Sequential Learning, Sequential Optimization

M2 course for tracks:
1- Probability & Statistics
2- Optimization
3- Data Science

Gilles Stoltz: CNRS researcher at HEC Paris
stoltz@hec.fr

Exam: Written exam in unlimited time (9h40 - [...])
On Wednesday March 29 at HEC Paris
No other option! (no project!) → Only 3 A4 2-sided cheat sheets

Beware: 5 ECTS for 18h in Data Science track:
A good deal but highly technical course:
definition / theorem / proof (sometimes: long and technical proofs)

6 × 3h from Feb. 1st to March 15 (except: Feb. 15)

Pre-requisits:
- Martingale theory
- Measure theory / Integration theory
- But no real knowledge of statistics is required

Lecture notes: On my website, "Enseignements" page
http://stoltz.perso.math.cnrs.fr

We will study:
- algorithms with associated upper bounds of performance
- lower bounds on the performance of any algorithm

Read them in details, meditate the proof techniques and summarize them on your cheat sheets

Also put FYI the statement of last year's exam
(was a disaster in the Data Science track: beware, work is needed!)

Send me an email for registration!

Course 1 / February 1st, 2017

0. Reminder on the Hoeffding-Azuma inequality (basic unconditional one ↑, conditional version ↗)

1. Framework and setting of stochastic bandits

2. A popular algorithm : UCB

↑ almost: up to a concentration lemma (Hoeffding-Azuma for a random number of summands)

↳ There are three exercises for you!

Lemma (Hoeffding): X random variable s.t. $X \in [a,b]$ a.s.

Then $\forall s \in \mathbb{R}$,

$$\ln \mathbb{E}\left[e^{s(X-\mathbb{E}X)} \right] = \ln \mathbb{E}\left[e^{sX}\right] - s\,\mathbb{E}[X] \leq \frac{s^2}{8}(b-a)^2$$

Proof (most elegant one I know of):

$$\Psi(s) = \ln \mathbb{E}\left[e^{sX}\right] \quad \text{defined for all } s \in \mathbb{R}$$

Ψ is differentiable at each $s \in \mathbb{R}$: cf. X bounded, thus $\eta \mapsto Xe^{\eta X}$ locally dominated around s by an integrable r.v. independent of η, thus $\eta \mapsto \mathbb{E}[e^{\eta X}]$ differentiable at s with derivative $\mathbb{E}[Xe^{sX}]$

with

$$\Psi'(s) = \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}$$

Similarly, Ψ is twice differentiable at each $s \in \mathbb{R}$, with:

$$\Psi''(s) = \frac{\mathbb{E}[X^2 e^{sX}]\,\mathbb{E}[e^{sX}] - \left(\mathbb{E}[Xe^{sX}]\right)^2}{\left(\mathbb{E}[e^{sX}]\right)^2} = \mathrm{Var}_Q(X)$$

under the probability Q defined by

$$\frac{dQ}{d\mathbb{P}}(w) = \frac{e^{sX(w)}}{\mathbb{E}[e^{sX}]}$$

$X \in [a,b]$: $\mathrm{Var}_Q(X) = \inf_{\mu \in \mathbb{R}} \mathbb{E}_Q\left[(X-\mu)^2\right]$

$$\leq \mathbb{E}_Q\left[\left(X - \tfrac{a+b}{2}\right)^2\right] = \frac{(b-a)^2}{4}$$

Taylor: $\exists x$ s.t. $\Psi(s) = \underbrace{\Psi(0)}_{=0} + s\underbrace{\Psi'(0)}_{=0} + \frac{s^2}{2}\underbrace{\Psi''(x)}_{\leq (b-a)^2/4}$

ie, $\ln \mathbb{E}\left[e^{sX}\right] \leq \frac{s^2}{8}(b-a)^2$

The Hoeffding-Azuma inequality

Theorem: Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration and let $(X_t)_{t\geq 1}$ be a sequence of adapted random variables (ie, $\forall t\geq 1$, X_t is $\mathcal{F}_t$-measurable), that are bounded: $\forall t$, $a_t \leq X_t \leq b_t$ a.s., where $a_t, b_t \in \mathbb{R}$.

