Lecture 2 — Spring 2021 term Lecture 2 — Spring 2021 term

Convex functions & comparison to the best convex vector

↳ efficient forecaster (but worse bound...)

$\mathcal{X} = \{ (p_1 \dots p_N) : \sum_j p_j = 1, \ \forall k \ p_k \geq 0 \}$ is the simplex of convex weight vectors.

Setting (reminder): At each round $t = 1, 2, \dots T$:

1. The Statistician and the opponent pick simultaneously $p_t \in \mathcal{X}$ and a convex loss function $\ell_t : \mathcal{X} \to [m, M]$
2. ℓ_t and p_t are revealed

↳ Regret: $R_T = \sum_{t=1}^{T} \ell_t(p_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(p)$ to be controlled in a uniform way

Fact: Convex functions are subdifferentiable on the interior of their domain of definition:

Let $f : \mathcal{D} \to \mathbb{R}$ be convex, where $\mathcal{D} \subseteq \mathbb{R}^N$ is convex:

$\forall x \in \mathring{\mathcal{D}}, \quad \exists \partial f_x \in \mathbb{R}^N \ \Big| \quad \forall y \in \mathcal{D},$

$$f(x) - f(y) \leq \partial f_x \cdot (x - y)$$

$\mathcal{G}(x) = \{$ the set of all possible such $\partial f_x \}$ is called the subgradient of f at x

If f is differentiable at x, then $\mathcal{G}(x) = \{ \nabla f(x) \}$.

Ex: $f(x) = |x|$ $\quad \mathcal{G}(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ \{1\} & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$

Application: If $p_t \in \mathring{\mathcal{X}}$ (ie, $p_{j,t} > 0 \ \forall j$) then $\exists \, \partial \ell_t(p_t) \in \mathbb{R}^N \ \Big|$

$$\forall p \in \mathcal{X}, \quad \ell_t(p_t) - \ell_t(p) \leq \partial \ell_t(p_t) \cdot (p_t - p)$$

Example: Meta-statistical framework: $\ell_t(p) = \left(\sum_j p_j f_{j,t} - y_t \right)^2$

ℓ_t is differentiable, $\nabla \ell_t(p) = \left(2 \left(\sum_j p_j f_{j,t} - y_t \right) f_{j,t} \right)_j$

↳ Gradients in $[-G, G]^N$ with $G = 2B^2$ if $f_{j,t}, y_t \in [0, B]$

<u>Strategy:</u> Exponentiated Gradients (EG) with learning rate $\eta > 0$

$p_1 = (1/N, \dots, 1/N)$ and denoting $\tilde{\ell}_{js}$ = j-th component of $\partial\ell_s(p_s)$:

$$p_{jt} = \exp\Big(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{js}\Big) \Big/ \sum_{k=1}^{N} \exp\Big(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{ks}\Big)$$

interpreted as pseudo-loss

Strategy easy to implement as soon as subgradients can be easily computed.

<u>Theorem:</u> Assume that the ℓ_t are picked such that

$$\forall t \ \forall j \ \forall p \in \mathring{\mathcal{X}} \qquad (\partial\ell_t(p))_j \in [-G, G]$$

Then the regret of EG tuned with $\eta > 0$ is controled as

$$R_T = \sum_{t=1}^{T} \ell_t(p_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(p) \leq \frac{\ln N}{\eta} + \eta \frac{G^2}{2} T$$

In particular, the choice $\eta = \frac{1}{G}\sqrt{\frac{2 \ln N}{T}}$ leads to $R_T \leq G\sqrt{2T \ln N}$.

<u>Proof:</u> Fix a p. Since by construction (of exponential weights) $p_t \in \mathring{\mathcal{X}}$, we have by convexity

$$\ell_t(p_t) - \ell_t(p) \leq \partial\ell_t(p_t) \cdot (p_t - p) \overset{\text{not.}}{=} \underbrace{\sum_{j=1}^{N} p_{jt} \tilde{\ell}_{jt}} - \underbrace{\sum_{j=1}^{N} p_j \tilde{\ell}_{jt}}$$

$\longrightarrow$ Reduction to linear losses: let's pretend that the opponent picks vectors of losses

$$(\tilde{\ell}_{1t} \dots \tilde{\ell}_{Nt}) \in [-G, G]$$

By the above inequality,

$$R_T \leq \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \tilde{\ell}_{jt} - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \sum_{j=1}^{N} p_j \tilde{\ell}_{jt}$$

$$= \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \tilde{\ell}_{jt} - \min_{k} \sum_{t=1}^{T} \tilde{\ell}_{kt} \leq \frac{\ln N}{\eta} + \eta \frac{(2G)^2}{8} T$$

This concludes the proof.

