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 $\infty \cdot 0$,

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

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 $\infty \cdot 0$,

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

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.

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

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

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

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

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

$\square$

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.

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

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

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

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

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

□

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,

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

□

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

Proof. By Proposition 1.6,

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

□

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

Proof. By Proposition 1.7,

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

□

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

Proof. By Proposition 1.7,

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

□

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 $f \in o(g)$. $\square$

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. $\square$

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 $f \succ h$.

$\square$

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 $f \succsim h$.

$\square$

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

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

Proposition 1.20. The binary relations defined on the set $\mathscr{F}$ 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 $\liminf_{n\to\infty} f(n)/g(n) \leq \limsup_{n\to\infty} f(n)/g(n)$. $\square$

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 $\text{Var}(X) \leq \sigma^2$.

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

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

Since $\mathbb{E}\left(e^{\lambda|X|}\right) < \infty$, 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,

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

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,

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

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. $\square$

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. $\square$

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 $\epsilon \geq 0$,

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

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,

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

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

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

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

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

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

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

□

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

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

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

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

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

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

Consider a sequence of independent random variables $(X_k : \Omega \to \mathbb{R} \mid k \in \mathbb{N}^+)$, 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 $X(\omega) \in \{1, \dots, m\}$ for every $\omega \in \Omega$. Consider a vector $p \in [0, 1]^m$ such that $p_i = \mathbb{P}(X = i)$ for every $i \in \{1, \dots, m\}$ and a random vector $P_t : \Omega \to [0, 1]^m$ such that $P_{t,i} = 1/t \sum_{k=1}^t \mathbb{I}_{\{X_k = i\}}$ for every $t \in \mathbb{N}^+$ and $i \in \{1, \dots, m\}$. 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 $k \in \{1, \dots, t\}$ and $\lambda \in \{-1,1\}^m$, let $Y_k^{(\lambda)} = \sum_{i=1}^{m} \lambda_i \mathbb{I}_{\{X_k = i\}} = \lambda_{X_k}$, so that $|Y_k^{(\lambda)}| \leq 1$ and

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

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

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

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

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

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

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

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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Suppose that there is a constant $c \in (0, \infty)$ such that $\int_{\mathbb{R}} |x| P_a(dx) \leq c$ for every action $a \in \mathcal{A}$. In that case, the mean μ_a^ν of action a is defined by $\mu_a^\nu = \int_{\mathbb{R}} x P_a(dx)$ and the highest mean μ_*^ν is defined by $\mu_*^\nu = \sup_a \mu_a^\nu$. If $\mu_a^\nu = \mu_*^\nu$ for some $a \in \mathcal{A}$, then ν is a stochastic bandit for the set of actions $\mathcal{A}$.

Definition 4.3. 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 4.1. 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 $A_t = \pi_t(X_0, \ldots, X_{t-1})$.

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,

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

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,

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

$\square$

Definition 4.4. 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 $A_t = \pi_t(X_0, \ldots, X_{t-1})$. 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 4.1 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

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

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

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

Because $\mathcal{L}_t$ is the law of X_t and $\mathcal{L}_t$ is the law of $\tilde{X}_t^{\nu,\pi}$,

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

□

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

Proof. For every $t \in \mathbb{N}^+$, A_t is $\sigma(X_0, \dots, X_{t-1})$-measurable. Therefore, almost surely for every $B \in \mathcal{B}(\mathbb{R})$,

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

Therefore, for every Borel function $h : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}^{\nu,\pi}\left(|h(X_t)|\right) < \infty$, recall that almost surely

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

The function $h : \mathbb{R} \to \mathbb{R}$ given by $h(x) = x$ is Borel. Since $\mathbb{E}^{\nu,\pi}\left(|X_t|\right) < \infty$, almost surely,

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

□

Proposition 4.3. If $t \in \mathbb{N}^+$ and $A_t = \pi_t(X_0, \dots, X_{t-1})$, 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 4.5. 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 4.6. For every $t \in \mathbb{N}^+$, the regret $R_t^{\nu,\pi}$ of policy π on ν after t time steps is given by

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

Definition 4.7. 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 4.8. The number of times $T^\pi_{t,a} : \Omega \to \{0, \dots, t\}$ that policy π chooses $a \in \mathcal{A}$ 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 $A_k = \pi_k(X_0, \dots, X_{k-1})$ for every $k \leq t$. Note that $\sum_a T^\pi_{t,a}(\omega) = t$ for every $\omega \in \Omega$.