Then (« probabilistic version »)

$$\forall \varepsilon > 0, \quad \mathbb{P}\left\{ \sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \geq \varepsilon \right\} \leq \exp\left(- \frac{2\varepsilon^2}{\sum_{t=1}^{T} (b_t - a_t)^2} \right)$$

or (« statistical version », totally equivalent)

$$\forall \delta \in (0,1), \quad \text{with probability at least } 1-\delta, \quad \sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \leq \sqrt{\frac{\sum_{t=1}^{T} (b_t - a_t)^2}{2} \ln \frac{1}{\delta}}$$

Note: Hoeffding's inequality is the special case when all X_t are independent and $\mathcal{F}_{t-1} = \sigma(X_1 \dots X_{t-1})$, so that $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[X_t]$.

Basic ingredient of the proof: extension of Hoeffding's lemma to conditional expectations

Lemma: X random variable s.t. $X \in [a,b]$ a.s.

Then, for all σ-algebras $\mathcal{G}$, for all $s \in \mathbb{R}$,

$$\ln \mathbb{E}\left[e^{s(X - \mathbb{E}[X \mid \mathcal{G}])} \mid \mathcal{G} \right] = \ln \left(\mathbb{E}[e^{sX} \mid \mathcal{G}] \right) - s\, \mathbb{E}[X \mid \mathcal{G}] \leq \frac{s^2}{8} (b-a)^2$$

(We will discuss the proof in a minute... let's first prove the theorem based on this lemma.)

Proof (of the theorem):

Markov-Chernoff bounding (= Markov's inequality after taking exponents):

We denote $S_T = \sum_{t=1}^{T} X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$

(S_T: martingale = sum of martingale increments or martingale differences)

The « probabilistic version » is about upper bounding $\mathbb{P}\{S_T > \varepsilon\}$:

$$\mathbb{P}\{S_T > \varepsilon\} = \mathbb{P}\{e^{sS_T} > e^{s\varepsilon}\} \leq e^{-s\varepsilon}\mathbb{E}[e^{sS_T}]$$

($= \mathbb{P}\{e^{sS_T} > e^{s\varepsilon}\}$: $\forall s > 0$; $\leq e^{-s\varepsilon}\mathbb{E}[e^{sS_T}]$: Markov's inequality)

We show by induction that $\mathbb{E}[e^{sS_T}] \leq \exp\left(\frac{s^2}{8}\sum_{t=1}^{T}(b_t - a_t)^2\right)$

- For $T=1$, true by the conditional version of Hoeffding's lemma and the fact that $S_1 = X_1 - \mathbb{E}[X_1 | \mathcal{F}_0]$ with $X_1 \in [a_1, b_1]$ + taking expectations
- For $T-1 \to T$, where $T \geq 2$:

  The extension of Hoeffding's lemma ensures that

$$\mathbb{E}\left[e^{s(X_T - \mathbb{E}[X_T | \mathcal{F}_{T-1}])} \,\middle|\, \mathcal{F}_{T-1}\right] \leq e^{s^2(b_T - a_T)^2/8}$$

  so that

$$\begin{aligned}\mathbb{E}[e^{sS_T}] &= \mathbb{E}\left[\mathbb{E}[e^{sS_T} | \mathcal{F}_{T-1}]\right] \\ &= \mathbb{E}\left[e^{sS_{T-1}}\mathbb{E}\left[e^{s(X_T - \mathbb{E}[X_T | \mathcal{F}_{T-1}])} \,\middle|\, \mathcal{F}_{T-1}\right]\right] \\ &\leq e^{s^2(b_T - a_T)^2/8} \times \mathbb{E}[e^{sS_{T-1}}] \quad \text{(by the induction hypothesis)} \\ &\leq \exp\left(s^2 \sum_{t \leq T} (b_t - a_t)^2/8\right)\end{aligned}$$

