu,\pi}_{1:t} : \mathbb{R}^t \to [0, \infty)$ given by

$$p^{\nu,\pi}_{1:t}(x_1, \dots, x_t) = \prod_{k=1}^{t} p_{\pi_k(x_0, \dots, x_{k-1})}(x_k),$$

where $x_0 = 0$. If λ^t is the product measure $\lambda \times \cdots \times \lambda$ on $(\mathbb{R}^t, \mathcal{B}(\mathbb{R}^t))$, then $p^{\nu,\pi}_{1:t} = d\mathcal{L}^{\nu,\pi}_{1:t}/d\lambda^t$ almost everywhere.

Proof. Consider the case where $t = 1$. For every $B \in \mathcal{B}(\mathbb{R})$, since $\pi_1(X_0) = \pi_1(0)$,

$$\mathcal{L}^{\nu,\pi}_{1:1}(B) = \mathbb{P}^{\nu,\pi}(X_1 \in B) = \mathbb{E}^{\nu,\pi}\left(P_{\pi_1(X_0)}(B)\right) = P_{\pi_1(0)}(B) = \int_B p_{\pi_1(0)} \, d\lambda = \int_B p^{\nu,\pi}_{1:1} \, d\lambda^1.$$

In order to employ induction, suppose there is a $t-1 \in \mathbb{N}^+$ such that $p^{\nu,\pi}_{1:t-1} = d\mathcal{L}^{\nu,\pi}_{1:t-1}/d\lambda^{t-1}$ almost everywhere. Since $p^{\nu,\pi}_{1:t} : \mathbb{R}^t \to [0, \infty)$ is $\mathcal{B}(\mathbb{R}^t)$-measurable, consider the measure $\mathcal{L}_{1:t} : \mathcal{B}(\mathbb{R}^t) \to [0, \infty]$ given by

$$\mathcal{L}_{1:t}(\Gamma) = \int_\Gamma p^{\nu,\pi}_{1:t} \, d\lambda^t.$$

Recall that $\mathcal{I}_t = \{B_1 \times \cdots \times B_t \mid B_k \in \mathcal{B}(\mathbb{R}) \text{ for every } k \in \{1, \dots, t\}\}$ is a π-system on $\mathbb{R}^t$ such that $\sigma(\mathcal{I}_t) = \mathcal{B}(\mathbb{R}^t)$. Therefore, if we show that $\mathcal{L}_{1:t}(I_t) = \mathcal{L}^{\nu,\pi}_{1:t}(I_t)$ for every $I_t \in \mathcal{I}_t$, then $\mathcal{L}_{1:t} = \mathcal{L}^{\nu,\pi}_{1:t}$, so that the proof will be complete.

Consider a set $I_t \in \mathcal{I}_t$ given by $I_t = B_1 \times \cdots \times B_t$. Because $\mathcal{L}^{\nu,\pi}_{1:t}$ is the joint law of $X_1, \dots, X_t$,

$$\mathcal{L}^{\nu,\pi}_{1:t}(I_t) = \mathbb{P}^{\nu,\pi}(X_1 \in B_1, \dots, X_t \in B_t) = \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_1 \in B_1, \dots, X_{t-1} \in B_{t-1}\}} \mathbb{I}_{\{X_t \in B_t\}}\right).$$

Let $A_t = \pi_t(X_0, \dots, X_{t-1})$. By taking out what is known,

$$\mathcal{L}^{\nu,\pi}_{1:t}(I_t) = \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_1 \in B_1, \dots, X_{t-1} \in B_{t-1}\}} \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_t \in B_t\}} \mid X_0, \dots, X_{t-1}\right)\right) = \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_1 \in B_1, \dots, X_{t-1} \in B_{t-1}\}} P_{A_t}(B_t)\right).$$

Because $\mathcal{L}^{\nu,\pi}_{1:t-1}$ is the joint law of $X_1, \dots, X_{t-1}$,

$$\mathcal{L}^{\nu,\pi}_{1:t}(I_t) = \int_{\mathbb{R}^{t-1}} \mathbb{I}_{B_1 \times \cdots \times B_{t-1}}(x_{1:t-1}) P_{\pi_t(0, x_{1:t-1})}(B_t) \; \mathcal{L}^{\nu,\pi}_{1:t-1}(dx_{1:t-1}).$$

By the inductive hypothesis and since $p_{\pi_t(0,x_{1:t-1})} = dP_{\pi_t(0,x_{1:t-1})}/d\lambda$ almost everywhere for every $x_{1:t-1} \in \mathbb{R}^{t-1}$,

$$\mathcal{L}^{\nu,\pi}_{1:t}(I_t) = \int_{\mathbb{R}^{t-1}} \mathbb{I}_{B_1 \times \cdots \times B_{t-1}}(x_{1:t-1}) p^{\nu,\pi}_{1:t-1}(x_{1:t-1}) \left(\int_{\mathbb{R}} \mathbb{I}_{B_t}(x_t) p_{\pi_t(0,x_{1:t-1})}(x_t) \; \lambda(dx_t) \right) \; \lambda^{t-1}(dx_{1:t-1}).$$

Since $p^{\nu,\pi}_{1:t}(x_{1:t}) = p^{\nu,\pi}_{1:t-1}(x_{1:t-1}) p_{\pi_t(0,x_{1:t-1})}(x_t)$ for every $x_{1:t} \in \mathbb{R}^t$ and Fubini's theorem,

$$\mathcal{L}^{\nu,\pi}_{1:t}(I_t) = \int_{\mathbb{R}^{t-1}} \int_{\mathbb{R}} \mathbb{I}_{B_1 \times \cdots \times B_t}(x_{1:t}) p^{\nu,\pi}_{1:t}(x_{1:t}) \; \lambda(dx_t) \; \lambda^{t-1}(dx_{1:t-1}) = \int_{I_t} p^{\nu,\pi}_{1:t} \; \lambda^t = \mathcal{L}_{1:t}(I_t).$$

□

Theorem 10.1. If $\nu' = (P'_a \mid a \in \mathcal{A})$ is a stochastic bandit such that $D(P_a, P'_a) < \infty$ for every $a \in \mathcal{A}$ and $t \in \mathbb{N}^+$,

$$D(\mathcal{L}^{\nu,\pi}_{1:t}, \mathcal{L}^{\nu',\pi}_{1:t}) = \sum_a D(P_a, P'_a) \mathbb{E}^{\nu,\pi}\left(T^\pi_{t,a}\right).$$