Definition 4.9. 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 $A_k = \pi_k(X_0, \dots, X_{k-1})$ 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 $A_k = \pi_k(X_0, \dots, X_{k-1})$ 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 4.4. If $t \in \mathbb{N}^+$, then $R^{\nu,\pi}_t \geq 0$.

Proof. Since $\Delta^{\nu}_a \geq 0$ and $\mathbb{E}^{\nu,\pi}\left(T^{\pi}_{t,a}\right) \geq 0$ for every $a \in \mathcal{A}$ and $t \in \mathbb{N}^+$, the claim is a consequence of Theorem 4.2. □

Proposition 4.5. Consider an action $a^* \in \mathcal{A}$ such that $\mu^{\nu}_{a^*} = \mu^{\nu}_*$. If $\pi_t = a^*$ for every $t \in \mathbb{N}^+$, then $R^{\nu,\pi}_t = 0$.

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

$$R^{\nu,\pi}_t = \sum_a \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}(T^{\pi}_{t,a}) = \Delta^{\nu}_{a^*} \mathbb{E}^{\nu,\pi}(T^{\pi}_{t,a^*}) = (\mu^{\nu}_* - \mu^{\nu}_{a^*}) \mathbb{E}^{\nu,\pi}(T^{\pi}_{t,a^*}) = 0.$$

□

Proposition 4.6. For every $t \in \mathbb{N}^+$, let $A_k = \pi_k(X_0, \dots, X_{k-1})$ for every $k \leq t$. If $R^{\nu,\pi}_t = 0$, then $\mu^{\nu}_{A_k} = \mu^{\nu}_*$ almost surely for every $k \leq t$.

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

$$R^{\nu,\pi}_t = \sum_a \Delta^{\nu}_a \mathbb{E}^{\nu,\pi}(T^{\pi}_{t,a}) = \sum_a \Delta^{\nu}_a \sum_{k=1}^{t} \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{A_k = a\}}\right) = \sum_{k=1}^{t} \mathbb{E}^{\nu,\pi}\left(\sum_a \mathbb{I}_{\{A_k = a\}} \Delta^{\nu}_a\right) = \sum_{k=1}^{t} \mathbb{E}^{\nu,\pi}\left(\Delta^{\nu}_{A_k}\right).$$

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

For convenience, let $R^{\nu,\pi}_0 = 0$.

Proposition 4.7. If $R^{\nu,\pi}_t = o(t)$, then

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

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

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

so that

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

□

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

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

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

Proposition 4.8. The number of times $T^{\nu,\pi}_{t,*} : \Omega \to \{0, \ldots, t\}$ that policy π chooses an optimal action on the stochastic bandit ν by time step $t \in \mathbb{N}^+$ is given by

$$T^{\nu,\pi}_{t,*}(\omega) = \sum_{a \mid \Delta^\nu_a = 0} T^\pi_{t,a}(\omega).$$

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

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

so that

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

□

Proposition 4.9. 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,

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

For every $t \in \mathbb{N}^+$, recall that $\sum_a T^\pi_{t,a} = t$. By Proposition 4.8,

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

so that

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

Therefore, considering a previous equation,

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

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

Definition 4.11. For a set of actions $\mathcal{A}$, an environment class $\mathcal{E}$ is a set of stochastic bandits for $\mathcal{A}$.

