Fundamental Concentration Inequalities

Guillaume Obozinski

Swiss Data Science Center
EPFL & ETH Zürich

RLSS, juillet 2019, Lille

What and why?

- **What is the concentration of measure phenomenon?**
  This refers to the phenomenon that there are certain ways to combine random variables that produce r.v. that are *concentrated* around their expectation. One of the main case of interest are averages of independent variables.
- **Why do we need it for reinforcement learning?**
  RL require to make decisions in the presence of "uncertain uncertainty", r.v.s whose distributions are not known initially. This requires to be able to produce confidence intervals (or confidence regions) for these r.v. in the environment that are not yet know, but that are typically being learned in the RL algorithm.
- **Why is the central limit theorem not sufficient?**
  The CLT only produces asymptotic CIs with an error which is a priori not quantified.

Union bound

Let $A_1, A_2, \ldots, A_k$ be events. We have

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_k) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_k)$$

Proof.

$$\mathbb{E}\left[1_{A_1 \cup A_2 \cup \ldots \cup A_k}\right] \leq \mathbb{E}\left[1_{A_1} + 1_{A_2} + \ldots + 1_{A_k}\right]. \qquad \square$$

Example

Let $X_t \sim \mathcal{N}(0, \sigma^2)$ (not necessarily independent)

$$\mathbb{P}(\max_t X_t > x) = \mathbb{P}\big(\bigcup_t \{X_t > x\}\big) \leq \sum_{t=1}^{T} \mathbb{P}(X_t > x) \leq T \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

So with probability $1 - \delta$, we have

$$X \leq \sigma \sqrt{2 \log \frac{T}{\delta}}$$

Markov, Chebychev and Chernoff

Markov inequality

If $X \geq 0$ a.s. and $t > 0$, then $$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}$$

Chebychev inequality

$$\forall t > 0, \qquad \mathbb{P}(X - \mathbb{E}[X] > t) \leq \frac{\mathrm{Var}(X)}{t^2}$$

Chernoff inequality

$$\mathbb{P}(X > t) \leq \inf_{r \geq 0} e^{-rt}\, \mathbb{E}\left[e^{rX}\right]$$

Note that $r \mapsto \mathbb{E}\left[e^{rX}\right]$ is the *moment generating function* (MGF) of X.

Cramér-Chernoff Method

$\forall r > 0,$

$$\mathbb{P}(X > t) \le e^{-rt}\,\mathbb{E}\left[e^{rX}\right] = \exp\left(\psi_X(r) - rt\right) \quad \text{for} \quad \psi_X(r) = \log \mathbb{E}\left[e^{rX}\right]$$

ψ is the log MGF of X, aka *cumulant generating function* if $\mathbb{E}[X] = 0$.

Since this true for all $r \ge 0$ if

$$\psi_X^*(t) = \sup_{r \ge 0} rt - \psi_X(r),$$

then we have

$$\mathbb{P}(X > t) \le \exp(-\psi_X^*(t))$$

- ψ_X^* is called the Cramér transform of X
- If $t \ge \mathbb{E}[X]$, then $\psi_X^*(t) = \sup_{r \in \mathbb{R}} rt - \psi_X(r)$,

  i.e., ψ_X^* is the Fenchel-Legendre conjugate of ψ_X.

Applying the Cramér-Chernoff to the Gaussian

Let $X \sim \mathcal{N}(0, \sigma^2)$, then

$$\mathbb{E}[e^{rX}] = e^{\frac{r^2\sigma^2}{2}}, \qquad \psi(r) = \frac{r^2\sigma^2}{2}, \qquad \psi^*(t) = \frac{t^2}{2\sigma^2},$$

So that $$1 - \Phi(t) := \mathbb{P}(X > t) \le e^{-\psi^*(t)} = e^{-\frac{t^2}{2\sigma^2}}.$$

But it is well-known that for all $t > 0$,

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} \le 1 - \Phi(t) \le \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}}.$$

In fact

$$\sup_{t \ge 0} \left(1 - \Phi(t)\right) e^{\frac{t^2}{2\sigma^2}} = \frac{1}{2}.$$

So the Cramér-Chernoff produces a relatively good bound.

MGF inequality for bounded r.v.