Proof. Consider the σ-finite measure $\lambda : \mathcal{B}(\mathbb{R}) \to [0, \infty]$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by $\lambda(B) = \sum_a P_a(B) + P'_a(B)$. Note that $P_a \ll \lambda$ and $P'_a \ll \lambda$ for every $a \in \mathcal{A}$. Let $p_a = dP_a/d\lambda$ almost everywhere and $p'_a = dP'_a/d\lambda$ almost everywhere for every $a \in \mathcal{A}$. For every $t \in \mathbb{N}^+$, consider the functions $p^{\nu,\pi}_{1:t} : \mathbb{R}^t \to [0, \infty)$ and $p^{\nu',\pi}_{1:t} : \mathbb{R}^t \to [0, \infty)$ given by

$$p^{\nu,\pi}_{1:t}(x_1, \dots, x_t) = \prod_{k=1}^{t} p_{\pi_k(x_0,\dots,x_{k-1})}(x_k),$$

$$p^{\nu',\pi}_{1:t}(x_1, \dots, x_t) = \prod_{k=1}^{t} p'_{\pi_k(x_0,\dots,x_{k-1})}(x_k),$$

where $x_0 = 0$. Recall that $p^{\nu,\pi}_{1:t} = d\mathcal{L}^{\nu,\pi}_{1:t}/d\lambda^t$ almost everywhere and $p^{\nu',\pi}_{1:t} = d\mathcal{L}^{\nu',\pi}_{1:t}/d\lambda^t$ almost everywhere, where λ^t is the product measure $\lambda \times \dots \times \lambda$ on $(\mathbb{R}^t, \mathcal{B}(\mathbb{R}^t))$. Furthermore, recall that $\mathcal{L}^{\nu,\pi}_{1:t} \ll \lambda^t$ and $\mathcal{L}^{\nu',\pi}_{1:t} \ll \lambda^t$.

For every $k \in \mathbb{N}^+$, let $A_k = \pi_k(X_0, \dots, X_{k-1})$. For every $t \in \mathbb{N}^+$, let D_t be given by

$$D_t = \sum_a D(P_a, P'_a)\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right) = \sum_{k=1}^{t} \mathbb{E}^{\nu,\pi}\left(\sum_a \mathbb{I}_{\{A_k=a\}} D(P_a, P'_a)\right) = \sum_{k=1}^{t} \mathbb{E}^{\nu,\pi}\left(D(P_{A_k}, P'_{A_k})\right) < \infty.$$

Consider the case where $t = 1$. Since $P_a(p'_a = 0) = 0$ for every $a \in \mathcal{A}$,

$$\mathcal{L}^{\nu,\pi}_{1:1}(p^{\nu',\pi}_{1:1} = 0) = \mathcal{L}^{\nu,\pi}_{1:1}(p'_{\pi_1(0)} = 0) = P_{\pi_1(0)}(p'_{\pi_1(0)} = 0) = 0.$$

Since $A_1 = \pi_1(X_0) = \pi_1(0)$,

$$D_1 = \mathbb{E}^{\nu,\pi}\left(D(P_{A_1}, P'_{A_1})\right) = D\left(P_{\pi_1(0)}, P'_{\pi_1(0)}\right) = \int_{\mathbb{R}} p_{\pi_1(0)} \log\left(\frac{p_{\pi_1(0)}}{p'_{\pi_1(0)}}\right) d\lambda = \int_{\mathbb{R}} p^{\nu,\pi}_{1:1} \log\left(\frac{p^{\nu,\pi}_{1:1}}{p^{\nu',\pi}_{1:1}}\right) d\lambda^1,$$

so that $p^{\nu,\pi}_{1:1} \log\left(p^{\nu,\pi}_{1:1}/p^{\nu',\pi}_{1:1}\right)$ is λ^1-integrable and $D_1 = D(\mathcal{L}^{\nu,\pi}_{1:1}, \mathcal{L}^{\nu',\pi}_{1:1})$.

In order to employ induction, suppose that $D_{t-1} = D(\mathcal{L}^{\nu,\pi}_{1:t-1}, \mathcal{L}^{\nu',\pi}_{1:t-1})$ for some $t - 1 \in \mathbb{N}^+$.

For every $x_{1:t} \in \mathbb{R}^t$, if $p^{\nu,\pi}_{1:t}(x_{1:t}) > 0$ and $p^{\nu',\pi}_{1:t}(x_{1:t}) = 0$, then $p^{\nu,\pi}_{1:t-1}(x_{1:t-1}) > 0$ and there is an action $a_t \in \mathcal{A}$ such that $p_{a_t}(x_t) > 0$. Furthermore, $p^{\nu',\pi}_{1:t-1}(x_{1:t-1}) = 0$ or $p^{\nu',\pi}_{1:t-1}(x_{1:t-1}) > 0$ and $p'_{a_t}(x_t) = 0$. Therefore,

$$\left\{p^{\nu,\pi}_{1:t} > 0, p^{\nu',\pi}_{1:t} = 0\right\} \subseteq \left(\left\{p^{\nu,\pi}_{1:t-1} > 0, p^{\nu',\pi}_{1:t-1} = 0\right\} \times \mathbb{R}\right) \cup \left(\bigcup_{a_t} \left\{p^{\nu,\pi}_{1:t-1} > 0, p^{\nu',\pi}_{1:t-1} > 0\right\} \times \left\{p_{a_t} > 0, p'_{a_t} = 0\right\}\right).$$

Let $l_t = \lambda^t\left(p^{\nu,\pi}_{1:t} > 0, p^{\nu',\pi}_{1:t} = 0\right)$. By an union bound,

$$l_t \le \lambda^t\left(\left\{p^{\nu,\pi}_{1:t-1} > 0, p^{\nu',\pi}_{1:t-1} = 0\right\} \times \mathbb{R}\right) + \sum_{a_t} \lambda^t\left(\left\{p^{\nu,\pi}_{1:t-1} > 0, p^{\nu',\pi}_{1:t-1} > 0\right\} \times \left\{p_{a_t} > 0, p'_{a_t} = 0\right\}\right).$$

Since λ^t is the product measure $\lambda \times \dots \times \lambda$ on $(\mathbb{R}^t, \mathcal{B}(\mathbb{R}^t))$,

$$l_t \le \lambda^{t-1}\left(p^{\nu,\pi}_{1:t-1} > 0, p^{\nu',\pi}_{1:t-1} = 0\right)\lambda(\mathbb{R}) + \sum_{a_t} \lambda^{t-1}\left(p^{\nu,\pi}_{1:t-1} > 0, p^{\nu',\pi}_{1:t-1} > 0\right)\lambda\left(p_{a_t} > 0, p'_{a_t} = 0\right).$$

Since $D_{t-1} = D(\mathcal{L}^{\nu,\pi}_{1:t-1}, \mathcal{L}^{\nu',\pi}_{1:t-1}) < \infty$ by the inductive hypothesis, note that $\lambda^{t-1}\left(p^{\nu,\pi}_{1:t-1} > 0, p^{\nu',\pi}_{1:t-1} = 0\right) = 0$. Since $D(P_{a_t}, P'_{a_t}) < \infty$, recall that $\lambda\left(p_{a_t} > 0, p'_{a_t} = 0\right) = 0$. Therefore, $\lambda^t\left(p^{\nu,\pi}_{1:t} > 0, p^{\nu',\pi}_{1:t} = 0\right) = l_t = 0$.

