Concentration of measure in probability and high-dimensional statistical learning

Guillaume Aubrun, Aurélien Garivier, Rémi Gribonval

remi.gribonval@inria.fr

http://perso.ens-lyon.fr/remi.gribonval

This week

- **Bounded difference (McDiarmid's) inequality**
- **The PAC framework for statistical learning**
- **Agnostic PAC bounds for ERM**
- **Sub-Gaussianity / sub-exponential variables**

Agnostic PAC bounds for empirical risk minimization

Probably Approximately Correct guarantees

- **Goal: establish PAC bounds:** $P(\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon) \geq 1 - \delta$
  - ✓ given a task (=loss+hypothesis class), bounds depend on
    - ✦ algorithm/principle
    - ✦ ***and data distribution***
- ***Agnostic* PAC bounds:** when no assumption needed on data distribution

Probably Approximately Correct guarantees

- **Goal: establish PAC bounds:** $P(\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon) \geq 1-\delta$
  - ✓ given a task (=loss+hypothesis class), bounds depend on
    - ✦ algorithm/principle
    - ✦ ***and data distribution***
- ***Agnostic*** **PAC bounds:** when no assumption needed on data distribution
- Notion of sample complexity (sharp or not) $n(\epsilon, \delta)$

Case study / exercice

- « Application » scenario
  - ✓ several vendors provide a spam detection tool
  - ✓ training set: mails correctly labeled as spam / non-spam
  - ✓ approach: select the tool with the least error
  - ✓ goal: predict how accurate it will be
- Exercice
  - ✓ formalize the problem
  - ✓ propose PAC bounds

``Formalization'' (last time)

- **Sample space:** {all possible mails}
- **Hypothesis class:** *finite* set of *binary* (SPAM / NOT SPAM) classifiers provided by all vendors
- **Loss:** binary (0 if correct, 1 if erroneous)
- **Training set**: some collection of labeled mails
- **Learning algorithm:** select spam detector with smallest (empirical) average loss
  - ✓ average loss= empirical risk
  - ✓ empirical risk minimization

Reminders and hints

- **Empirical risk minimization**

$$\hat{\mathcal{R}}_n(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, h).$$

$$\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{\mathcal{R}}_n(h)$$

- **Use Hoeffding's inequality and the union bound**

Behaviour of the empirical risk

- **Given a fixed hypothesis h**
  - ✓ ***Empirical risk = empirical average*** over *n* (i.i.d.) samples

$$X_i = \ell((x_i, y_i), h)$$

  - ✓ ***Expectation = true risk***

$$\mu := \mathbb{E}[X_i] = \mathcal{R}(h)$$

  - ✓ ***Bounded*** (binary) loss: can use Hoeffding's inequality

$$P(|\bar{X}_n - \mu| > t) \leq 2e^{-\frac{2nt^2}{(b-a)^2}}$$

How to handle multiple hypotheses ?

- **If I know that h1 is best:**
  - ✓ except with probability at most $e^{-2n\epsilon^2}$ it holds that

$$\mathcal{R}(h^\star) = \mathcal{R}(h_1) \leq \hat{\mathcal{R}}_n(h_1) + \epsilon$$

- **If I don't know which is best**
  - ✓ except with probability at most $2e^{-2n\epsilon^2}$ it holds that

$$\hat{\mathcal{R}}_n(h_1) - \epsilon \leq \mathcal{R}(h_1) \leq \hat{\mathcal{R}}_n(h_1) + \epsilon$$