Bernoulli r.v. X

For $X_B \sim \mathrm{Ber}(\theta)$, we have $\mathbb{E}\left[e^{sX_B}\right] = 1 - \theta + \theta e^s$

Any bounded r.v. X on $[0,1]$

If $\mathbb{E}[X] = \theta$, $\forall s \in \mathbb{R}$, we have $\mathbb{E}\left[e^{sX}\right] \leq \mathbb{E}\left[(1-X) + Xe^s\right] = 1 - \theta + \theta e^s$

So

$$\mathbb{E}\left[e^{sX}\right] \leq \mathbb{E}\left[e^{sX_B}\right] = 1 - \theta + \theta e^s.$$

And

$$\mathbb{E}\left[e^{s(X-\theta)}\right] \leq \mathbb{E}\left[e^{s(X_B-\theta)}\right] = \left(1 - \theta + \theta e^s\right)e^{-s\theta} = e^{\phi(s)},$$

with

$$\phi(s) := \log\left(1 - \theta + \theta e^s\right) - s\theta.$$

Key inequality (Hoeffding's Lemma)

Let $\phi(s) := \log\left(1-\theta+\theta e^s\right) - s\theta$. We have $\boxed{\phi(s) \leq \frac{s^2}{8}}$.

Proof.

By Taylor-Lagrange $\phi(s) = \phi(0) + s\phi'(0) + \frac{s^2}{2}\phi''(t)$ with $t \in (0, s)$.

$$\phi'(t) + \theta = \frac{\theta e^t}{1-\theta+\theta e^t} = \frac{1}{1+\alpha e^{-t}} \quad \text{with} \quad \alpha = \tfrac{1-\theta}{\theta}.$$

$$\phi''(t) = \frac{\alpha e^{-t}}{(1+\alpha e^{-t})^2} = \phi'(t)\big(1-\phi'(t)\big) \leq \frac{1}{4} \quad \text{since} \quad \phi'(t) \leq 1.$$

So $\phi(0) = 0, \phi'(0) = 0$ and, by T.-L.,

$$\phi(s) \leq \frac{s^2}{2}\phi''(t) \leq \frac{s^2}{8}.$$

$\square$

Bounded r.v. are sub-Bernoulli and thus sub-Gaussian

Let

- X be a r.v. on $[0,1]$ with $\mathbb{E}[X] = \theta$
- $X_B \sim \mathrm{Ber}(\theta)$
- $X_G \sim \mathcal{N}(0, \frac{1}{4})$
- $\phi(s) := \log\left(1 - \theta + \theta e^s\right) - s\theta.$

Then $\mathbb{E}\left[e^{s(X-\theta)}\right] \leq \mathbb{E}\left[e^{s(X_B-\theta)}\right] = e^{\phi(s)} \leq e^{\frac{s^2}{8}} = \mathbb{E}\left[e^{sX_G}\right].$

Bounded r.v. are sub-Bernoulli and thus sub-Gaussian

Let

- X be a r.v. on $[0,1]$ with $\mathbb{E}[X] = \theta$
- $X_B \sim \mathrm{Ber}(\theta)$
- $X_G \sim \mathcal{N}(0, \frac{1}{4})$
- $\phi(s) := \log\big(1 - \theta + \theta e^s\big) - s\theta.$

Then $\forall s \geq 0$, $\mathbb{E}\big[e^{s(X-\theta)}\big] \leq \mathbb{E}\big[e^{s(X_B-\theta)}\big] = e^{\phi(s)} \leq e^{\frac{s^2}{8}} = \mathbb{E}\big[e^{sX_G}\big]$.

Now, let

- Y be a random variable on the interval $[a,b]$
- $X := \dfrac{Y-a}{b-a} \in [0,1]$ so that $Y = (b-a)X + a$.
- $\widetilde{Y} = Y - \mathbb{E}[Y], \quad \widetilde{X} = X - \mathbb{E}[X], \quad \widetilde{X}_B = X_B - \mathbb{E}[X_B],$

We have $\widetilde{Y} = (b-a)\,\widetilde{X}$ and

$$\mathbb{E}\big[e^{s\widetilde{Y}}\big] = \mathbb{E}\big[e^{s(b-a)\widetilde{X}}\big] \leq \mathbb{E}\big[e^{s(b-a)\widetilde{X}_B}\big] = e^{\phi(s(b-a))} \leq e^{\frac{s^2(b-a)^2}{8}}.$$

Hoeffding inequality

Let X_i be *independent* bounded r.v. such that

- $\mathbb{E}[X_i] = 0$ and X_i has support in $[a_i, b_i]$.

Let $\tau^2 := \frac{1}{n}\sum_i \tau_i^2$ with $\tau_i^2 := \frac{1}{4}(b_i - a_i)^2$. Note that $\mathrm{Var}(X_i) \leq \tau_i^2$.

Then $\forall x \geq 0$, $\quad \mathbb{P}\Big(\overline{X} \geq x\Big) \leq \exp\left(-\frac{nx^2}{2\tau^2}\right) \quad$ with $\quad \overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$.