Substituting above:

$$\mathbb{P}\{S_T > \varepsilon\} \leq \inf_{s>0} \exp\left(-s\varepsilon + s^2 \sum_{t \leq T} \frac{(b_t - a_t)^2}{8}\right)$$

strictly convex function to minimize in the exponent: minimum achieved at s^* such that $s^* \sum_{t \leq T}(b_t - a_t)^2/4 = \varepsilon$ (gradient vanishes), i.e. $s^* = 4\varepsilon / \sum_{t \leq T}(b_t - a_t)^2$

$$= \exp\left(-2\varepsilon^2 / \sum_{t \leq T}(b_t - a_t)^2\right).$$

→ It only remains to prove the extension of Hoeffding's lemma to conditional expectations.

Back to Hoeffding's lemma with conditional expectations :

Proof 1? Can we take the proof of Hoeffding's lemma we just saw and replace all $\mathbb{E}$ by $\mathbb{E}[\,\cdot\,|\mathcal{G}_j]$?

$\psi(s) = \ln \mathbb{E}[e^{sX}|\mathcal{G}_j]$ → The theorem of differentiation under $\mathbb{E}[\,]$ only requires dominated convergence, which holds true for $\mathbb{E}[\,\cdot\,|\mathcal{G}_j]$ as well. Thus, we also have a theorem of differentiation under $\mathbb{E}[\,\cdot\,|\mathcal{G}_j]$:

a.s., $\psi''(s)$ exists and equals $$\psi''(s) = \frac{\mathbb{E}[X^2 e^{sX}|\mathcal{G}_j]\,\mathbb{E}[e^{sX}|\mathcal{G}_j] - \left(\mathbb{E}[Xe^{sX}|\mathcal{G}_j]\right)^2}{\left(\mathbb{E}[e^{sX}|\mathcal{G}_j]\right)^2}$$

= some conditional variance under a different probability measure ?

Yes, using the notion of « regular conditional probability », which always exists in our case, we could perform a change of measure again and identify a conditional variance.

But **PLEASE!** this is extremely heavy maths... I can't inflict that to you....

Proof 2 Too bad for elegance, let's get back to the original proof of Hoeffding's (unconditional) lemma, which only relies on calculus :

$Y = X - \mathbb{E}[X|\mathcal{G}_j] \in [A, B]$ where $A = a - \mathbb{E}[X|\mathcal{G}_j] \leq 0$, $B = b - \mathbb{E}[X|\mathcal{G}_j] \geq 0$ are both $\mathcal{G}_j$-measurable and $B - A = b - a > 0$

$$Y = \frac{B-Y}{B-A}A + \frac{Y-A}{B-A}B$$

(convex weights)

Since $y \mapsto e^{sy}$ is convex :

$$e^{sY} \leq \frac{B-Y}{B-A}e^{sA} + \frac{Y-A}{B-A}e^{sB}$$

Taking $\mathbb{E}[\,\cdot\,|\mathcal{G}_j]$: using $\mathbb{E}[Y|\mathcal{G}_j] = 0$ and A, B $\mathcal{G}_j$-measurable :

$$\mathbb{E}[e^{sY}|\mathcal{G}_j] \leq \frac{B}{B-A}e^{sA} - \frac{A}{B-A}e^{sB}$$

← note that $B/B-A$ and $-A/B-A$ are convex weights

Now, by a function study (the very same as the one we performed in the proof of the unconditional version of Hoeffding's lemma) — or even by the latter lemma itself:

$$\forall u,v \in \mathbb{R},\ \forall s \in \mathbb{R},\ \forall p \in [0,1],\quad \ln\left(p e^{su} + (1-p)e^{sv}\right) \leq s(pu + (1-p)v) + \frac{s^2}{8}(v-u)^2$$

($\ln$ of expected value of e^{sZ} where $Z = \begin{cases} u & \text{wp } p \\ v & \text{wp } 1-p \end{cases}$; expected value of Z; range is $[u,v]$)

In particular,

$$\frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \leq \exp\left(s\left(\frac{ba}{b-a} - \frac{ab}{b-a}\right) + \frac{s^2}{8}(b-a)^2\right) = \exp\left(\frac{s^2}{8}(b-a)^2\right)$$