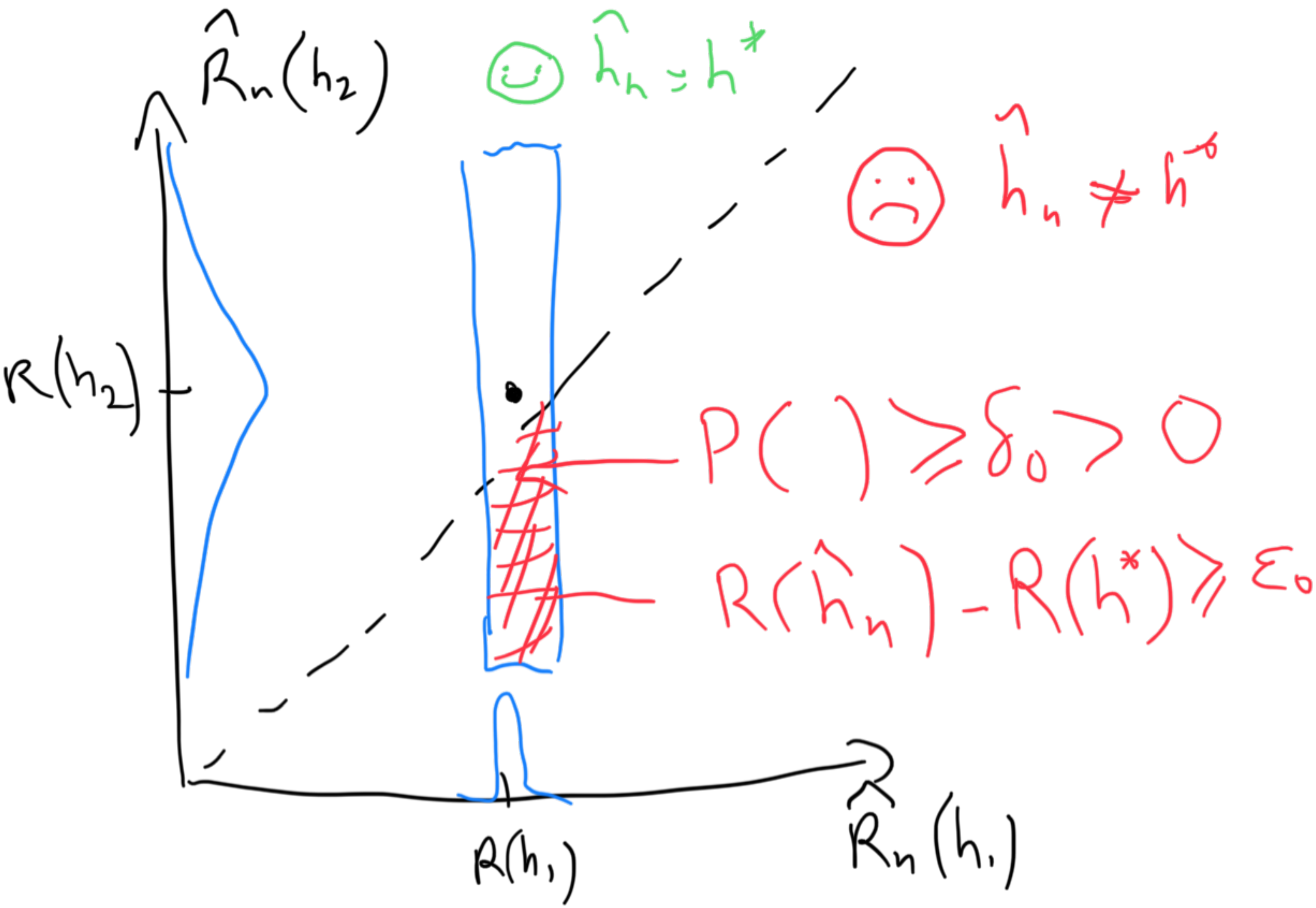
  - ✓ except with probability at most $2e^{-2n\epsilon^2}$ it holds that

$$\hat{\mathcal{R}}_n(h_2) - \epsilon \leq \mathcal{R}(h_2) \leq \hat{\mathcal{R}}_n(h_2) + \epsilon$$

  - ✓ …
  - ✓ except with probability at most $2Ke^{-2n\epsilon^2}$ it holds that

$$\hat{\mathcal{R}}_n(h_i) - \epsilon \leq \mathcal{R}(h_i) \leq \hat{\mathcal{R}}_n(h_i) + \epsilon \text{ for all } 1 \leq i \leq K$$