Since $\mathcal{L}^{\nu,\pi}_{1:t} \ll \lambda^t$, note that $\mathcal{L}^{\nu,\pi}_{1:t}(p^{\nu,\pi}_{1:t} > 0, p^{\nu',\pi}_{1:t} = 0) = 0$. Therefore, completing this step,

$$0 = \mathcal{L}^{\nu,\pi}_{1:t}(p^{\nu,\pi}_{1:t} > 0, p^{\nu',\pi}_{1:t} = 0) = \int_{\{p^{\nu,\pi}_{1:t} > 0, p^{\nu',\pi}_{1:t} = 0\}} p^{\nu,\pi}_{1:t} \, d\lambda^t = \int_{\{p^{\nu',\pi}_{1:t} = 0\}} p^{\nu,\pi}_{1:t} \, d\lambda^t = \mathcal{L}^{\nu,\pi}_{1:t}(p^{\nu',\pi}_{1:t} = 0).$$

It remains to show that $p^{\nu,\pi}_{1:t} \log\left(p^{\nu,\pi}_{1:t}/p^{\nu',\pi}_{1:t}\right)$ is λ^t-integrable and that

$$D_t = \int_{\mathbb{R}^t} p^{\nu,\pi}_{1:t} \log\left(\frac{p^{\nu,\pi}_{1:t}}{p^{\nu',\pi}_{1:t}}\right) d\lambda^t.$$

Since $\mathcal{L}^{\nu,\pi}_{1:t-1}$ is the joint law of $X_1, \dots, X_{t-1}$,

$$D_t = D_{t-1} + \mathbb{E}^{\nu,\pi}\left(D(P_{A_t}, P'_{A_t})\right) = D_{t-1} + \int_{\mathbb{R}^{t-1}} D(P_{\pi_t(0,x_{1:t-1})}, P'_{\pi_t(0,x_{1:t-1})})\ \mathcal{L}^{\nu,\pi}_{1:t-1}(dx_{1:t-1}).$$

Since $D(P_a, P'_a) < \infty$ for every $a \in \mathcal{A}$,

$$D_t = D_{t-1} + \int_{\mathbb{R}^{t-1}} \int_{\mathbb{R}} p_{\pi_t(0,x_{1:t-1})}(x_t) \log\left(\frac{p_{\pi_t(0,x_{1:t-1})}(x_t)}{p'_{\pi_t(0,x_{1:t-1})}(x_t)}\right)\ \lambda(dx_t)\ \mathcal{L}^{\nu,\pi}_{1:t-1}(dx_{1:t-1}).$$

Since $p^{\nu,\pi}_{1:t-1} = d\mathcal{L}^{\nu,\pi}_{1:t-1}/d\lambda^{t-1}$ almost everywhere and $p^{\nu,\pi}_{1:t}(x_{1:t}) = p^{\nu,\pi}_{1:t-1}(x_{1:t-1}) p_{\pi_t(0,x_{1:t-1})}(x_t)$,

$$D_t = D_{t-1} + \int_{\mathbb{R}^{t-1}} \int_{\mathbb{R}} p^{\nu,\pi}_{1:t}(x_{1:t}) \log\left(\frac{p_{\pi_t(0,x_{1:t-1})}(x_t)}{p'_{\pi_t(0,x_{1:t-1})}(x_t)}\right)\ \lambda(dx_t)\ \lambda^{t-1}(dx_{1:t-1}).$$

Since the function under consideration is λ^t-integrable, by Fubini's theorem,

$$D_t = D_{t-1} + \int_{\mathbb{R}^t} p^{\nu,\pi}_{1:t}(x_{1:t}) \log\left(\frac{p_{\pi_t(0,x_{1:t-1})}(x_t)}{p'_{\pi_t(0,x_{1:t-1})}(x_t)}\right)\ \lambda^t(dx_{1:t}).$$

Since $p^{\nu,\pi}_{1:t} = d\mathcal{L}^{\nu,\pi}_{1:t}/d\lambda^t$ almost everywhere and $\mathcal{L}^{\nu,\pi}_{1:t}$ is the joint law of $X_1, \dots, X_t$,

$$D_t = D_{t-1} + \int_{\mathbb{R}^t} \log\left(\frac{p_{\pi_t(0,x_{1:t-1})}(x_t)}{p'_{\pi_t(0,x_{1:t-1})}(x_t)}\right)\ \mathcal{L}^{\nu,\pi}_{1:t}(dx_{1:t}) = D_{t-1} + \mathbb{E}^{\nu,\pi}\left(\log\left(\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\right)\right).$$

By the inductive hypothesis, since $p^{\nu,\pi}_{1:t-1} = d\mathcal{L}^{\nu,\pi}_{1:t-1}/d\lambda^{t-1}$ almost everywhere,

$$D_{t-1} = \int_{\mathbb{R}^{t-1}} \log\left(\frac{p^{\nu,\pi}_{1:t-1}(x_{1:t-1})}{p^{\nu',\pi}_{1:t-1}(x_{1:t-1})}\right)\ \mathcal{L}^{\nu,\pi}_{1:t-1}(dx_{1:t-1}) = \mathbb{E}^{\nu,\pi}\left(\log\left(\frac{p^{\nu,\pi}_{1:t-1}(X_1, \dots, X_{t-1})}{p^{\nu',\pi}_{1:t-1}(X_1, \dots, X_{t-1})}\right)\right).$$

By the definition of the functions $p^{\nu,\pi}_{1:t-1}$ and $p^{\nu',\pi}_{1:t-1}$,

$$D_{t-1} = \mathbb{E}^{\nu,\pi}\left(\log\left(\prod_{k=1}^{t-1} \frac{p_{A_k}(X_k)}{p'_{A_k}(X_k)}\right)\right) = \sum_{k=1}^{t-1} \mathbb{E}^{\nu,\pi}\left(\log\left(\frac{p_{A_k}(X_k)}{p'_{A_k}(X_k)}\right)\right).$$

By combining the equation above with a previous equation,

$$D_t = \sum_{k=1}^{t} \mathbb{E}^{\nu,\pi}\left(\log\left(\frac{p_{A_k}(X_k)}{p'_{A_k}(X_k)}\right)\right) = \mathbb{E}^{\nu,\pi}\left(\log\left(\prod_{k=1}^{t} \frac{p_{A_k}(X_k)}{p'_{A_k}(X_k)}\right)\right) = \mathbb{E}^{\nu,\pi}\left(\log\left(\frac{p^{\nu,\pi}_{1:t}(X_1, \dots, X_t)}{p^{\nu',\pi}_{1:t}(X_1, \dots, X_t)}\right)\right).$$

Because $\mathcal{L}^{\nu,\pi}_{1:t}$ is the joint law of $X_1, \dots, X_t$ and $p^{\nu,\pi}_{1:t} = d\mathcal{L}^{\nu,\pi}_{1:t}/d\lambda^t$ almost everywhere,

$$D_t = \int_{\mathbb{R}^t} \log\left(\frac{p^{\nu,\pi}_{1:t}(x_{1:t})}{p^{\nu',\pi}_{1:t}(x_{1:t})}\right)\ \mathcal{L}^{\nu,\pi}_{1:t}(dx_{1:t}) = \int_{\mathbb{R}^t} p^{\nu,\pi}_{1:t}(x_{1:t}) \log\left(\frac{p^{\nu,\pi}_{1:t}(x_{1:t})}{p^{\nu',\pi}_{1:t}(x_{1:t})}\right)\ \lambda^t(dx_{1:t}),$$

which implies that $p^{\nu,\pi}_{1:t} \log\left(p^{\nu,\pi}_{1:t}/p^{\nu',\pi}_{1:t}\right)$ is λ^t-integrable and that $D_t = D(\mathcal{L}^{\nu,\pi}_{1:t}, \mathcal{L}^{\nu',\pi}_{1:t})$. □