Bound proved last week in the linear case

Application: Oracle inequalities → From individual sequences to stochastic sequences

Question: Let $(Y_1 \dots Y_T)$ be a sequence of iid random variables taking values in an arbitrary set $\mathcal{Y}$; and $Q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a bounded loss function, convex in its first argument, with bounded subgradients:

with either $Q \geq 0$ or $\forall \theta,\ Q(\theta, Y) \in \mathbb{L}^1$ for all expectations below to be defined

$$G = \sup_{y \in \mathcal{Y}} \sup_{p \in \mathcal{X}^\circ} \| \partial Q(p, y) \|_\infty < +\infty$$

Aim: Construct $\hat{\theta}_T = \hat{\theta}_T(Y_1 \dots Y_T)$ such that

$$\mathbb{E}\left[Q(\hat{\theta}_T, Y) \right] \leq \inf_{\theta \in \mathcal{X}} \mathbb{E}\left[Q(\theta, Y) \right] + \varepsilon_T$$

where Y is independent of the Y_t with the same distribution and $\varepsilon_T \searrow 0$.

(The expectation $\mathbb{E}$ is w.r.t. $Y_1 \dots Y_T$ and Y.)

Typical machine learning method: Empirical risk minimization, possibly in a regularized way:

$$\hat{\theta}_T \in \operatorname*{argmin}_{\theta \in \mathcal{X}} \left\{ \frac{1}{T} \sum_{t=1}^{T} Q(\theta, Y_t) + \lambda \operatorname{reg}(\theta) \right\}$$

with $\operatorname{reg}(\theta) = \|\theta\|_2$ or $\|\theta_1\|$ or...

and λ to be tuned (called the regularization factor)

Our method: (as in our proof of Sion's lemma)

1) Pretend data is sequential, while of course it is batch

$\tilde{\theta}_1 = (1/N, \dots 1/N)$ and by induction, with $\eta = \dfrac{1}{G}\sqrt{\dfrac{2 \ln N}{T}}$:

Strongly depends on the order picked

$$\tilde{\theta}_{j,t} = \exp\left(-\eta \sum_{s=1}^{t-1} \left(\partial Q(\tilde{\theta}_s, Y_s) \right)_j \right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\partial Q(\tilde{\theta}_s, Y_s) \right)_k \right)$$

2) Consider an average : $\hat{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \tilde{\theta}_t$

(note : we do not use y_T !)

Guarantee : $\varepsilon_T = G\sqrt{\frac{2\ln N}{T}}$

Proof : By the theorem on EG (exponentiated gradient) and given our choice of η :

$$\sum_{t=1}^{T} Q(\tilde{\theta}_t, y_t) - \inf_{\theta \in \mathcal{X}} \sum_{t=1}^{T} Q(\theta, y_t) \leqslant G\sqrt{2T\ln N}$$

Thus : $\forall \theta \in \mathcal{X}$, $\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, y_t) \leqslant \frac{1}{T}\sum_{t=1}^{T} Q(\theta, y_t) + \varepsilon_T$

with $\varepsilon_T = G\sqrt{\frac{2\ln N}{T}}$

Now, since $y_1 \dots y_T$ and y are iid :

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\theta, y_t)\right] = \mathbb{E}\left[Q(\theta, y)\right]$$

We conclude the proof by showing that $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, y_t)\right] \geqslant \mathbb{E}\left[Q(\hat{\theta}_T, y)\right]$

Indeed : $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, y_t)\right] = \mathbb{E}\left[\underbrace{\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, y)}_{\geqslant Q\left(\frac{1}{T}\sum_t \tilde{\theta}_t,\, y\right)}\right]$

$\tilde{\theta}_t$ only depends on $y_1 \dots y_{t-1}$ and $(y_1 \dots y_{t-1}, y_t)$ and $(y_1 \dots y_{t-1}, y)$ have the same distribution

by convexity and because the second argument is now the same for everyone !

We have shown: $\forall \theta$,

$$\mathbb{E}\left[Q(\hat{\theta}_T, y)\right] \leq \mathbb{E}\left[Q(\theta, y)\right] + \varepsilon_T$$

from which the conclusion follows.

Extension : Oracle inequality for stationary data.

A stationary sequence $(y_1, y_2, \dots)$ is by definition a sequence of random variables such that

$\forall k \geqslant 1$, $\forall t \geqslant 1$, $(y_1 \dots y_k)$ and $(y_{t+1}, \dots y_{t+k})$ have the same distribution.

 Lecture 2 — Spring 2021 term Lecture 2 — Spring 2021 term

The strategy is much more complicated to state and we need first a piece of notation

* For a sequence of observations $z_1, \dots z_t \in \mathcal{Y}$, EG outputs the following weights, which we re-define as functions of the information available:

$$\psi_1^{EG} = (1/N, \dots 1/N)$$

and for $t \geq 2$:

$$\psi_{jt}^{EG}(z_1, \dots z_{t-1}) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\partial Q(\psi_s^{EG}(z_1, \dots z_{s-1}), z_s)\right)_j\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\partial Q(\psi_s^{EG}(z_1 \dots z_{s-1}), z_s)\right)_k\right)}$$

j-th component of $\psi_t^{EG}(z_1 \dots z_{t-1}) \in \mathbb{R}^N$

! ψ_t^{EG} is an extremely complicated function of its arguments $z_1 \dots z_{t-1}$ (cf. the various calls to ψ_s^{EG}, with $s \leq t-1$, in its definition).