Definition 4.12. For a set of actions $\mathcal{A}$ and an environment class $\mathcal{E}$, consider a probability triple $(\mathcal{E}, \mathcal{G}, \mathbb{Q})$ such that $R_t^{\cdot,\pi} : \mathcal{E} \to [0, \infty]$ is $\mathcal{G}$-measurable for every policy π and time step $t \in \mathbb{N}^+$. The Bayesian regret B_t^π of policy π after $t \in \mathbb{N}^+$ time steps is given by

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

Definition 4.13. 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 $Z_a(x) = x - \mu_a^\nu$ 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_{t_1} and X_{t_2} are independent in $(\Omega, \mathcal{F}, \mathbb{P}^{\nu,\pi})$ for every $t_1 \in \mathbb{N}^+$ and $t_2 \in \mathbb{N}^+$ such that $t_1 < t_2 \leq mn$.

Proof. Consider $t_1 \in \mathbb{N}^+$ and $t_2 \in \mathbb{N}^+$ such that $t_1 < t_2 \leq mn$. For every $B_1 \in \mathcal{B}(\mathbb{R})$ and $B_2 \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}^{\nu,\pi}\left(X_{t_1} \in B_1, X_{t_2} \in B_2\right) = \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_{t_1} \in B_1\}}\mathbb{I}_{\{X_{t_2} \in B_2\}}\right) = \mathbb{E}^{\nu,\pi}\left(\mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_{t_1} \in B_1\}}\mathbb{I}_{\{X_{t_2} \in B_2\}} \mid X_0, \ldots, X_{t_2-1}\right)\right).$$

For every $t \in \mathbb{N}^+$ such that $t \leq mn$, let $a_t = ((t-1) \bmod n) + 1$. By taking out what is known,

$$\mathbb{P}^{\nu,\pi}\left(X_{t_1} \in B_1, X_{t_2} \in B_2\right) = \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_{t_1} \in B_1\}}\mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_{t_2} \in B_2\}} \mid X_0, \ldots, X_{t_2-1}\right)\right) = \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_{t_1} \in B_1\}}P_{a_{t_2}}(B_2)\right).$$

By Proposition 5.2, for every $B_1 \in \mathcal{B}(\mathbb{R})$ and $B_2 \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}^{\nu,\pi}\left(X_{t_1} \in B_1, X_{t_2} \in B_2\right) = \mathbb{E}^{\nu,\pi}\left(\mathbb{I}_{\{X_{t_1} \in B_1\}}\right)P_{a_{t_2}}(B_2) = \mathbb{P}^{\nu,\pi}\left(X_{t_1} \in B_1\right)\mathbb{P}^{\nu,\pi}\left(X_{t_2} \in B_2\right).$$

□

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 $t \le mn$, 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 $\lambda \in \mathbb{R}$,

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

□

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 $t \ge mn$,

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

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 $t > mn$,

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

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

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

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

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

so that

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

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

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

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

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

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

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

□

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

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

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

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

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

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

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

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

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

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

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

Because the function $f : (0,\infty) \to (0,\infty)$ 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.

Definition 5.4. A policy π restarts to the policy π' after $t \in \mathbb{N}$ steps if, for all $k \in \mathbb{N}^+$ and $(x_0, \ldots, x_{t+k-1}) \in \mathbb{R}^{t+k}$,

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

Proposition 5.6. If a policy π restarts to the policy π' after $t \in \mathbb{N}$ steps, then

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

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

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

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

where $A_{t+1} = \pi_{t+1}(X_0, \ldots, X_t) = \pi'_1(0)$. 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 $B_1, \ldots, B_k \in \mathcal{B}(\mathbb{R})$,

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

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

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

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

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

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

By taking out what is known,

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

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

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

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

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

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

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

□

Proposition 5.7. 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 5.6, 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 5.5. 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 5.8. 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 5.7, 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 5.9. 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 5.8,

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

□

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 5.10. 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 5.9,

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

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 does not necessarily provide the best regret upper bound.

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

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