11 Relative lower 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 ν under the policy π.

Theorem 11.1. Suppose that $\Delta^{\nu}_{a'} > 0$ for some action $a' \in \mathcal{A}$ and consider a stochastic bandit $\nu' = (P'_a \mid a \in \mathcal{A})$ such that $P'_a = P_a$ for every $a \neq a'$. Furthermore, suppose that $\mu^{\nu'}_* = \mu^{\nu'}_{a'} > \mu^{\nu}_*$ and that $D(P_{a'}, P'_{a'}) \in (0, \infty)$. In that case, for every time step $t > 1$,

$$R^{\nu',\pi}_t \geq \frac{t}{4} \min(\Delta^{\nu}_{a'}, \mu^{\nu'}_* - \mu^{\nu}_*) e^{-D(P_{a'}, P'_{a'})\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a'}\right)} - R^{\nu,\pi}_t.$$

Proof. Consider an action $a' \in \mathcal{A}$ such that $\Delta^{\nu}_{a'} > 0$ and let $t > 1$. By Theorem 4.2 and Markov's inequality,

$$R^{\nu,\pi}_t = \sum_a \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right) \geq \Delta^{\nu}_{a'} \mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a'}\right) \geq \frac{t}{2} \Delta^{\nu}_{a'} \mathbb{P}^{\nu,\pi}\left(T^{\pi}_{t,a'} \geq \frac{t}{2}\right).$$

For every $a \neq a'$, note that $\Delta^{\nu'}_a = \mu^{\nu'}_* - \mu^{\nu'}_a = \mu^{\nu'}_* - \mu^{\nu}_a \geq \mu^{\nu'}_* - \mu^{\nu}_*$. Since $\Delta^{\nu'}_{a'} = \mu^{\nu'}_* - \mu^{\nu'}_{a'} = 0$, by Theorem 4.2,

$$R^{\nu',\pi}_t = \sum_{a \neq a'} \Delta^{\nu'}_a \mathbb{E}^{\nu',\pi}\left(T^{\pi}_{t,a}\right) \geq (\mu^{\nu'}_* - \mu^{\nu}_*)\left(t - \mathbb{E}^{\nu',\pi}\left(T^{\pi}_{t,a'}\right)\right) = (\mu^{\nu'}_* - \mu^{\nu}_*)\mathbb{E}^{\nu',\pi}\left(t - T^{\pi}_{t,a'}\right),$$

where we also used the fact that $t = \sum_a \mathbb{E}^{\nu',\pi}\left(T^{\pi}_{t,a}\right) = \mathbb{E}^{\nu',\pi}\left(T^{\pi}_{t,a'}\right) + \sum_{a \neq a'} \mathbb{E}^{\nu',\pi}\left(T^{\pi}_{t,a}\right)$.

By Markov's inequality and since $\mathbb{P}^{\nu',\pi}\left(T^{\pi}_{t,a'} \leq t/2\right) \geq \mathbb{P}^{\nu',\pi}\left(T^{\pi}_{t,a'} < t/2\right)$,

$$R^{\nu',\pi}_t \geq \frac{t}{2}(\mu^{\nu'}_* - \mu^{\nu}_*)\mathbb{P}^{\nu',\pi}\left(t - T^{\pi}_{t,a'} \geq \frac{t}{2}\right) = \frac{t}{2}(\mu^{\nu'}_* - \mu^{\nu}_*)\mathbb{P}^{\nu',\pi}\left(T^{\pi}_{t,a'} \leq \frac{t}{2}\right) \geq \frac{t}{2}(\mu^{\nu'}_* - \mu^{\nu}_*)\mathbb{P}^{\nu',\pi}\left(T^{\pi}_{t,a'} < \frac{t}{2}\right).$$

By combining the previous inequalities,

$$R^{\nu,\pi}_t + R^{\nu',\pi}_t \geq \frac{t}{2}\Delta^{\nu}_{a'}\mathbb{P}^{\nu,\pi}\left(T^{\pi}_{t,a'} \geq \frac{t}{2}\right) + \frac{t}{2}(\mu^{\nu'}_* - \mu^{\nu}_*)\mathbb{P}^{\nu',\pi}\left(T^{\pi}_{t,a'} < \frac{t}{2}\right).$$

Since $ab + cd \geq \min(a, c)(b + d)$ for every $a \in \mathbb{R}$, $b \geq 0$, $c \in \mathbb{R}$, and $d \geq 0$,

$$R^{\nu,\pi}_t + R^{\nu',\pi}_t \geq \frac{t}{2}\min(\Delta^{\nu}_{a'}, \mu^{\nu'}_* - \mu^{\nu}_*)\left(\mathbb{P}^{\nu,\pi}\left(T^{\pi}_{t,a'} \geq \frac{t}{2}\right) + \mathbb{P}^{\nu',\pi}\left(T^{\pi}_{t,a'} < \frac{t}{2}\right)\right).$$

Because the random variable $T^{\pi}_{t,a'}$ is $\sigma(X_1, \dots, X_{t-1})$-measurable, recall that there is a $\mathcal{B}(\mathbb{R}^{t-1})/\mathcal{B}(\mathbb{R})$-measurable function $f^{\pi}_{t-1} : \mathbb{R}^{t-1} \to \mathbb{R}$ such that $T^{\pi}_{t,a'}(\omega) = f^{\pi}_{t-1}(X_1(\omega), \dots, X_{t-1}(\omega))$ for every $\omega \in \Omega$. If $\mathcal{L}^{\nu,\pi}_{1:t-1}$ denotes the joint law of $X_1, \dots, X_{t-1}$ under $\mathbb{P}^{\nu,\pi}$ and $\mathcal{L}^{\nu',\pi}_{1:t-1}$ denotes the joint law of $X_1, \dots, X_{t-1}$ under $\mathbb{P}^{\nu',\pi}$,

$$R^{\nu,\pi}_t + R^{\nu',\pi}_t \geq \frac{t}{2}\min(\Delta^{\nu}_{a'}, \mu^{\nu'}_* - \mu^{\nu}_*)\left(\mathcal{L}^{\nu,\pi}_{1:t-1}\left(f^{\pi}_{t-1} \geq \frac{t}{2}\right) + \mathcal{L}^{\nu',\pi}_{1:t-1}\left(f^{\pi}_{t-1} < \frac{t}{2}\right)\right).$$

By Theorem 9.1, since $\mathcal{L}^{\nu,\pi}_{1:t-1}$ and $\mathcal{L}^{\nu',\pi}_{1:t-1}$ are probability measures on the measurable space $(\mathbb{R}^{t-1}, \mathcal{B}(\mathbb{R}^{t-1}))$,

$$R^{\nu,\pi}_t + R^{\nu',\pi}_t \geq \frac{t}{2}\min(\Delta^{\nu}_{a'}, \mu^{\nu'}_* - \mu^{\nu}_*)\frac{e^{-D(\mathcal{L}^{\nu,\pi}_{1:t-1}, \mathcal{L}^{\nu',\pi}_{1:t-1})}}{2} = \frac{t}{4}\min(\Delta^{\nu}_{a'}, \mu^{\nu'}_* - \mu^{\nu}_*)e^{-D(\mathcal{L}^{\nu,\pi}_{1:t-1}, \mathcal{L}^{\nu',\pi}_{1:t-1})}.$$