Proof.
$$
\begin{aligned}
\mathbb{P}\left(\sum_i X_i \geq nx\right) &= \mathbb{P}\Big(\exp\left(s\sum_i X_i\right) \geq \exp(snx)\Big) \\
&\leq e^{-snx}\,\mathbb{E}\Big[\prod_i e^{sX_i}\Big] = e^{-snx}\prod_i \mathbb{E}\big[e^{sX_i}\big] \\
&\overset{\text{sub-G}}{\leq} \exp\left(-snx + \tfrac{s^2}{8}\sum_i (b_i - a_i)^2\right) \\
&= \exp\left(-snx + \tfrac{s^2}{2} n\tau^2\right)
\end{aligned}
$$

Thus $\quad \mathbb{P}\left(\sum_i X_i \geq nx\right) \leq \exp\left(-\frac{nx^2}{2\tau^2}\right) \quad$ by setting $\quad s = \frac{x}{n\tau^2} \geq 0$

which minimizes the RHS w.r.t. s. □

Comparing Hoeffding with the CLT

Let X_i be *independent* bounded r.v. such that

- $\mathbb{E}[X_i] = 0$ and X_i has support in $[a_i, b_i]$.
- Let $\tau^2 := \frac{1}{n}\sum_i \tau_i^2$ with $\tau_i^2 := \frac{1}{4}(b_i - a_i)^2$.
- Let $\sigma^2 := \frac{1}{n}\sum_i \sigma_i^2$ with $\sigma_i^2 = \mathrm{Var}(X_i) \leq \tau_i^2$.

By the CLT:

$$\sqrt{n}\,\overline{X} \xrightarrow{(d)} X^* \quad \text{with} \quad X^* \sim \mathcal{N}(0, \sigma^2)$$

We can compare:

Hoeffding: $\mathbb{P}\left(\sqrt{n}\,\overline{X} > x\right) \leq \exp\left(-\frac{x^2}{2\tau^2}\right)$

CLT: $\mathbb{P}\left(\sqrt{n}\,\overline{X} \geq x\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(X^* \geq x\right) \leq \frac{1}{x}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$

High probability statement of Hoeffding's inequality

As before let $\tau^2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2$.

Hoeffding inequality

$$\mathbb{P}\Big(\overline{X} > x\Big) \leq \exp\Big(-\frac{nx^2}{2\tau^2}\Big)$$

By setting the RHS to δ, we obtain the following reformulation.

High probability statement:

With probability $1-\delta$, $\quad \overline{X} \leq \sqrt{\frac{\tau^2}{n}\cdot 2\log\Big(\frac{1}{\delta}\Big)}.$

Or equivalently $\quad \sum_{i=1}^{n} X_i \leq \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}\sqrt{2\log\Big(\frac{1}{\delta}\Big)}.$

Sharper than Hoeffding: the Chernoff-Hoeffding inequality

If X_i are independent r.v. on [0,1] with $\mathbb{E}[X_i] = \theta_i$, then

$$\mathbb{P}\big(\tfrac{1}{n}\textstyle\sum_i X_i \geq q\big) \leq \exp\big(-n\,\mathrm{KL}(q\|\theta)\big)$$

$$\text{with} \quad \mathrm{KL}(q\|\theta) = q\log\tfrac{q}{\theta} + (1-q)\log\tfrac{1-q}{1-\theta}.$$

Proof

$$\begin{aligned}
\mathbb{P}\big(\textstyle\sum_i X_i \geq nq\big) &= \mathbb{P}\Big(\exp\big(s\textstyle\sum_i X_i\big) \geq \exp(snq)\Big)\\
&\leq e^{-snq}\,\mathbb{E}\Big[\textstyle\prod_i e^{sX_i}\Big] = e^{-snq}\textstyle\prod_i \mathbb{E}\big[e^{sX_i}\big]\\
&= e^{-snq}\textstyle\prod_i \big(1-\theta_i+\theta_i e^s\big)\\
&\leq e^{-snq}\ \big(1-\theta+\theta e^s\big)^n \quad \text{with} \quad \theta = \tfrac{1}{n}\textstyle\sum \theta_i,
\end{aligned}$$

by the arithmetico-geometric inequality.