* We define $\tilde{\theta}_t = \psi_t^{EG}(y_{T+2-t}, \dots y_T)$ for $t \geq 2$

and $\tilde{\theta}_1 = (1/N, \dots 1/N)$

We recommend: $\hat{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \tilde{\theta}_t$

! Computing $\hat{\theta}_T$ is costly: the computation of each $\tilde{\theta}_t$ requires $t-1$ steps, namely the computation of the

$$\psi_s^{EG}(y_{T+2-t}, \dots y_{T-t+s}) \quad \text{for } 2 \leq s \leq t$$

no such intermediate quantity appears twice

$\Rightarrow$ about $\sum_{t=1}^{T} (t-1) = O(T^2/2)$ computation steps required.

 Lecture 2 — Spring 2021 term Lecture 2 — Spring 2021 term

What does EG guarantee?

$\forall \theta \in \mathcal{X}$, $$\frac{1}{T}\sum_{t=1}^{T} Q\big(\psi_t^{EG}(y_1,\dots y_{t-1}),\, y_t\big)$$ still $G\sqrt{(2\ln N)/T}$

$$\leqslant \frac{1}{T}\sum_{t=1}^{T} Q(\theta, y_t) \quad + \quad \varepsilon_T$$

We take expectations in both side to conclude:

- Stationarity implies in particular that all the y_t have the same distribution (but they are NOT independent in general); hence

$$\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T} Q(\theta, y_t)\Big] = \mathbb{E}\big[Q(\theta, y_{T+1})\big]$$

- it also implies that $\big(\psi_t^{EG}(y_1,\dots y_{t-1}),\, y_t\big)$

and $\big(\psi_t^{EG}(y_{T+2-t},\dots y_T),\, y_{T+1}\big) = \big(\tilde\theta_t,\, y_{T+1}\big)$ this defines $\tilde\theta_t$

have the same distribution, so that

$$\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T} Q\big(\psi_t^{EG}(y_1,\dots y_{t-1}),\, y_t\big)\Big]$$

$$= \mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde\theta_t, y_{T+1})\Big] \geqslant \mathbb{E}\big[Q(\hat\theta_T, y_{T+1})\big]$$

convexity of $\theta \mapsto Q(\theta, y_{T+1})$

- thus we proved, for our choice of $\hat\theta_T = \frac{1}{T}\sum_{t=1}^{T}\tilde\theta_t$

$$\mathbb{E}\big[Q(\hat\theta_T, y_{T+1})\big] \leqslant \inf_{\theta\in\mathcal{X}} \mathbb{E}\big[Q(\theta, y_{T+1})\big] + \varepsilon_T$$

same ε_T as for the iid case.

REMINDER: When does a limit in distribution entail a limit for the expectations?

(1) In general, $Y_T \xrightarrow{\mathcal{L}} Y$ does not imply $\mathbb{E}[f(Y_T)] \to \mathbb{E}[f(Y)]$ if f is just assumed continuous.

Ex: f = identity, $Y_T = T^2 \, \mathbb{1}_{\{U_T \in [0,1/T]\}}$ where $U_1 \dots U_T \dots$ iid $\sim \mathcal{U}_{[0,1]}$

We have $\begin{cases} Y_T \xrightarrow{\mathbb{P}} 0 \\ \mathbb{E}[Y_T] = T \longrightarrow +\infty \end{cases}$

(2) Definition of $\xrightarrow{\mathcal{L}}$ is for continuous and bounded functions:

Let $(Y_T)_{T \geq 1}$ be a sequence of $\mathbb{R}^N$-valued random variables

$Y_T \xrightarrow{\mathcal{L}} Y$ if by definition, for all continuous and bounded functions $f : \mathbb{R}^N \to \mathbb{R}$, we have

$$\mathbb{E}[f(Y_T)] \longrightarrow \mathbb{E}[f(Y)].$$

(3) Uniform asymptotic integrability [uai]: is a sufficient condition for (1) to hold

Def: $(Z_T)_{T \geq 1}$ is uai if

$$\lim_{L \to +\infty} \limsup_{T \to +\infty} \mathbb{E}\Big[\|Z_T\| \, \mathbb{1}_{\{\|Z_T\| > L\}} \Big] = 0$$

Lemma: If $(Y_T)_{T \geq 1}$ is a sequence of random vectors in $\mathbb{R}^N$ and $f : \mathbb{R}^N \to \mathbb{R}$ is a continuous function such that