By Theorem 10.1, since $D(P_a, P'_a) = 0$ for every $a \neq a'$ and $D(P_{a'}, P'_{a'}) < \infty$,

$$D(\mathcal{L}^{\nu,\pi}_{1:t-1}, \mathcal{L}^{\nu',\pi}_{1:t-1}) = \sum_a D(P_a, P'_a)\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t-1,a}\right) = D(P_{a'}, P'_{a'})\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t-1,a'}\right) \leq D(P_{a'}, P'_{a'})\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a'}\right).$$

By returning to a previous inequality,

$$R^{\nu,\pi}_t + R^{\nu',\pi}_t \geq \frac{t}{4}\min(\Delta^{\nu}_{a'}, \mu^{\nu'}_* - \mu^{\nu}_*)e^{-D(P_{a'}, P'_{a'})\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a'}\right)}.$$

□

12 Minimax lower bounds

Consider a number of actions $n \in \mathbb{N}^+$ and an environment class $\mathcal{E}$ for the set of actions $\mathcal{A} = \{1, \dots, n\}$. Let $(\Omega, \mathcal{F}, \mathbb{P}^{\nu,\pi})$ denote a canonical triple for a stochastic bandit $\nu \in \mathcal{E}$ and a policy $\pi = (\pi_t : \mathbb{R}^t \to \mathcal{A} \mid t \in \mathbb{N}^+)$.

Definition 12.1. The worst-case regret $R_t^{\mathcal{E},\pi}$ of policy π on the class $\mathcal{E}$ after $t \in \mathbb{N}^+$ time steps is given by

$$R_t^{\mathcal{E},\pi} = \sup_{\nu \in \mathcal{E}} R_t^{\nu,\pi}.$$

Definition 12.2. The minimax regret $R_t^{\mathcal{E},*}$ of the environment class $\mathcal{E}$ after $t \in \mathbb{N}^+$ time steps is given by

$$R_t^{\mathcal{E},*} = \inf_{\pi} R_t^{\mathcal{E},\pi}.$$

Definition 12.3. A policy π is minimax optimal on the environment class $\mathcal{E}$ after $t \in \mathbb{N}^+$ time steps if $R_t^{\mathcal{E},\pi} = R_t^{\mathcal{E},*}$.

Definition 12.4. The Gaussian measure $P : \mathcal{B}(\mathbb{R}) \to [0,1]$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ is given by

$$P(B) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_B e^{-\frac{(x-\mu)^2}{2\sigma^2}} \; \mathrm{Leb}(dx),$$

where π denotes the circle constant (as opposed to a policy), so that P is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Definition 12.5. A stochastic bandit $\nu = (P_a \mid a \in \mathcal{A})$ is a Gaussian bandit with variance $\sigma^2 > 0$ if P_a is the Gaussian measure with mean μ_a^ν and variance σ^2 for every $a \in \mathcal{A}$.

Definition 12.6. Let $\mathcal{E}_{\mathcal{N}}^{n,\sigma^2}$ denote the set of Gaussian bandits with variance σ^2 for the set of actions $\mathcal{A} = \{1, \dots, n\}$.

Theorem 12.1. The minimax regret $R_t^{\mathcal{E}_{\mathcal{N}}^{n,1},*}$ of the environment class $\mathcal{E}_{\mathcal{N}}^{n,1}$ after $t > 1$ time steps is at least

$$R_t^{\mathcal{E}_{\mathcal{N}}^{n,1},*} \geq \frac{1}{27}\sqrt{(n-1)t}.$$

Proof. The claim is trivial if $n = 1$. Therefore, suppose that $n > 1$. For some $t > 1$, let $\Delta = \sqrt{(n-1)/4t} > 0$ and consider an arbitrary policy π for the set of actions $\mathcal{A} = \{1, \dots, n\}$.

Let $\nu = (P_a \mid a \in \mathcal{A})$ denote a Gaussian bandit with variance 1 such that $\mu_1^\nu = \Delta$ and $\mu_a^\nu = 0$ for every $a > 1$. Note that $\Delta_1^\nu = 0$ and $\Delta_a^\nu = \mu_*^\nu - \mu_a^\nu = \Delta$ for every $a > 1$.

Let $a' \in \mathcal{A}$ denote an action such that $a' > 1$ and $\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right) = \min_{a>1} \mathbb{E}^{\nu,\pi}\left(T_{t,a}^\pi\right)$. Let $\nu' = (P'_a \mid a \in \mathcal{A})$ denote a Gaussian bandit with variance 1 such that $\mu_a^{\nu'} = \mu_a^\nu$ for every $a \neq a'$ and $\mu_{a'}^{\nu'} = 2\Delta$. Note that $\Delta_1^{\nu'} = \Delta$, $\Delta_{a'}^{\nu'} = 0$, and $\Delta_a^{\nu'} = 2\Delta$ for every $a > 1$ such that $a \neq a'$.

For every $a \in \mathcal{A}$, P_a and P'_a are Gaussian measures with variance 1, so that $D(P_a, P'_a) = (\mu_a^\nu - \mu_a^{\nu'})^2/2$. Therefore, by Theorem 11.1,

$$R_t^{\nu,\pi} + R_t^{\nu',\pi} \geq \frac{t}{4}\min(\Delta_{a'}^\nu, \mu_*^{\nu'} - \mu_*^\nu) e^{-D(P_{a'}, P'_{a'})\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)} = \frac{t}{4}\Delta e^{-2\Delta^2 \mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)}.$$

Since $t = \sum_a \mathbb{E}^{\nu,\pi}\left(T_{t,a}^\pi\right)$ and $\mathbb{E}^{\nu,\pi}\left(T_{t,a}^\pi\right) \geq \mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)$ for every $a > 1$ such that $a \neq a'$,

$$t = \mathbb{E}^{\nu,\pi}\left(T_{t,1}^\pi\right) + \mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right) + \sum_{a>1 \mid a \neq a'} \mathbb{E}^{\nu,\pi}\left(T_{t,a}^\pi\right) \geq \mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right) + (n-2)\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right) = (n-1)\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right),$$

so that $\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right) \leq t/(n-1)$. By returning to a previous inequality,

$$R_t^{\nu,\pi} + R_t^{\nu',\pi} \geq \frac{t}{4}\Delta e^{-2\Delta^2 \mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)} \geq \frac{t}{4}\Delta e^{-\frac{2\Delta^2 t}{n-1}}.$$

Since $\max(x, y) \geq (x+y)/2$ for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$ and $\Delta = \sqrt{(n-1)/4t}$,

$$\max(R_t^{\nu,\pi}, R_t^{\nu',\pi}) \geq \frac{R_t^{\nu,\pi} + R_t^{\nu',\pi}}{2} \geq \frac{t}{8}\Delta e^{-\frac{2\Delta^2 t}{n-1}} = \frac{e^{-\frac{1}{2}}}{16}\sqrt{(n-1)t} \geq \frac{1}{27}\sqrt{(n-1)t}.$$