WHITEBOARD



$\hat{R}_n(h_2)$
$\hat{h}_n = h^*$
$\hat{h}_n \neq h^*$
$R(h_2)$
$P(\ ) \geq \delta_0 > 0$
$R(\hat{h}_n) - R(h^*) \geq \varepsilon_0$
$R(h_1)$
$\hat{R}_n(h_1)$

Agnostic PAC bounds for ERM learning with finite bounded class

Summary: ERM with bounded loss $0 \leq \ell(z,h) \leq B$ and finite hypothesis class

- Agnostic **uniform convergence**: for any n, $t > 0$ and $\mathbb{P}$
$$P(\max_{h \in \mathcal{H}} |\hat{\mathcal{R}}_n(h) - \mathcal{R}(h)| \geq t) \leq 2|\mathcal{H}| \cdot e^{-2nt^2/B^2}.$$
- Agnostic **PAC bound**: for any $n, \epsilon > 0$ and $\mathbb{P}$
$$P(\mathcal{R}(\hat{h}_n) - \mathcal{R}(h^*) \geq \epsilon) \leq 2|\mathcal{H}| \cdot e^{-\frac{n\epsilon^2}{2B^2}}$$
- Agnostic (*upper* bound on) sample complexity: precision ϵ, probability level δ, as soon as
$$n \geq \frac{2B^2}{\epsilon^2} \cdot (\log 2|\mathcal{H}| + \log 2/\delta).$$

Agnostic PAC bounds for ERM learning with finite bounded class

Summary: ERM with bounded loss $0 \leq \ell(z,h) \leq B$ and finite hypothesis class

- Agnostic **uniform convergence**: for any n, $t > 0$ and $\mathbb{P}$

$$P(\max_{h\in\mathcal{H}} |\hat{\mathcal{R}}_n(h) - \mathcal{R}(h)| \geq t) \leq 2|\mathcal{H}| \cdot e^{-2nt^2/B^2}.$$

- Agnostic **PAC bound**: for any $n, \epsilon > 0$ and $\mathbb{P}$

$$P(\mathcal{R}(\hat{h}_n) - \mathcal{R}(h^*) \geq \epsilon) \leq 2|\mathcal{H}| \cdot e^{-\frac{n\epsilon^2}{2B^2}}$$

- Agnostic (*upper* bound on) sample complexity: precision ϵ, probability level δ, as soon as

$$n \geq \frac{2B^2}{\epsilon^2} \cdot (\log 2|\mathcal{H}| + \log 2/\delta) .$$

sharpness? lower-bounds, information theory (with A. Garivier)
unbounded loss? sub-gaussiannity **(next)**
infinite hypothesis class? VC-dim (with A. Garivier)

Sub-gaussian random variables

Reminders of Lecture 1

- **Markov's inequality**

$$\text{if } Z \geq 0 \text{ then} : \mathbb{P}(Z > t) \leq \frac{\mathbb{E}[Z]}{t}, \ \forall t > 0$$

- **Chebyshev's inequality**

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > t) \leq \frac{\mathtt{Var}[Z]}{t^2}, \ \forall t > 0$$

- **Chernoff's bound**

$$\mathbb{P}(Z > t) \leq \frac{\mathbb{E}[e^{\lambda Z}]}{e^{\lambda t}}, \ \forall t, \lambda > 0$$

Reminders of Lecture 1

- **Markov's inequality**

$$\text{if } Z \geq 0 \text{ then} : \mathbb{P}(Z > t) \leq \frac{\mathbb{E}[Z]}{t}, \ \forall t > 0$$

- **Chebyshev's inequality**

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > t) \leq \frac{\mathtt{Var}[Z]}{t^2}, \ \forall t > 0$$

- **Chernoff's bound**

Moment generating function

$$\mathbb{P}(Z > t) \leq \frac{\mathbb{E}[e^{\lambda Z}]}{e^{\lambda t}}, \ \forall t, \lambda > 0$$