- $Y_T \xrightarrow{\mathcal{L}} Y$
- $\big(f(Y_T)\big)_{T \geq 1}$ is uai,

then: a) $f(Y) \in \mathbb{L}^1$ and $f(Y_T) \in \mathbb{L}^1$ for all T sufficiently large

b) $\mathbb{E}[f(Y_T)] \longrightarrow \mathbb{E}[f(Y)]$

Proof: For a): by uai, $\exists L_0 \mid \limsup_{T \to +\infty} \mathbb{E}\Big[|f(Y_T)| \, \mathbb{1}_{\{|f(Y_T)| > L_0\}} \Big] < +\infty$

in particular, $\limsup_{T \to +\infty} \mathbb{E}\big[|f(Y_T)|\big] < +\infty$, thus, $\left(\mathbb{E}\left[\,|f(y_T)|\,\right]\right)_{T \geq T_0}$ is bounded for some T_0

By Skorokhod's theorem, we can assume with no loss of generality that

$f(y_T) \longrightarrow f(y)$ as

so that Fatou's lemma indicates $\mathbb{E}\left[\,|f(y)|\,\right] \leq \underbrace{\liminf_{T \to +\infty} \mathbb{E}\left[\,|f(y_T)|\,\right]}_{< +\infty}$

For b): We denote by $\varphi_L : x \in \mathbb{R} \longmapsto \max\{-L, \min\{x, L\}\}$ the clipping operator

$$\Big|\mathbb{E}[f(y)] - \mathbb{E}[f(y_T)]\Big| \leq \Big|\mathbb{E}[f(y)] - \mathbb{E}[\varphi_L(f(y))]\Big| + \Big|\mathbb{E}[\varphi_L(f(y))] - \mathbb{E}[\varphi_L(f(y_T))]\Big| + \Big|\mathbb{E}[\varphi_L(f(y_T))] - \mathbb{E}[f(y_T)]\Big|$$

$\forall L$: tends to 0 as $y_T \xrightarrow{\mathcal{L}} y$ and $\varphi_L \circ f$ is continuous and bounded

Thus,

$$\forall L, \quad \limsup_{T \to +\infty} \Big|\mathbb{E}[f(y)] - \mathbb{E}[f(y_T)]\Big| \leq \Big|\mathbb{E}[f(y)] - \mathbb{E}[\varphi_L(f(y))]\Big| + \limsup_{T \to +\infty} \Big|\mathbb{E}[\varphi_L(f(y_T))] - \mathbb{E}[f(y_T)]\Big|$$

Taking $\limsup_{L \to +\infty}$ in both sides (well, the left-hand side is independent of L) and using that $\limsup_{T \to +\infty} a_T + b_T \leq \limsup_{T \to +\infty} a_T + \limsup_{T \to +\infty} b_T$ for non-negative sequences (a_T) and (b_T):

$$\limsup_{T \to +\infty} \Big|\mathbb{E}[f(y)] - \mathbb{E}[f(y_T)]\Big| \leq \underbrace{\limsup_{L \to +\infty} \Big|\mathbb{E}[f(y)] - \mathbb{E}[\varphi_L(f(y))]\Big|}_{= 0 \text{ by dominated convergence}} + \underbrace{\limsup_{L \to +\infty} \limsup_{T \to +\infty} \Big|\mathbb{E}[f(y_T)] - \mathbb{E}[\varphi_L(f(y_T))]\Big|}_{\text{to be handled by the u.i. property}}$$

By definition of φ_L,

$$\mathbb{E}[f(y_T)] - \mathbb{E}[\varphi_L(f(y_T))]$$
$$= \mathbb{E}\left[(f(y_T)+L)\,\mathbb{1}_{\{f(y_T)<-L\}}\right] + \mathbb{E}\left[(f(y_T)-L)\,\mathbb{1}_{\{f(y_T)>L\}}\right]$$

In particular, $\big|\,\mathbb{E}[f(y_T)] - \mathbb{E}[\varphi_L(f(y_T))]\,\big|$

$$\leqslant \mathbb{E}\Big[\underbrace{|f(y_T)+L|}_{\leqslant |f(y_T)| \text{ on this event}}\,\mathbb{1}_{\{f(y)<-L\}}\Big] + \mathbb{E}\Big[\underbrace{|f(y_T)-L|}_{\leqslant |f(y_T)| \text{ on this event}}\,\mathbb{1}_{\{f(y_T)>L\}}\Big]$$

$$\leqslant \mathbb{E}\left[|f(y_T)|\,\mathbb{1}_{\{|f(y_T)|>L\}}\right]$$

whose $\limsup\limits_{L\to+\infty} \limsup\limits_{T\to+\infty} \ldots$ equals 0 by the very definition of uai for $(f(y_T))_{T\geqslant 1}$

This concludes the proof.