In summary, we have shown that for every policy π, number of actions $n > 1$, and time step $t > 1$, it is possible to find Gaussian bandits ν and ν' with variance 1 such that either $R_t^{\nu,\pi} \geq \sqrt{(n-1)t}/27$ or $R_t^{\nu',\pi} \geq \sqrt{(n-1)t}/27$. Therefore, for every policy π, number of actions $n \in \mathbb{N}^+$, and time step $t > 1$, we know that $R_t^{\mathcal{E}_{\mathcal{N}}^{n,1},\pi} \geq \sqrt{(n-1)t}/27$. Consequently, $R_t^{\mathcal{E}_{\mathcal{N}}^{n,1},*} = \inf_\pi R_t^{\mathcal{E}_{\mathcal{N}}^{n,1},\pi} \geq \sqrt{(n-1)t}/27$. □

13 Asymptotic lower bounds

Consider a number of actions $n \in \mathbb{N}^+$ and an environment class $\mathcal{E}$ for the set of actions $\mathcal{A} = \{1, \dots, n\}$. Let $(\Omega, \mathcal{F}, \mathbb{P}^{\nu,\pi})$ denote a canonical triple for a stochastic bandit $\nu \in \mathcal{E}$ and a policy $\pi = (\pi_t : \mathbb{R}^t \to \mathcal{A} \mid t \in \mathbb{N}^+)$.

Definition 13.1. A policy $\pi = (\pi_t : \mathbb{R}^t \to \mathcal{A} \mid t \in \mathbb{N}^+)$ is consistent over the environment class $\mathcal{E}$ if

$$\lim_{t\to\infty} \frac{R_t^{\nu,\pi}}{t^p} = 0$$

for every stochastic bandit $\nu \in \mathcal{E}$ and constant $p > 0$.

Definition 13.2. The environment class $\mathcal{E}$ is unstructured if $\mathcal{E} = \prod_a \mathcal{M}_a$, where $\mathcal{M}_a$ is a set of probability measures on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for every $a \in \mathcal{A}$.

Definition 13.3. If $\mathcal{E} = \prod_a \mathcal{M}_a$ is an unstructured environment class and $l \in \mathbb{R}$, then the set $\mathcal{M}_a^l$ is given by

$$\mathcal{M}_a^l = \left\{ P_a \in \mathcal{M}_a \mid \int_{\mathbb{R}} x \ P_a(dx) > l \right\}.$$

Definition 13.4. An unstructured environment class $\mathcal{E} = \prod_a \mathcal{M}_a$ is well-unstructured if:

- For every $a \in \mathcal{A}$, if $P_a \in \mathcal{M}_a$ and $P'_a \in \mathcal{M}_a$ are measures such that $P_a \neq P'_a$, then $D(P_a, P'_a) \in (0, \infty)$.
- For every stochastic bandit $\nu \in \mathcal{E}$ and action $a \in \mathcal{A}$, if $\Delta_a^\nu > 0$, then $\mathcal{M}_a^{\mu_*^\nu} \neq \emptyset$.

Proposition 13.1. The environment class $\mathcal{E}_{\mathcal{N}}^{n,1}$ is well-unstructured.

Proof. For every $a \in \mathcal{A}$, let $\mathcal{M}_a$ denote the set of Gaussian measures with variance 1, so that $\mathcal{E}_{\mathcal{N}}^{n,1} = \prod_a \mathcal{M}_a$. For every $a \in \mathcal{A}$, recall that if $P_a \in \mathcal{M}_a$ is a Gaussian measure with mean $\mu \in \mathbb{R}$ and variance 1 and $P'_a \in \mathcal{M}_a$ is a Gaussian measure with mean $\mu' \in \mathbb{R}$ and variance 1, then $D(P_a, P'_a) = (\mu - \mu')^2/2$. Therefore, if $P_a \neq P'_a$, then $D(P_a, P'_a) \in (0, \infty)$. Furthermore, $\mathcal{M}_a^\mu \neq \emptyset$ for every $a \in \mathcal{A}$ and $\mu \in \mathbb{R}$. □

Theorem 13.1. If $\mathcal{E} = \prod_a \mathcal{M}_a$ is a well-unstructured environment class and a policy π is consistent over $\mathcal{E}$, then

$$\liminf_{t\to\infty} \frac{R_t^{\nu,\pi}}{\log(t)} \geq \sum_{a \mid \Delta_a^\nu > 0} \frac{\Delta_a^\nu}{\inf_{P'_a \in \mathcal{M}_a^{\mu_*^\nu}} D(P_a, P'_a)}$$

for every stochastic bandit $\nu = (P_a \in \mathcal{M}_a \mid a \in \mathcal{A})$.

Proof. Consider a policy π that is consistent over $\mathcal{E}$ and a stochastic bandit $\nu = (P_a \in \mathcal{M}_a \mid a \in \mathcal{A})$. The claim is trivial if $\Delta_a^\nu = 0$ for every $a \in \mathcal{A}$, so suppose that $n > 1$ and $\Delta_{a'}^\nu > 0$ for at least one action $a' \in \mathcal{A}$.

For any action $a' \in \mathcal{A}$ such that $\Delta_{a'}^\nu > 0$, consider a stochastic bandit $\nu' = (P'_a \in \mathcal{M}_a \mid a \in \mathcal{A})$ such that $P'_a = P_a$ for every $a \neq a'$ and $P'_{a'} \in \mathcal{M}_{a'}^{\mu_*^\nu}$, so that $\mu_*^{\nu'} = \mu_{a'}^{\nu'} > \mu_*^\nu$ and $D(P_{a'}, P'_{a'}) \in (0, \infty)$.

By Theorem 11.1, for every $t > 1$,

$$R_t^{\nu,\pi} + R_t^{\nu',\pi} \geq \frac{t}{4} \min(\Delta_{a'}^\nu, \mu_*^{\nu'} - \mu_*^\nu) e^{-D(P_{a'}, P'_{a'}) \mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)}.$$

Because the right side of the inequality above is positive,

$$\log\left(R_t^{\nu,\pi} + R_t^{\nu',\pi}\right) \geq \log(t) - \log(4) + \log\left(\min(\Delta_{a'}^\nu, \mu_*^{\nu'} - \mu_*^\nu)\right) - D(P_{a'}, P'_{a'}) \mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right).$$

By rearranging and dividing both sides of the inequality above by $\log(t)$,

$$D(P_{a'}, P'_{a'}) \frac{\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)}{\log(t)} \geq \frac{\log(t) - \log(4) + \log\left(\min(\Delta_{a'}^\nu, \mu_*^{\nu'} - \mu_*^\nu)\right) - \log\left(R_t^{\nu,\pi} + R_t^{\nu',\pi}\right)}{\log(t)}.$$

By taking the limit inferior when $t \to \infty$ and the superadditivity of the limit inferior,

$$D(P_{a'}, P'_{a'}) \liminf_{t\to\infty} \frac{\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)}{\log(t)} \geq 1 + \liminf_{t\to\infty} -\frac{\log\left(R_t^{\nu,\pi} + R_t^{\nu',\pi}\right)}{\log(t)}.$$