Summarizing:

$$\mathbb{E}[e^{sY} \mid \mathcal{G}_j] \leq \exp\left(\frac{s^2}{8}(b-a)^2\right) = \mathbb{E}[e^{sX} \mid \mathcal{G}_j] \times \exp(-s\,\mathbb{E}[X \mid \mathcal{G}_j])$$

Proof 3 My preferred (not only because I found by myself): Hoeffding's lemma in its unconditional version ENTAILS the conditional version! This is because Hoeffding's lemma holds for all probability distributions — we should play with this fact.

For all $A \in \mathcal{G}_j$ s.t. $\mathbb{P}(A) > 0$, let $\mathbb{P}_A = \mathbb{P}(\cdot \mid A)$, the conditional distribution given the event A.

The unconditional version of Hoeffding's lemma ensures that

$$\forall A \in \mathcal{G}_j \text{ s.t. } \mathbb{P}(A) > 0,\ \forall s \in \mathbb{R},\quad \ln \mathbb{E}_A[e^{sX}] \leq s\,\mathbb{E}_A[X] + \frac{s^2}{8}(b-a)^2$$

Why do we consider the $\mathbb{E}_A$? Because $\mathbb{E}[X \mid \mathcal{G}_j]$ is the unique $\mathcal{G}_j$-measurable random variable such that

$$\forall A \in \mathcal{G}_j,\quad \mathbb{E}[X \mathbb{1}_A] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}_j]\mathbb{1}_A]$$

or, equivalently,

$$\forall A \in \mathcal{G}_j \text{ s.t. } \mathbb{P}(A) > 0,\quad \mathbb{E}_A[X] = \mathbb{E}_A[\mathbb{E}[X \mid \mathcal{G}_j]].$$

Now, consider the event

$$H = \left\{ \mathbb{E}[e^{sX} \mid \mathcal{G}_j] - e^{s\,\mathbb{E}[X \mid \mathcal{G}_j]} e^{s^2(b-a)^2/8} > 0 \right\} \in \mathcal{G}_j$$

We want to prove that $\mathbb{P}(H) = 0$. We proceed by contradiction: if we had $\mathbb{P}(H) > 0$, then $\mathbb{P}_H$ would be defined and

$$\mathbb{E}_H[e^{sX}] = \mathbb{E}_H[\mathbb{E}[e^{sX}|\mathcal{G}_j]] > \mathbb{E}_H[e^{s\mathbb{E}[X|\mathcal{G}_j]}]\, e^{s^2(b-a)^2/8}$$

(= : a property of conditional expectations; > : by definition of H and because a function > 0 on a set with probability > 0 has an expectation > 0)

$$\geq e^{s\,\mathbb{E}_H[\mathbb{E}[X|\mathcal{G}_j]]}\, e^{s^2(b-a)^2/8} = e^{s\,\mathbb{E}_H[X]}\, e^{s^2(b-a)^2/8}$$

(≥ : (unconditional) Jensen's inequality)

which would be in contradiction with the unconditional Hoeffding's lemma:

$$\ln \mathbb{E}_H[e^{sX}] \leq s\,\mathbb{E}_H[X] + \frac{s^2}{8}(b-a)^2.$$

We conclude that $\mathbb{P}(H) = 0$.

A final remark: A better version of Hoeffding's lemma in its conditional form is the following one:

> **Hoeffding's lemma:** X random variable s.t. there exists G $\mathcal{G}_j$-measurable and $a, b \in \mathbb{R}$ with: $G + a \leq X \leq G + b$ a.s.
>
> Then $\forall s \in \mathbb{R}$, $\ln \mathbb{E}[e^{sX}|\mathcal{G}_j] - s\mathbb{E}[X|\mathcal{G}_j] \leq \frac{s^2}{8}(b-a)^2$

→ In my first statement, I only considered $G \equiv 0$.
Why can we consider general G's?