Let $\psi(s) = n\log\big(1-\theta+\theta e^s\big)$. Then $\psi'(s^*) - nq = 0$ iff

$$\frac{\theta e^{s^*}}{1-\theta+\theta e^{s^*}} = q \quad \Leftrightarrow \quad e^{s^*} = \frac{q}{1-q}\,\frac{1-\theta}{\theta}.$$

Sharper than Hoeffding: the Chernoff-Hoeffding inequality

We found $\psi'(s^*) - nq = 0$ iff

$$\frac{\theta e^{s^*}}{1-\theta+\theta e^{s^*}} = q \quad \Leftrightarrow \quad e^{s^*} = \frac{q}{1-q}\frac{1-\theta}{\theta}.$$

$$\begin{aligned}\log \mathbb{P}\big(\textstyle\sum_i X_i \geq nq\big) &\leq n\log\left(\frac{\theta e^{s^*}}{q}\right) - s^* nq \\ &\leq n\log\tfrac{\theta}{q} + s^* n\,(1-q) = n\log\tfrac{\theta}{q} + n(1-q)\big[\log\tfrac{1-\theta}{1-q} - \log\tfrac{\theta}{q}\big] \\ &= -nq\log\tfrac{q}{\theta} - n(1-q)\log\tfrac{1-q}{1-\theta} = -n\,\mathrm{KL}(q\|\theta)\end{aligned}$$

Bennett's inequality

Let X_i be *independent* bounded r.v. such that

- $\mathbb{E}[X_i] = 0$ and $\mathbb{P}(X_i \leq 1) = 1$.
- Let $\sigma^2 := \frac{1}{n}\sum_i \sigma_i^2$ with $\sigma_i^2 = \mathrm{Var}(X_i) \leq \tau_i^2$.

Then

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_i X_i > x\Big) \leq \exp\Big(-n\,\sigma^2\, h\Big(\frac{x}{\sigma^2}\Big)\Big)$$

$$\text{for } h(u) = (1+u)\log(1+u)$$

Or equivalently

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_i X_i > x\Big) \leq \exp\Big(-n\,(\sigma^2 + x)\,\log\Big(1 + \frac{x}{\sigma^2}\Big)\Big)$$

see, e.g. Boucheron et al. (2003) for a proof.

Bernstein's Inequality

Bennett's inequality: $\mathbb{P}\Big(\frac{1}{n}\sum_i X_i > x\Big) \le \exp\Big(-n\,\sigma^2\,h\Big(\frac{x}{\sigma^2}\Big)\Big)$

for $h(u) = (1+u)\log(1+u)$ but $h(u) \ge \frac{1}{2}\frac{u^2}{1+u/3}$ which implies

Bernstein's inequality

$$\mathbb{P}\Big(\frac{1}{n}\sum_i X_i > x\Big) \le \exp\Big(-\frac{nx^2}{2(\sigma^2 + x/3)}\Big)$$

compare with Hoeffding's inequality

$$\mathbb{P}\Big(\frac{1}{n}\sum_i X_i > x\Big) \le \exp\Big(-\frac{nx^2}{2\tau^2}\Big)$$

- If $x \ll \sigma^2$ this captures the right asymptotic variance
- If $\sigma^2 + x/3 \ge \tau^2$ then this is worse than Hoeffding
- But when $\sigma^2 + x/3 < \tau^2$ it captures relevant behavior for small σ^2
  - e.g. $\mathrm{Bin}(n, \lambda/n) \to \mathrm{Poisson}(\lambda)$ with tail in $e^{-\lambda}$.

High probability statement of Berstein's inequality

Bernstein's inequality

$$\mathbb{P}\left(\frac{1}{n}\sum_i X_i > x\right) \leq \exp\left(-\frac{nx^2}{2(\sigma^2 + x/3)}\right)$$

By solving for x in $t = nx^2 / \big(2(\sigma^2 + x/3)\big)$ we get

$$x = \frac{t}{3n} + \sqrt{\frac{t^2}{9n^2} + \frac{2\sigma^2 t}{n}} \geq \frac{t}{3n} + \sqrt{\frac{2\sigma^2 t}{n}},$$

we get

$$\mathbb{P}\left(\frac{1}{n}\sum_i X_i > \sqrt{\frac{2\sigma^2 t}{n}} + \frac{t}{3n}\right) \leq e^{-t}$$

So that with probability $1 - \delta$, we have

$$\frac{1}{n}\sum_i X_i > \sqrt{\frac{2\sigma^2 \log(\frac{1}{\delta})}{n}} + \frac{\log(\frac{1}{\delta})}{3n}$$

References I

Boucheron, S., Lugosi, G., and Bousquet, O. (2003). Concentration inequalities. In *Summer School on Machine Learning*, pages 208–240. Springer.

Massart, P. (2003). Concentration inequalities and model selection. Lectures from the 33rd Saint-Flour Summer School on Probability Theory. *Lecture Notes in Mathematics*, 1896.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.