By the relationship between the limit inferior and the limit superior,

$$D(P_{a'}, P'_{a'}) \liminf_{t\to\infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a'}\right)}{\log(t)} \geq 1 - \limsup_{t\to\infty} \frac{\log\left(R^{\nu,\pi}_t + R^{\nu',\pi}_t\right)}{\log(t)}.$$

For every $p > 0$, because the policy π is consistent over the environment class $\mathcal{E}$,

$$0 = \lim_{t\to\infty} \frac{R^{\nu,\pi}_t}{t^p} + \lim_{t\to\infty} \frac{R^{\nu',\pi}_t}{t^p} = \lim_{t\to\infty} \frac{R^{\nu,\pi}_t + R^{\nu',\pi}_t}{t^p}.$$

Therefore, for every $p > 0$ and $\epsilon > 0$ there is a $T > 1$ such that $t \geq T$ implies $(R^{\nu,\pi}_t + R^{\nu',\pi}_t)/t^p < \epsilon$. Since $R^{\nu,\pi}_t + R^{\nu',\pi}_t > 0$ by a previous inequality, by rearranging and taking the logarithm,

$$\log\left(R^{\nu,\pi}_t + R^{\nu',\pi}_t\right) \leq \log\left(\epsilon t^p\right) = \log\left(\epsilon\right) + p\log\left(t\right).$$

By dividing both sides by $\log(t)$, for every $p > 0$ and $\epsilon > 0$ there is a $T > 1$ such that $t \geq T$ implies

$$\frac{\log\left(R^{\nu,\pi}_t + R^{\nu',\pi}_t\right)}{\log\left(t\right)} \leq \frac{\log\left(\epsilon\right)}{\log\left(t\right)} + p.$$

Therefore, $\limsup_{t\to\infty} \log\left(R^{\nu,\pi}_t + R^{\nu',\pi}_t\right) / \log(t) \leq p$ for every $p > 0$. By returning to a previous inequality,

$$D(P_{a'}, P'_{a'}) \liminf_{t\to\infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a'}\right)}{\log(t)} \geq 1 - \limsup_{t\to\infty} \frac{\log\left(R^{\nu,\pi}_t + R^{\nu',\pi}_t\right)}{\log(t)} \geq 1.$$

In summary, for every action $a \in \mathcal{A}$ such that $\Delta^{\nu}_a > 0$ and $P'_a \in \mathcal{M}^{\mu^{\nu}_*}_a$,

$$D(P_a, P'_a) \liminf_{t\to\infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right)}{\log(t)} \geq 1.$$

For every action $a \in \mathcal{A}$ such that $\Delta^{\nu}_a > 0$, unless the expression on the left side below is $0 \cdot \infty$,

$$\left(\inf_{P'_a \in \mathcal{M}^{\mu^{\nu}_*}_a} D(P_a, P'_a)\right) \liminf_{t\to\infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right)}{\log(t)} = \inf_{P'_a \in \mathcal{M}^{\mu^{\nu}_*}_a} \left(D(P_a, P'_a) \liminf_{t\to\infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right)}{\log(t)}\right) \geq 1.$$

Therefore, for every action $a \in \mathcal{A}$ such that $\Delta^{\nu}_a > 0$,

$$\liminf_{t\to\infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right)}{\log(t)} \geq \frac{1}{\inf_{P'_a \in \mathcal{M}^{\mu^{\nu}_*}_a} D(P_a, P'_a)}.$$

By Theorem 4.2 and the superadditivity of the limit inferior,

$$\liminf_{t\to\infty} \frac{R^{\nu,\pi}_t}{\log(t)} = \liminf_{t\to\infty} \sum_{a | \Delta^{\nu}_a > 0} \Delta^{\nu}_a \frac{\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right)}{\log(t)} \geq \sum_{a | \Delta^{\nu}_a > 0} \Delta^{\nu}_a \liminf_{t\to\infty} \frac{\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right)}{\log(t)} \geq \sum_{a | \Delta^{\nu}_a > 0} \frac{\Delta^{\nu}_a}{\inf_{P'_a \in \mathcal{M}^{\mu^{\nu}_*}_a} D(P_a, P'_a)}.$$

□

Proposition 13.2. If a policy π is consistent over the environment class $\mathcal{E}^{n,1}_{\mathcal{N}}$ and $\nu \in \mathcal{E}^{n,1}_{\mathcal{N}}$, then

$$\liminf_{t\to\infty} \frac{R^{\nu,\pi}_t}{\log(t)} \geq 2 \sum_{a | \Delta^{\nu}_a > 0} \frac{1}{\Delta^{\nu}_a}.$$

Proof. For every $a \in \mathcal{A}$, let $\mathcal{M}_a$ denote the set of Gaussian measures with variance 1, so that $\mathcal{E}^{n,1}_{\mathcal{N}} = \prod_a \mathcal{M}_a$. For every stochastic bandit $\nu = (P_a \in \mathcal{M}_a \mid a \in \mathcal{A})$ and action $a \in \mathcal{A}$,

$$\inf_{P'_a \in \mathcal{M}^{\mu^{\nu}_*}_a} D(P_a, P'_a) = \inf_{\mu' > \mu^{\nu}_*} \frac{(\mu^{\nu}_a - \mu')^2}{2} = \frac{(\mu^{\nu}_a - \mu^{\nu}_*)^2}{2} = \frac{(-\Delta^{\nu}_a)^2}{2} = \frac{(\Delta^{\nu}_a)^2}{2}.$$

By Theorem 13.1, since the environment class $\mathcal{E}_{\mathcal{N}}^{n,1}$ is well-unstructured,

$$\liminf_{t\to\infty} \frac{R_t^{\nu,\pi}}{\log(t)} \geq \sum_{a|\Delta_a^\nu>0} \frac{\Delta_a^\nu}{\inf_{P_a'\in\mathcal{M}_a^{\mu_*^\nu}} D(P_a, P_a')} = 2 \sum_{a|\Delta_a^\nu>0} \frac{1}{\Delta_a^\nu}.$$

□

Definition 13.5. A policy π is asymptotically optimal on a well-unstructured environment class $\mathcal{E} = \prod_a \mathcal{M}_a$ if

$$\lim_{t\to\infty} \frac{R_t^{\nu,\pi}}{\log(t)} = \sum_{a|\Delta_a^\nu>0} \frac{\Delta_a^\nu}{\inf_{P_a'\in\mathcal{M}_a^{\mu_*^\nu}} D(P_a, P_a')}$$

for every stochastic bandit $\nu = (P_a \in \mathcal{M}_a \mid a \in \mathcal{A})$.

14 Finite-time lower bounds

Consider a number of actions $n \in \mathbb{N}^+$ and the environment class $\mathcal{E}_{\mathcal{N}}^{n,1}$ for the set of actions $\mathcal{A} = \{1, \dots, n\}$. Let $(\Omega, \mathcal{F}, \mathbb{P}^{\nu,\pi})$ denote a canonical triple for a stochastic bandit $\nu \in \mathcal{E}_{\mathcal{N}}^{n,1}$ and a policy $\pi = (\pi_t : \mathbb{R}^t \to \mathcal{A} \mid t \in \mathbb{N}^+)$.