Proof 2: $Y = X - \mathbb{E}[X|\mathcal{G}_j] = (X-G) - \mathbb{E}[(X-G)|\mathcal{G}_j]$ still $\in [A, B]$ and everything goes through

Proof 3: Some adaptation is needed as X may be unbounded! Only $X - G$ is bounded between a and b.

So, using the same proof scheme, we can conclude that

$$\ln \mathbb{E}[e^{s(X-G)}|\mathcal{G}_j] \leq s\,\mathbb{E}[X - G|\mathcal{G}_j] + \frac{s^2}{8}(b-a)^2$$

But I am not sure we may write e.g.

$$\mathbb{E}[X-G \mid \mathcal{G}_j] = \mathbb{E}[X \mid \mathcal{G}_j] - G$$

We may, under an additional assumption that X is integrable.

Some issues with the $\mathbb{E}[e^{s(X-G)} \mid \mathcal{G}_j]$ term: I would be more comfortable with adding the assumption that $e^{sX} \in \mathbb{L}^1$ ∀s.

Hence:

Exercise 0: Can anyone make this proof #3 work without these extra assumptions?

Now, we get back to the Hoeffding-Azuma inequality:

Hoeffding-Azuma: holds therefore for adapted processes such that for all t, there exist $(a_t, b_t) \in \mathbb{R}^2$ $\mathcal{G}_t$ $\mathcal{F}_{t-1}$-measurable with $G_t + a_t \leq X_t \leq G_t + b_t$ a.s.

Stochastic bandits. Finitely many arms.

Setting:

K arms indexed by $1, 2, \dots K$

With each arm j is associated a probability distribution ν_j (over $\mathbb{R}$) with an expectation

At each round $t = 1, 2, \dots$

- The decision-maker picks $I_t \in \{1, \dots K\}$, possibly at random
- She gets a reward Y_t drawn at random according to ν_{I_t} (given I_t)
- This is the only feedback she gets / the only observation she has access to.

Aim:

We denote by $\mu_i = \mathbb{E}(\nu_i)$ the expectation of ν_i

(note: operator $\mathbb{E}$ vs. expectation $\mathbb{E}$ of an expression involving random variables.)

Pseudo-regret $\bar{R}_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$ to be controlled

where $\mu^* = \max_{j \le K} \mu_j$

Useful notation:

$\Delta_a = \mu^* - \mu_a$ gap of arm a

$\Delta_a = 0$: a is an optimal arm (there can be several of them)

$\Delta_a > 0$: a is a suboptimal arm

$N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = a\}}$ total number of times that a is pulled.

Upper confidence bound [UCB] algorithm: Very popular!

For $t = 1, 2 \dots K$

- Pull arm $I_t = t$, get a reward Y_t

For $t = K+1, K+2, \dots$

- Pull an arm $I_t \in \underset{j \in \{1, \dots K\}}{\operatorname{argmax}} \left\{ \hat{\mu}_{j,t-1} + \sqrt{\frac{2 \ln t}{N_j(t-1)}} \right\}$

(the - breaking rule: pick the element with smallest index)

where $N_j(t-1) = \sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = j\}}$

and where $\hat{\mu}_{j,t-1} = \frac{1}{N_j(t-1)} \sum_{s=1}^{t-1} Y_s \mathbb{1}_{\{I_s = j\}}$

always ≥ 1 since each arm was tried sequentially during rounds $1, 2, \dots K$

- Get a reward Y_t

Theorem: If the distributions ν_j have supports all included in $[0,1]$, then the pseudo-regret of UCB is smaller than

$$\overline{R}_T \leq \sum_{i : \Delta_i > 0} \left(\frac{8 \ln T}{\Delta_i} + 2 \right)$$

where $\Delta_i = \mu^* - \mu_i$ is the gap of arm i.