(4) A useful sufficient condition for uai:

Lemma: If $(Z_T)_{T\geqslant 1}$ is bounded in $\mathbb{L}^p$ for some $p>1$,

ie, $\sup\limits_{T\geqslant 1} \mathbb{E}[\|Z_T\|^p] \overset{\text{not.}}{=} B < +\infty$.

Then $(Z_T)_{T\geqslant 1}$ is uai.

Proof: $\varphi(x) = x^p$: as $p>1$, $\varphi(x) \gg x$

thus, $\forall M>0,\ \exists L \mid \forall x\geqslant L,\ \varphi(x) \geqslant Mx$

Taking $M=B\gamma$: $\exists L_\gamma \mid \forall T\geqslant 1,\ B \geqslant \mathbb{E}[\|Z_T\|^p] \geqslant \mathbb{E}\left[\|Z_T\|^p\,\mathbb{1}_{\{\|Z_T\|>L_\gamma\}}\right]$

$$\geqslant B\gamma\,\mathbb{E}\left[\|Z_T\|\,\mathbb{1}_{\{\|Z_T\|>L_\gamma\}}\right]$$

Thus, for this L: $\sup\limits_{T\geqslant 1} \mathbb{E}\left[\|Z_T\|\,\mathbb{1}_{\{\|Z_T\|>L_\gamma\}}\right] \leqslant \dfrac{1}{\gamma}$

and the uai property follows by letting $\gamma\to+\infty$.
(even something stronger than that follows)

OPTIMALITY OF THE $\sqrt{\frac{T}{2}\ln N}$ BOUND: Asymptotic lower bound

- In the case of linear losses
- By homogeneity we may assume that $\ell_{jt} \in [0,1]$ for all t and j (i.e., $m = 0$ and $M = 1$)

Theorem:
$$\liminf_{N \to +\infty} \ \liminf_{T \to +\infty} \ \inf_{\text{algorithms}} \ \sup_{\ell_{jt} \in [0,1]} \frac{\sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_t \ell_{kt}}{\sqrt{\frac{T}{2}\ln N}} \geq 1$$

Stronger to have an infimum here than to prove $\forall$ algo, $\liminf \liminf \sup \ldots \geq 1$

The opponent does not need to react, it suffices to consider fixed-in-advance sequences (individual sequences)

Proof: We lower bound the $\sup_{\ell_{jt} \in [0,1]}$ by an $\mathbb{E}[\ \ ]$

with losses L_{jt} drawn iid $\sim \mathrm{Ber}(1/2)$

Using that for any algorithm:

$$\forall t \geq 1, \quad \mathbb{E}\Big[\sum_j p_{jt} L_{jt}\Big] = \sum_j \mathbb{E}[p_{jt}]\, \underbrace{\mathbb{E}[L_{jt}]}_{=1/2} = \frac{1}{2}\,\mathbb{E}\Big[\underbrace{\sum_j p_{jt}}_{=1 \text{ a.s.}}\Big] = \frac{1}{2}$$

p_{jt} is measurable wrt past losses L_{ks}, $s \leq t-1$ and is thus independent of the L_{jt}

↳ Not so surprising: means that we cannot predict efficiently fair coin tosses

* But * by a crowd effect (due to the central limit theorem) we will show that

$$\mathbb{E}\Big[\min_k \sum_{t=1}^{T} L_{kt}\Big] = T/2 - \text{something of order } \sqrt{T \ln N}$$

even though we "individually" have $\forall k, \ \mathbb{E}\Big[\sum_{t=1}^{T} L_{kt}\Big] = T/2$

More formally:

$$\inf_{\text{algo}} \sup_{\ell_{jt} \in [0,1]} \frac{\sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_t \ell_{kt}}{\sqrt{\frac{T}{2}\ln N}} \geq \inf_{\text{algo}} \frac{\mathbb{E}\Big[\sum_{t,j} p_{jt} L_{jt}\Big] - \mathbb{E}\Big[\min_k \sum_{t=1}^{T} L_{kt}\Big]}{\sqrt{(T/2)\ln N}}$$

$$= \frac{1}{\sqrt{2\ln N}}\, \mathbb{E}\left[\max_{k=1\ldots N} \frac{\sum_{t=1}^{T} (1/2 - L_{kt})}{\frac{1}{2}\sqrt{T}}\right]$$

 Lecture 2 — Spring 2021 term Lecture 2 — Spring 2021 term

We denote
$$Z_{kT} = \frac{\sum_{t=1}^{T} \left(\tfrac{1}{2} - L_{kt}\right)}{\tfrac{1}{2}\sqrt{T}} \qquad \begin{array}{l} \longleftarrow \text{ sum of iid random var} \\ \longleftarrow \ \sigma\sqrt{T} \end{array}$$

and note that Z_{kT} is a central-limit-theorem statistic: $\tfrac{1}{2} - L_{kt}$ are iid, centered with $\mathrm{Var}\left(\tfrac{1}{2} - L_{kt}\right) = 1/4$