Bounding the moment generating function

- **Case of *bounded* variables**
  - **Hoeffding's lemma**, assuming $a \leq Z \leq b,\ \mu := \mathbb{E}(Z)$

$$\mathbb{E}(e^{\lambda(Z-\mu)}) \leq e^{\lambda^2(b-a)^2/8}, \quad \forall \lambda > 0$$

  - worst-case over *all* bounded variables
  - what if
    - we have more information (e.g. controlled variance) ?
    - unbounded variables ?
- **Observation:** controlling the moment generating function is enough to get ***Hoeffding's inequality***

Beyond bounded variables: sub-Gaussianity (scalar variables)

- **Definition:**
  - ✓ a ***centered*** random variable Z is **sub-Gaussian** with parameter $\sigma > 0$ if

$$\mathbb{E}[e^{\lambda Z}] \leq e^{\lambda^2 \sigma^2 / 2}, \ \forall \lambda \in \mathbb{R}$$

  - ✓ a random variable *X* that admits an expectation is sub-Gaussian if $X - \mathbb{E}[X]$ is sub-Gaussian
- **Property:** if X is sub-Gaussian with parameter $\sigma > 0$ then for each t>0

$$P(X - \mathbb{E}[X] > t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Proof: HOMEWORK

Sub-gaussianity
Examples & counter-examples (1)

- **Gaussian variables:** if $Z \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathbb{E}(e^{\lambda Z}) = e^{\lambda\mu + \lambda^2\sigma^2/2}$.

Sub-gaussianity
Examples & counter-examples (1)

- **Gaussian variables:** if $Z \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathbb{E}(e^{\lambda Z}) = e^{\lambda\mu + \lambda^2\sigma^2/2}$.
- **Bounded variables**: why ?

Sub-gaussianity
Examples & counter-examples (1)

- **Gaussian variables:** if $Z \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathbb{E}(e^{\lambda Z}) = e^{\lambda\mu + \lambda^2\sigma^2/2}$.
- **Bounded variables**: why ?
- **Weighted sums** of independent subG. variables

> **Additivity property of sub-Gaussian random variables**: if X_i are independent sub-Gaussian with parameters σ_i and $\lambda_i \in \mathbb{R}$ then $\sum_{i=1}^n \lambda_i X_i$ is sub-Gaussian with parameter $\sqrt{\sum_i \lambda_i^2 \sigma_i^2}$.

✓ Proof: HOMEWORK

Sub-gaussianity
Examples & counter-examples (1)

- **Gaussian variables:** if $Z \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathbb{E}(e^{\lambda Z}) = e^{\lambda\mu+\lambda^2\sigma^2/2}$.
- **Bounded variables**: why ?
- **Weighted sums** of independent subG. variables

> **Additivity property of sub-Gaussian random variables**: if X_i are independent sub-Gaussian with parameters σ_i and $\lambda_i \in \mathbb{R}$ then $\sum_{i=1}^n \lambda_i X_i$ is sub-Gaussian with parameter $\sqrt{\sum_i \lambda_i^2 \sigma_i^2}$.

  - ✓ Proof: HOMEWORK
- **Rademacher variables** $P(Z = +1) = P(Z = -1) = 1/2$
  - ✓ why ? which $\sigma > 0$
  - ✓ EXERCISE: direct proof ?

EXERCISE: Rademacher variables

EXERCISE: Rademacher variables

- Hints:
  - ✓ develop moment generating function into power series
  - ✓ use that $(2k)! \geq 2^k k!$

$$\mathbb{E}[e^{\lambda Z}] = \frac{1}{2}\left(e^{\lambda} + e^{-\lambda}\right) = \cosh\lambda = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!}$$

$$\leq \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2^k k!}$$

$$= e^{\lambda^2/2}$$

$$\sigma^2 = 1$$

Sub-gaussianity
Examples & counter-examples (2)