This regret bound is obtained via the following proposition:

Proposition: If the distributions ν_j have supports all included in $[0,1]$, then

$$\forall i \text{ s.t. } \Delta_i > 0, \quad \mathbb{E}[N_i(T)] \leq \frac{8 \ln T}{\Delta_i^2} + 2.$$

Exercise 1. The bounds above are called distribution-dependent because they depend heavily on the distributions ν_i at hand (via the gaps $\Delta_i = \mu^* - \mu_i$).
Show the following distribution-free bound (that only

depends on the support $[0,1]$, not on the specific distributions ν_i at hand): for the UCB algorithm,

$$\sup_{\nu_1,\dots,\nu_K \text{ with supports in } [0,1]} \bar{R}_T \leq O\left(\sqrt{TK\ln T}\right).$$

Hint: For small values of Δ_i, the bound of the Proposition can be worse than the trivial T bound...

Proof [of the theorem based on the Proposition]:

$$\bar{R}_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$

where by definition of the bandit model, ← Given I_t, Y_t is drawn at random according to ν_{I_t}

$$\mathbb{E}[Y_t \mid I_t] = \mu_{I_t}$$

thus (by the tower rule)

$$\mathbb{E}[Y_t] = \mathbb{E}[\mu_{I_t}] = \sum_j \mu_j \, \mathbb{E}\left[\mathbb{1}_{\{I_t=j\}}\right]$$

Summing over t:

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \sum_{j=1}^{K} \mu_j \, \mathbb{E}[N_j(T)]$$

and (in view of $T = \sum_j \mathbb{E}[N_j(T)]$)

$$\bar{R}_T = \sum_j (\mu^* - \mu_j)\,\mathbb{E}[N_j(T)] = \sum_{j=1}^{K} \Delta_j \, \mathbb{E}[N_j(T)] = \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{E}[N_j(T)]$$

↓ it suffices to consider the suboptimal arms...

We conclude by substituting $\mathbb{E}[N_j(T)] \leq \frac{8\ln T}{\Delta_j^2} + 2$ and by bounding $2\Delta_j \leq 2$.

Note: Keep in mind the rewriting $\bar{R}_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$ as we will often use it!

Proof [of the Proposition]: We fix an optimal arm $a^* \in \{1, \dots K\}$, i.e. s.t. $\mu_{a^*} = \mu^*$.

→ It will show why this algorithm is called UCB:

Because $\hat{\mu}_{j,t-1} + \sqrt{\frac{2\ln t}{N_j(t-1)}}$ will indeed appear as an upper confidence bound on μ_j

- $\hat{\mu}_{j,t-1}$: estimate based on the raw performance ↔ exploitation of the results
- $\sqrt{\frac{2\ln t}{N_j(t-1)}}$: larger for arms j not much sampled so far ↔ forces some exploration

The UCB algorithm realizes some compromise / trade-off between exploitation & exploration.

We compare these statements to the Hoeffding-Azuma inequality (⚠) in the next lecture

LEMMA: $\forall j$, $\forall t \geq j$ (so that $N_j(t) \geq 1$)

$$\forall \delta \in (0,1), \quad \mathbb{P}\left\{ \mu_j > \hat{\mu}_{j,t} - \sqrt{\frac{\ln(1/\delta)}{2N_j(t)}} \right\} \geq 1 - t\delta.$$

[If $Y_i \sim \nu_i$ supported by $[0,1]$, $\mu_i = \mathbb{E}[Y_i]$, $1-Y_i \sim \nu_i'$ then $\mu_i' = 1 - \mu_i$ as well and ν_i' supported by $[0,1]$]

reduce → By symmetry: $\forall \delta \in (0,1)$,

$$\mathbb{P}\left\{ \mu_j < \hat{\mu}_{j,t} + \sqrt{\frac{\ln(1/\delta)}{2N_j(t)}} \right\} \geq 1 - t\delta$$

→ Application:

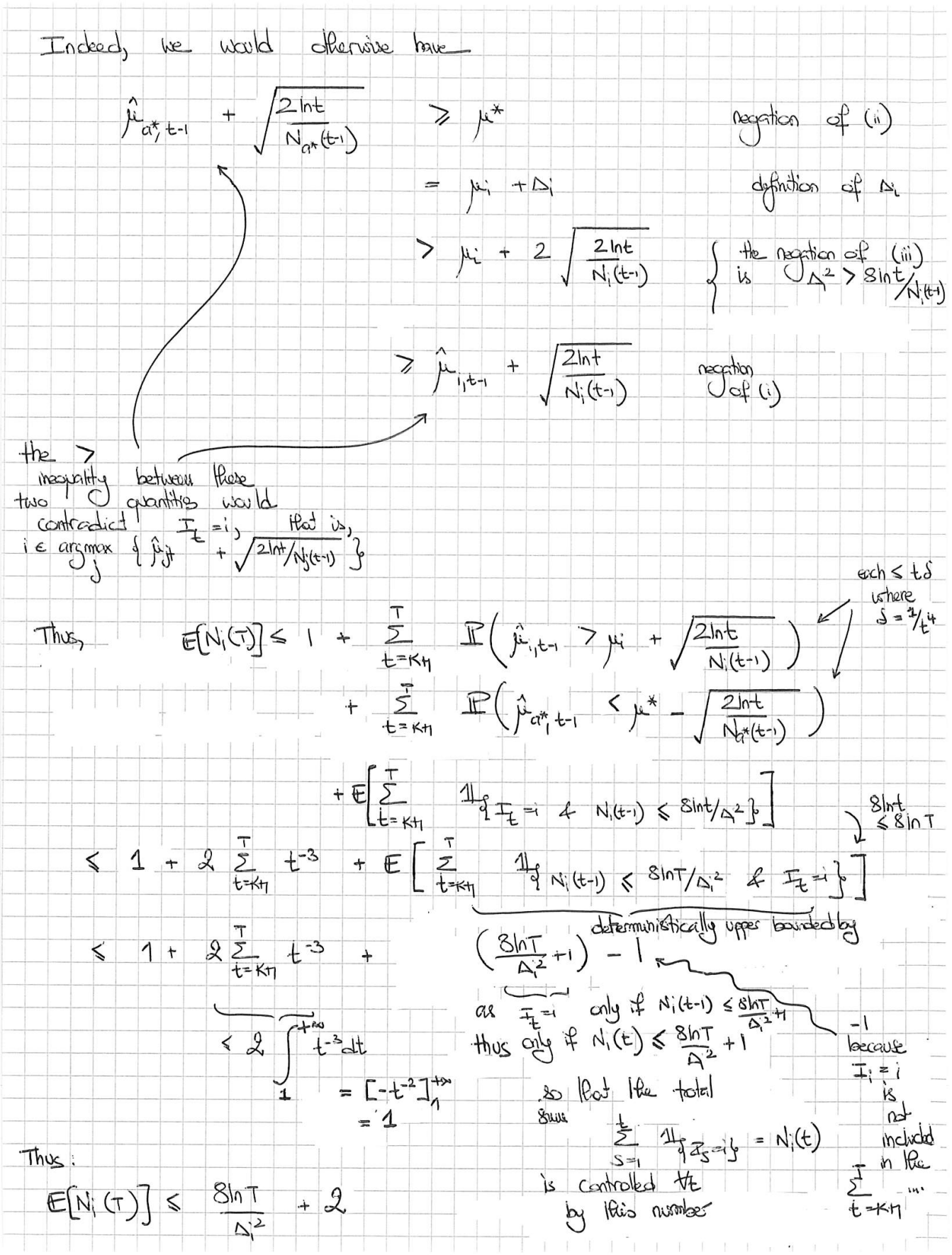
$$N_i(T) = 1 + \sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i\}}$$

We show below that $t \geq K+1$ and $I_t = i$ entails one of the following:

(i) $\hat{\mu}_{i,t-1} > \mu_i + \sqrt{\frac{2\ln t}{N_i(t-1)}}$ [$\mu_i <$ lower confidence bound]

(ii) $\hat{\mu}_{a^*,t-1} < \mu^* - \sqrt{\frac{2\ln t}{N_{i^*}(t-1)}}$ [$\mu^* >$ upper confidence bound]