For each k, $\quad Z_{kT} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

By independence,
$$Z_T = \begin{pmatrix} Z_{1T} \\ \vdots \\ Z_{NT} \end{pmatrix} \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0, I_N)$$

Take $f_N(x_1, \ldots x_N) = \max_{k \leq N} x_k$; we have shown so far:
$$\inf_{\text{algo}} \ \sup_{\ell_{jt} \in [0,1]} \frac{\sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_t \ell_{kt}}{\sqrt{\frac{T}{2}\ln N}} \geq \frac{1}{\sqrt{2 \ln N}} \, \mathbb{E}\left[f_N(Z_T) \right]$$

The $f_N(Z_T)$ are bounded by N in $\mathbb{L}^2$-norm:
$$\mathbb{E}\left[f_N(Z_T)^2 \right] = \mathbb{E}\left[\max_{k \leq N} Z_{kT}^2 \right] \leq \sum_{k=1}^{N} \underbrace{\mathbb{E}\left[Z_{kT}^2 \right]}_{\text{each } = 1,} = N$$
cf. Z_{kT} central-limit-theorem statistic

Thus they form a uai sequence and we have
$$\lim_{T \to +\infty} \frac{1}{\sqrt{2 \ln N}} \, \mathbb{E}\left[f_N(Z_T) \right] = \frac{1}{\sqrt{2 \ln N}} \, \mathbb{E}\left[\max_{k \leq N} Z_k \right]$$

So far:
$$\liminf_{T \to +\infty} \ \inf_{\text{algo}} \ \sup_{\ell_{jt} \in [0,1]} \frac{\sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_t \ell_{kt}}{\sqrt{(T/2) \ln N}}$$
$$\geq \underbrace{\frac{1}{\sqrt{2 \ln N}} \, \mathbb{E}\left[\max_{k \leq N} Z_k \right]}$$

The proof is concluded by: $\qquad \liminf_{N \to +\infty} \ \ldots \ \geq 1$

↑ details on the next pages $\qquad$ (this is even a true limit)

Asymptotics for the expectation of the maximum of standard Gaussian random variables

We prove that for $G_1\, G_2 \dots$ iid $\sim \mathcal{N}(0,1)$,

$$\mathbb{E}\Big[\max_{k \le N} G_k\Big] \sim \sqrt{2 \ln N}$$

(1) Easy part (useless for the main result proved above):

$$\mathbb{E}\Big[\max_{k \le N} G_k\Big] \leqslant \sqrt{2 \ln N}$$

as a special case of the following lemma.

Lemma: If $V_1\, V_2 \dots V_N$ are such that $\mathbb{E}\big[e^{\lambda V_j}\big] \leqslant e^{\lambda^2 \sigma^2/2}$ $\forall j$ $\forall \lambda > 0$

then $\mathbb{E}\Big[\max_{k \le N} V_k\Big] \leqslant \sqrt{2\sigma^2 \ln N}$

Notes:
- Valid even without independence (independence will be needed for the lower bound)
- If $G_k \sim \mathcal{N}(0,\sigma^2)$ then $\mathbb{E}\big[e^{\lambda G_k}\big] = e^{\lambda^2 \sigma^2/2}$
- If $G_k \in [m, M]$ with $\mathbb{E}[G_k] = 0$ then by Hoeffding's lemma $\mathbb{E}\big[e^{\lambda G_k}\big] \leqslant e^{\lambda^2 (M-m)^2/8}$
- We call sub-Gaussian the random variables V_k satisfying the assumptions of the lemma.

Proof: Proof technique called "Pisier's argument":

$$\mathbb{E}\Big[\max_{k \le N} V_k\Big] \underset{\text{Jensen's inequality}}{\leqslant} \frac{1}{\lambda} \ln \mathbb{E}\Big[\exp\Big(\lambda \max_{k \le N} V_k\Big)\Big] = \frac{1}{\lambda} \ln \mathbb{E}\Big[\max_{k \le N} e^{\lambda V_k}\Big]$$

$$\underset{\text{max of non-neg quantities} \leqslant \text{their sum}}{\leqslant} \frac{1}{\lambda} \ln\Big(\sum_{k \le N} \mathbb{E}\big[e^{\lambda V_k}\big]\Big)$$

$$\underset{\text{by the sub-Gaussian assumption}}{\leqslant} \frac{1}{\lambda} \ln\Big(N e^{\lambda^2 \sigma^2/2}\Big) = \frac{\ln N}{\lambda} + \lambda \frac{\sigma^2}{2}$$

$$= \sqrt{2\sigma^2 \ln N}$$

Choose $\lambda^* = \sqrt{\dfrac{2 \ln N}{\sigma^2}}$

(2) "Difficult" part (and the one that we need!):

$$\liminf_{N\to+\infty} \frac{\mathbb{E}\big[\max_{k\leq N} G_k\big]}{\sqrt{2\ln N}} \geq 1$$