- **Chi-square variables** $X \sim \mathcal{N}(0,1),\ Z = X^2$

$$\mathbb{E}[e^{\lambda(Z-1)}] = \begin{cases} \frac{1}{\sqrt{1-2\lambda}} & \lambda \in [0, 1/2) \\ +\infty & \lambda \geq 1/2 \end{cases}$$

  - see e.g. Foundations of Machine Learning (C.14)

- **Do we loose all concentration properties ?**
  - ✓ upcoming: notion of *sub-exponential* random variables
  - ✓ application: *Johnson-Lindenstrauss lemma*

Sub-exponential random variables

sub-Gaussian *vs* sub-exponential

- **Definition:**
  - ✓ a ***centered*** random variable Z is **sub-Gaussian** with parameter $\sigma > 0$ if

$$\mathbb{E}[e^{\lambda Z}] \leq e^{\lambda^2 \sigma^2 / 2}, \ \forall \lambda \in \mathbb{R}$$

  - ✓ a random variable *X* that admits an expectation is sub-Gaussian if $X - \mathbb{E}[X]$ is sub-Gaussian

sub-Gaussian *vs* sub-exponential

- **Definition:**
  - ✓ a ***centered*** random variable Z is **sub-Gaussian** with **sub-exponential** parameter $\sigma > 0$ if parameters $\nu, b > 0$

$$\mathbb{E}[e^{\lambda Z}] \leq e^{\lambda^2 \nu^2/2}, \ \forall \lambda \in \mathbb{R}$$

$$\in [-1/b, 1/b]$$

  - ✓ a random variable *X* that admits an expectation is sub-Gaussian if exponential $X - \mathbb{E}[X]$ is sub-Gaussian **sub-exponential**

Properties of sub-exponential variables

- **Concentration:** if *Z* is sub-exponential then

$$\mathbb{P}(Z \geq \mu + t) \leq \begin{cases} e^{-t^2/2\nu^2}, & \text{if } 0 \leq t \leq \nu^2/b \\ e^{-t/2b}, & \text{for } t > \nu^2/b \end{cases}$$

  - ✓ Hence the name-subexponential
  - ✓ Proof: EXERCISE

- **Additivity**

Additivity property of sub-exponential random variables: if X_i are sub-exponential with parameters ν_i, b_i and $\lambda_i \in \mathbb{R}$ then $\sum_{i=1}^n \lambda_i X_i$ is sub-exponential with parameter $\nu \leq ??$ and $b \geq ??$.

  - ✓ Proof: Home practice

Characterizations

Theorem 1 (Characterizing sub-Exponential variables, cf Vershynin, Prop 2.7.1)**.** *Assume Z is zero mean. Then the following properties are equivalent:*

(1) *there are ν, b such that $\mathbb{E}(e^{\lambda Z}) \leq e^{\lambda^2 \nu^2/2}$ for all $|\lambda| < 1/b$.*

(2) **sub-exponential tails***: there are $c_0, c_1 > 0$ such that*

$$\mathbb{P}(|Z| \geq t) \leq c_0 e^{-c_1 t}, \quad \forall t > 0$$

(3) **moment growth***: there is $c_2 > 0$ such that*

$$\left[\mathbb{E}(|Z|^k)\right]^{1/k} \leq c_2 k, \quad \forall k \geq 1$$

(4) *there is $c_3 > 0$ such that $\mathbb{E}(e^{\lambda |Z|}) \leq e^{c_3 \lambda}$ for $0 \leq \lambda \leq 1/c_3$.*

(5) *there is $c_4 > 0$ such that $\mathbb{E}(e^{c_4 |Z|}) < \infty$.*

Characterizations

Theorem 1 (Characterizing sub-Exponential variables, cf Vershynin, Prop 2.7.1). *Assume Z is zero mean. Then the following properties are equivalent:*