(iii) $N_i(t-1) \leq \frac{8\ln t}{\Delta^2}$ [i not played often enough yet]



$$\hat{\mu}_{a^*,t-1} + \sqrt{\frac{2\ln t}{N_{a^*}(t-1)}} \geq \mu^*$$

$$= \mu_i + \Delta_i$$

$$> \mu_i + 2\sqrt{\frac{2\ln t}{N_i(t-1)}}$$

$$\geq \hat{\mu}_{i,t-1} + \sqrt{\frac{2\ln t}{N_i(t-1)}}$$

$$\mathbb{E}[N_i(T)] \leq 1 + \sum_{t=K+1}^{T} \mathbb{P}\left(\hat{\mu}_{i,t-1} > \mu_i + \sqrt{\frac{2\ln t}{N_i(t-1)}}\right) + \sum_{t=K+1}^{T} \mathbb{P}\left(\hat{\mu}_{a^*,t-1} < \mu^* - \sqrt{\frac{2\ln t}{N_{a^*}(t-1)}}\right) + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \text{ and } N_i(t-1) \leq 8\ln t/\Delta_i^2\}}\right]$$

$$\leq 1 + 2\sum_{t=K+1}^{T} t^{-3} + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}_{\{N_i(t-1) \leq 8\ln T/\Delta_i^2 \text{ and } I_t = i\}}\right]$$

$$\leq 1 + 2\sum_{t=K+1}^{T} t^{-3} + \left(\frac{8\ln T}{\Delta_i^2} + 1\right) - 1$$

$$\leq 2\int_1^{+\infty} t^{-3}\,dt = \left[-t^{-2}\right]_1^{+\infty} = 1$$

$$\mathbb{E}[N_i(T)] \leq \frac{8\ln T}{\Delta_i^2} + 2$$

Proof of the lemma (Hoeffding-Azuma inequality with a random number of summands):

We only start (we will finish it next week!)

$$Z_t = \sum_{s=1}^{t} (Y_s - \mu_a) \mathbb{1}_{\{I_s = a\}}$$

(0) $(Z_t)_{t\geq 0}$ is a martingale wrt. $(\mathcal{F}_t)_{t\geq 0} = (\sigma(Y_1, \dots Y_t))$ where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ trivial σ-algebra

Indeed: each I_t is $\mathcal{F}_{t-1}$-measurable (picked based only on past payoffs)

thus Z_t is $\mathcal{F}_t$-adapted

Showing that it is a martingale amounts to showing

$$\mathbb{E}[(Y_t - \mu_a) \mathbb{1}_{\{I_t = a\}} \mid Y_1 \dots Y_{t-1}] \overset{?}{=} 0 \quad \text{a.s.}$$

but since I_t is $\mathcal{F}_{t-1}$-measurable, this quantity equals

$$\mathbb{E}[(Y_t - \mu_a) \mathbb{1}_{\{I_t = a\}} \mid Y_1 \dots Y_{t-1}, I_t]$$

$$= (\mathbb{E}[Y_t \mid I_t, Y_1 \dots Y_{t-1}] - \mu_a) \mathbb{1}_{\{I_t = a\}}$$

(by the bandit model, Y_t is drawn independently at random given I_t; thus by the very bandit model, this conditional expectation equals μ_{I_t})

$$= (\mu_{I_t} - \mu_a) \mathbb{1}_{\{I_t = a\}} = 0 \quad \text{as as desired}$$

Exercise 2: To conclude the proof, show that

(1) For all $x \in \mathbb{R}$, $\left(M_t\right) = \left(\exp\left(x Z_t - \frac{x^2}{8} N_a(t)\right)\right)_{t\geq 0}$ is an adapted $(\mathcal{F}_t)_{t\geq 0}$ supermartingale

↳ in particular $\mathbb{E}[M_t] \leq 1$ for all t

(2) $\forall \varepsilon > 0$, $\forall \ell \geq 1$, $\mathbb{P}\{Z_t \geq \varepsilon \text{ and } N_a(t) = \ell\} \leq e^{-2\varepsilon^2/\ell}$

and then conclude.