Credits: Pascal Massart + Mathieu Even, M2 student following this course in 2019 – 20

$$M_N \overset{\text{def.}}{=} \left(\max_{k\leq N} G_k \Big/ \sqrt{2\ln N}\right)^+ \longleftarrow \text{non-negative part}$$

Given that $0 \geq \mathbb{E}\Big[\big(\max_{k\leq N} G_k\big)\,\mathbb{1}_{\{\max_{k\leq N} G_k \leq 0\}}\Big]$

$$\geq \mathbb{E}\Big[G_1\,\mathbb{1}_{\{\max_{k\leq N} G_k \leq 0\}}\Big] \geq \mathbb{E}\big[G_1\,\mathbb{1}_{\{G_1\leq 0\}}\big]$$

$$\underset{\text{(Cauchy-Schwarz)}}{\geq} -\sqrt{\mathbb{E}[G_1^2]\;\mathbb{P}\{G_1\leq 0\}} = -1/\sqrt{2}.$$

we have

$$\liminf_{N\to+\infty} \frac{\mathbb{E}\big[\max_{k\leq N} G_k\big]}{\sqrt{2\ln N}} = \liminf_{N\to+\infty} \mathbb{E}[M_N]$$

By Fatou's lemma, it thus suffices to show that $\liminf_{N\to+\infty} M_N \geq 1$ a.s.

For $\varepsilon\in(0,1)$:

$$\mathbb{P}\big(M_N \leq \sqrt{1-\varepsilon}\big) \overset{\substack{\text{cf. positive part}\\ \text{in the def of } M_N\\ \downarrow}}{=} \mathbb{P}\Big(\forall k\leq N,\quad G_k \leq \sqrt{2(1-\varepsilon)\ln N}\Big)$$

$$= \Big(\Phi\big(\sqrt{2(1-\varepsilon)\ln N}\big)\Big)^N \underset{1+u\leq e^u}{\leq} \exp\big(-N(1-\Phi(\dots))\big)$$

where Φ is the cdf of the $\mathcal{N}(0,1)$ distribution.

As $1-\Phi(x) \underset{x\to+\infty}{\sim} e^{-x^2/2}\big/ x\sqrt{2\pi}$ (proof by two integration by parts, see below)

we have

$$N\Big(1-\Phi\big(\sqrt{2(1-\varepsilon)\ln N}\big)\Big) \sim N\exp\big(-(1-\varepsilon)\ln N\big) \Big/ 2\sqrt{\pi(1-\varepsilon)\ln N}$$

$$= N^{\varepsilon}\Big/ 2\sqrt{\pi(1-\varepsilon)\ln N}$$

Thus $\sum_N \exp\big(-N(1-\Phi(\;))\big)$ is convergent

therefore, $\sum_N \mathbb{P}\big(M_N\leq\sqrt{1-\varepsilon}\big)$ is also convergent

 Lecture 2 — Spring 2021 term Lecture 2 — Spring 2021 term

By the Borel-Cantelli lemma: $\{ M_N \leqslant \sqrt{1-\varepsilon} \}$ a.s. takes place only finitely many times

i.e. $\liminf\limits_{N\to\infty} M_N \geqslant \sqrt{1-\varepsilon}$ a.s.

Since this is true for all $\varepsilon>0$, we get the desired $\liminf\limits_{N\to\infty} M_N \geqslant 1$ a.s.
(with $\varepsilon_n = 1/n \to 0$)

(3) Proof of $1-\Phi(x) \underset{x\to+\infty}{\sim} e^{-x^2/2} \big/ x\sqrt{2\pi}$

First IbP: (integration by parts)

$$1-\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-t^2/2}\,dt = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} t\, e^{-t^2/2}\, \frac{1}{t}\,dt$$

$$= \underbrace{\left[-\frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t} \right]_x^{+\infty}}_{= e^{-x^2/2} / x\sqrt{2\pi}} - \underbrace{\frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \frac{e^{-t^2/2}}{t^2}\,dt}$$

Second IbP: $$0 \leqslant \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \frac{e^{-t^2/2}}{t^2}\,dt = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} t\, e^{-t^2/2}\, \frac{1}{t^3}\,dt$$

$$= \left[-\frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t^3} \right]_x^{+\infty} - \int_x^{+\infty} \underbrace{\qquad}_{\geqslant 0}\,dt$$

$$\leqslant \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x^3}$$

So that the $e^{-x^2/2} \big/ x\sqrt{2\pi}$ was the dominating term in what the first IbP yielded, hence the result.