(1) *there are ν, b such that $\mathbb{E}(e^{\lambda Z}) \leq e^{\lambda^2\nu^2/2}$ for all $|\lambda| < 1/b$.*
(2) **sub-exponential tails**: *there are $c_0, c_1 > 0$ such that*

$$\mathbb{P}(|Z| \geq t) \leq c_0 e^{-c_1 t}, \quad \forall t > 0$$

(3) **moment growth**: *there is $c_2 > 0$ such that*

$$\left[\mathbb{E}(|Z|^k)\right]^{1/k} \leq c_2 k, \quad \forall k \geq 1$$

(4) *there is $c_3 > 0$ such that $\mathbb{E}(e^{\lambda|Z|}) \leq e^{c_3\lambda}$ for $0 \leq \lambda \leq 1/c_3$.*
(5) *there is $c_4 > 0$ such that $\mathbb{E}(e^{c_4|Z|}) < \infty$.*

Theorem 2 (Characterizing sub-Gaussian variables, cf Vershynin, Prop 2.5.2). *Assume Z is zero mean. Then the following properties are equivalent:*

(1) *there is σ such that $\mathbb{E}(e^{\lambda Z}) \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$.*
(2) **sub-gaussian tails**: *there are $c_0, c_1 > 0$ such that*

$$\mathbb{P}(|Z| \geq t) \leq c_0 P(|X| \geq t), \quad \forall t > 0, \ \textit{with } X \sim N(0, c_1)$$

(3) **moment growth**

$$\left[\mathbb{E}(|Z|^k)\right]^{1/k} \leq c_2 \sqrt{k}, \quad \forall k \geq 1$$

(4) *there is c_3 such that $\mathbb{E}(e^{\lambda^2 Z^2}) \leq e^{c_3^2\lambda^2}$ for $|\lambda| \leq 1/c_3$.*
(5) *there is c_4 such that $\mathbb{E}(e^{c_4 Z^2}) < \infty$.*

Bernstein's condition

- **Theorem**
  - ✓ denote $\mu = \mathbb{E}(Z)$ and $V = \mathtt{Var}(Z)$
  - ✓ assume $\mathbb{E}(|Z-\mu|^k) \leq \frac{1}{2}k!Vb^{k-2}, \quad \text{for } k = 3, 4, \ldots$
  - ✓ then
    - $\mathbb{E}(e^{\lambda(Z-\mu)}) \leq e^{\frac{\lambda^2 V}{2(1-|\lambda|b)}}$ for all $|\lambda| < 1/b$
    - Z is sub-exponential with parameters $\nu = \sqrt{2}\sqrt{V}$ and $2b$.
- **Proof sketch:**
  - ✦ Develop into power series and use moments and definition
- **Exercice:**
  - ✓ assume $|Z-\mu| \leq b$ and $V = \mathtt{Var}(Z) \leq b^2$
  - ✓ check Bernstein's condition
  - ✓ compare to Hoeffding's inequality
- Home practice (to go further): compare to Bennett's inequality

That’s all folks !

Exploiting variance information via Bennett's inequality

- **Assumptions & notations**
  - ✓ n *independent* random variables $X_1, \ldots, X_n$ satisfy

    $$\mathbb{E}[X_i] = 0 \quad X_i \leq c \qquad \sigma^2 := \frac{1}{n}\sum_{i=1}^{n} \mathtt{Var}(X_i)$$

  - ✓ Then for each t>0

    $$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i > t\right) \leq \exp\left(-\frac{n\sigma^2}{c^2}\theta(tc/\sigma^2)\right)$$

    where $\theta(u) := (1+u)\log(1+u) - u$

- Home practice
  - proof
  - comparison to Hoeffding's inequality in the small and large deviation regimes to be expressed e.g. as $t \ll t_0$

Hints

- Show that for any t>0, $$\mathbb{E}[e^{tX_i}] \leq e^{f(\texttt{Var}(X_i)/c^2)}$$
  - ✓ where $$f(x) = \log\left(\frac{1}{1+x}e^{-ctx} + \frac{x}{1+x}e^{ct}\right), \ \forall x \geq 0$$
- Show that f is concave
- Use Chernoff's bounding technique