Definition 14.1. For every stochastic bandit $\nu \in \mathcal{E}_{\mathcal{N}}^{n,1}$, the environment class $\mathcal{E}^\nu$ is given by

$$\mathcal{E}^\nu = \{\nu' \in \mathcal{E}_{\mathcal{N}}^{n,1} \mid \mu_a^{\nu'} \in [\mu_a^\nu, \mu_a^\nu + 2\Delta_a^\nu] \text{ for every } a \in \mathcal{A}\}.$$

Theorem 14.1. Consider a stochastic bandit $\nu \in \mathcal{E}_{\mathcal{N}}^{n,1}$. If there is a policy $\pi = (\pi_t : \mathbb{R}^t \to \mathcal{A} \mid t \in \mathbb{N}^+)$, a time step $t > 1$, a constant $C > 0$, and a constant $p \in (0, 1)$ such that $R_t^{\nu',\pi} \leq Ct^p$ for every $\nu' \in \mathcal{E}^\nu$, then, for every $\epsilon \in (0, 1]$,

$$R_t^{\nu,\pi} \geq \frac{2}{(1+\epsilon)^2} \sum_{a \mid \Delta_a^\nu > 0} \max\left(\frac{(1-p)\log(t) + \log(\epsilon\Delta_a^\nu/8C)}{\Delta_a^\nu}, 0\right).$$

Proof. Consider a stochastic bandit $\nu = (P_a \mid a \in \mathcal{A})$ such that $\nu \in \mathcal{E}_{\mathcal{N}}^{n,1}$ and let $\epsilon \in (0, 1]$. The claim is trivial if $\Delta_a^\nu = 0$ for every $a \in \mathcal{A}$, so suppose that $n > 1$ and $\Delta_{a'}^\nu > 0$ for at least one action $a' \in \mathcal{A}$.

Suppose that there is a policy $\pi = (\pi_t : \mathbb{R}^t \to \mathcal{A} \mid t \in \mathbb{N}^+)$, a time step $t > 1$, a constant $C > 0$, and a constant $p \in (0, 1)$ such that $R_t^{\nu',\pi} \leq Ct^p$ for every $\nu' \in \mathcal{E}^\nu$.

For any action $a' \in \mathcal{A}$ such that $\Delta_{a'}^\nu > 0$, consider a stochastic bandit $\nu' = (P'_a \mid a \in \mathcal{A})$ such that $P'_a = P_a$ for every $a \neq a'$. Let $P'_{a'}$ be a Gaussian measure with mean $\mu_{a'}^{\nu'} = \mu_{a'}^\nu + \Delta_{a'}^\nu(1+\epsilon)$ and variance 1. Note that $\mu_{a'}^{\nu'} > \mu_{a'}^\nu + \Delta_{a'}^\nu = \mu_*^\nu$ and $\mu_{a'}^{\nu'} \leq \mu_{a'}^\nu + 2\Delta_{a'}^\nu$, so that $\nu' \in \mathcal{E}^\nu$ and $\mu_*^{\nu'} = \mu_{a'}^{\nu'} > \mu_*^\nu$.

By Theorem 11.1, since $D(P_{a'}, P'_{a'}) = (\mu_{a'}^\nu - \mu_{a'}^{\nu'})^2/2 = (\Delta_{a'}^\nu)^2(1+\epsilon)^2/2$,

$$R_t^{\nu,\pi} + R_t^{\nu',\pi} \geq \frac{t}{4}\min(\Delta_{a'}^\nu, \mu_*^{\nu'} - \mu_*^\nu)e^{-D(P_{a'},P'_{a'})\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)} = \frac{t}{4}\epsilon\Delta_{a'}^\nu e^{-\frac{1}{2}(\Delta_{a'}^\nu)^2(1+\epsilon)^2\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)},$$

where we also used the fact that $\min(\Delta_{a'}^\nu, \mu_*^{\nu'} - \mu_*^\nu) = \min(\Delta_{a'}^\nu, \mu_{a'}^\nu + \Delta_{a'}^\nu + \epsilon\Delta_{a'}^\nu - \mu_*^\nu) = \min(\Delta_{a'}^\nu, \epsilon\Delta_{a'}^\nu) = \epsilon\Delta_{a'}^\nu$.

Since $\nu \in \mathcal{E}^\nu$ and $\nu' \in \mathcal{E}^\nu$,

$$2Ct^p \geq R_t^{\nu,\pi} + R_t^{\nu',\pi} \geq \frac{t}{4}\epsilon\Delta_{a'}^\nu e^{-\frac{1}{2}(\Delta_{a'}^\nu)^2(1+\epsilon)^2\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right)}.$$

Since the right side of the inequality above is positive, by taking the logarithm,

$$\log(2C) + p\log(t) \geq \log(t) + \log(\epsilon\Delta_{a'}^\nu/4) - \frac{1}{2}(\Delta_{a'}^\nu)^2(1+\epsilon)^2\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right).$$

By rearranging terms, since $(\Delta_{a'}^\nu)^2(1+\epsilon)^2 > 0$,

$$\mathbb{E}^{\nu,\pi}\left(T_{t,a'}^\pi\right) \geq \frac{2}{(\Delta_{a'}^\nu)^2(1+\epsilon)^2}\left((1-p)\log(t) + \log(\epsilon\Delta_{a'}^\nu/8C)\right).$$

In summary, for every $a \in \mathcal{A}$ such that $\Delta_a^\nu > 0$,

$$\mathbb{E}^{\nu,\pi}\left(T_{t,a}^\pi\right) \geq \max\left(\frac{2}{(\Delta_a^\nu)^2(1+\epsilon)^2}\left((1-p)\log(t) + \log(\epsilon\Delta_a^\nu/8C)\right), 0\right).$$

By Theorem 4.2,

$$R_t^{\nu,\pi} = \sum_{a \mid \Delta_a^\nu > 0} \Delta_a^\nu \mathbb{E}^{\nu,\pi}\left(T_{t,a}^\pi\right) \geq \sum_{a \mid \Delta_a^\nu > 0} \Delta_a^\nu \max\left(\frac{2}{(\Delta_a^\nu)^2(1+\epsilon)^2}\left((1-p)\log(t) + \log(\epsilon\Delta_a^\nu/8C)\right), 0\right).$$

By rearranging terms,

$$R_t^{\nu,\pi} \geq \frac{2}{(1+\epsilon)^2} \sum_{a \mid \Delta_a^\nu > 0} \max\left(\frac{(1-p)\log(t) + \log(\epsilon\Delta_a^\nu/8C)}{\Delta_a^\nu}, 0\right).$$

□

Acknowledgements

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

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] Pollard, D. *A User's Guide to Measure Theoretic Probability.* Cambridge University Press, 2002.

[5] Rivasplata, O. *Subgaussian random variables: An expository note.* 2012.

[6] Wainwright, M. J. *High-Dimensional Statistics - A Non-Asymptotic Viewpoint.* Cambridge University Press, 2019.