Exercice — **Finite-time lower bound.** (also known as: **Non-asymptotic lower bound**)

It's a difficult exercise, try to do as much as possible!
Part of the difficulty comes from the manipulation of Kullback-Leibler divergences, we will study their properties in detail in a few weeks, when studying the lower bounds for stochastic bandits.

Let $\mathbb{P}_0$ be a probability such that the losses L_{kt} are iid $\sim \mathcal{B}(1/2)$ $\forall k,t$

For $j \in \{1, \dots N\}$ let $\mathbb{P}_j$ be a probability such that:

- all losses L_{kt} are independent
- $L_{kt} \sim \mathrm{Ber}(1/2)$ $\forall t, \quad \forall k \neq j$
- $L_{jt} \sim \mathrm{Ber}(1/2 - \varepsilon)$ $\forall t$

(1) Show that

$$\sup_{\ell_{jt} \in [0,1]} \left\{ \sum_{j,t} p_{jt} \ell_{jt} - \min_k \sum_{t=1}^{T} \ell_{kt} \right\} \geq T\varepsilon \max_{k=1,\dots N} \left\{ 1 - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_k [p_{kt}] \right\}$$

REMINDER or **QUICK INTRODUCTION TO A NEW OBJECT:**

Recall that the Kullback-Leibler divergence between two probability distributions μ and ν equals:

$$KL(\nu, \mu) = \begin{cases} +\infty & \text{if } \nu \text{ not absolutely continuous wrt } \mu \\ \int \left(\frac{d\nu}{d\mu} \ln \frac{d\nu}{d\mu} \right) d\mu \quad \in [0, +\infty] \end{cases}$$

(KL well-defined as $x \mapsto x \ln x$ is $\geq -1/e$ over $(0, +\infty)$)

$$= \int \left(\ln \frac{d\nu}{d\mu} \right) d\nu \quad \text{if } \nu \ll \mu$$

KL has the following properties:

- $KL(\nu \otimes \nu', \mu \otimes \mu') = KL(\nu, \mu) + KL(\nu', \mu')$ ⟵ by Fubini-Tonelli
- $\forall$ random variable X, $KL(\nu^X, \mu^X) \leq KL(\nu, \mu)$ ⟵ data-processing inequality (we take it for granted and will prove it later)

We denote by

$$kl(p,q) = KL(\mathrm{Ber}(p), \mathrm{Ber}(q)) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$$

the KL-divergence between two Bernoulli distributions.

 Lecture 2 — Spring 2021 term Lecture 2 — Spring 2021 term

(2) Prove FANO's lemma as revisited by Lucien Birgé:

let $(\Omega, \mathcal{F})$ be a measurable space

let $Q_1, \dots Q_N$ be probability distributions over Ω

let $(A_1 \dots A_N)$ be a partition of Ω

then

$$\min_{j=1\dots N} Q_j(A_j) \leqslant \max\left\{ \frac{2e}{2e+1}, \frac{\overline{K}}{\ln N} \right\}$$

where $\overline{K} = \dfrac{1}{N-1} \displaystyle\sum_{j=2}^{N} KL(Q_j, Q_1)$

Hints: Consider $p = \dfrac{1}{N-1} \displaystyle\sum_{j=2}^{N} Q_j(A_j)$ and $q = \dfrac{1}{N-1} \displaystyle\sum_{j=2}^{N} Q_1(A_j)$

and denote $a = \min_{j=1..N} Q_j(A_j)$. Show that:

- $kl(p,q) \leqslant \overline{K}$ / to that end, first prove that KL is jointly convex (it is a consequence of the data-processing inequality, explain why)
- $kl(p,q) / \ln N \geqslant a$ for $a \geqslant 2e/2e+1$

Hint: Show that $kl(\mathbb{E}_{\mathbb{P}} Z, \mathbb{E}_{\mathbb{Q}} Z) \leqslant KL(\mathbb{P}, \mathbb{Q})$ for any random variable Z on $(\Omega, \mathcal{F})$ equipped with $\mathbb{P}$ or $\mathbb{Q}$

(3) Extend FANO's lemma to random variables so as to show

$$\min_{k=1,\dots N} \mathbb{E}_k\left[\frac{1}{T} \sum_{t=1}^{T} p_{kt} \right] \leqslant \max\left\{ \frac{2e}{2e+1}, \frac{\overline{K}'}{\ln N} \right\}$$

where $\overline{K}' = \dfrac{1}{N-1} \displaystyle\sum_{j=2}^{N} KL(\mathbb{P}_j^L, \mathbb{P}_1^L)$ with $L = (L_{kt})_{\substack{t \leq T \\ k \leq N}}$.

(4) Show that $\overline{K}' \leqslant 5T\varepsilon^2$ for $\varepsilon \leqslant 1/10$.

(5) Conclude to the following theorem:

Theorem: For all strategies, for all $N \geqslant 2$ and $T \geqslant 17 \ln N$,

$$\sup R_T \geqslant 0.06 \sqrt{T \ln N}$$