Concentration Inequalities

DIMA Genova

Summer school 2017

Andreas Maurer

Tail bounds

Z a real random variable. We look for bounds of the form

$$\mathsf{Pr}\left\{Z - E\left[Z\right] > t\right\} \leq \Xi\left(t\right)$$

where we hope that Ξ exponentially decreasing

Applications

- Statistics
- Mathematical physics
- Learning Theory
- Economy
- Computer science

Problem setup

$(\Omega, \Sigma, \mu) = \prod_{i=1}^{n} (\Omega_i, \Sigma_i, \mu_i)$ a product of probability spaces
$\mathcal{A} =$ the algebra of bounded measurable functions $f : \Omega \rightarrow \mathbb{R}$

X_i a random variable distributed as μ_i, so
$\mathbf{X} = (X_1, ..., X_n)$ is a vector of independent varianbles

Objective: for $f \in \mathcal{A}$ and $t > 0$, bound

$$\Pr\{f(\mathbf{X}) - E[f(\mathbf{X})] > t\} = \Pr\{f - E[f] > t\}$$

Markov inequality and exponential moment method

Theorem (*Markov inequality*): If $f \in L_1(\mu)$, $f \geq 0$ and $t > 0$ then

$$\Pr\{f > t\} \leq \frac{E[f]}{t}.$$

Theorem (*Chebychev inequality*): If $f \in L_2(\mu)$ and $t > 0$ then

$$\Pr\{|f - E[f]| > t\} \leq \frac{\sigma^2(f)}{t^2}.$$

Theorem (*Exponential moment method*): If $f \in L_\infty[\mu]$ and $t > 0$ then

$$\Pr\{f - E[f] > t\} \leq \inf_{\beta \geq 0} \exp\left(\ln E\left[e^{\beta(f - E[f])}\right] - \beta t\right).$$

$\ln E\left[e^{\beta(f-E[f])}\right]$ is called the *moment generating function (mgf)*.

Entropy and the moment generating function

Theorem : If $f \in \mathcal{A}$ and $\beta \in \mathbb{R}$ then

$$\ln E\left[e^{\beta f}\right] = \sup_{\rho = \text{p-density on } \Omega} \beta E\left[\rho f\right] - E\left[\rho \ln \rho\right]$$

and the supremum is attained by the density

$$\rho_{\beta f} = \frac{e^{\beta f}}{E\left[e^{\beta f}\right]}.$$

Terminology:

Partition function	$Z_{\beta f} = E\left[e^{\beta f}\right]$
Thermal measure	$d\mu_{\beta f} = \rho_{\beta f} d\mu = e^{\beta f} d\mu / Z_{\beta f}$
Thermal expectation	$E_{\beta f}\left[g\right] = E\left[g e^{\beta f}\right] / Z_{\beta f}$
Entropy	$\text{Ent}_f\left(\beta\right) = E\left[\rho_{\beta f} \ln \rho_{\beta f}\right] = \beta E_{\beta f}\left[f\right] - \ln Z_{\beta f}$

Entropy and concentration

Theorem: For $f \in \mathcal{A}$ and $\beta \geq 0$

$$\begin{aligned} \ln E\left[e^{\beta(f-E[f])}\right] &= \beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma \\ \Pr\{f - E[f] > t\} &\leq \inf_{\beta \geq 0} \exp\left(\beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma - \beta t\right). \end{aligned}$$

This is the reason why we want to bound the entropy!

Entropy, thermal variance and fluctuations

$$\sigma^2_{\beta f}(g) = E_{\beta f}\left[\left(g - E_{\beta f}[g]\right)^2\right].$$

Formulas:

$$\begin{aligned} \frac{d}{d\beta} \ln E\left[e^{\beta f}\right] &= E_{\beta f}[f] \\ \frac{d}{d\beta} E_{\beta f}[g] &= E_{\beta f}[fg] - E_{\beta f}[f]\, E_{\beta f}[g]. \end{aligned}$$

Theorem (fluctuation representation of entropy): For $f \in \mathcal{A}$ and $\beta \geq 0$

$$\mathrm{Ent}_f(\beta) = \int_0^\beta \int_t^\beta \sigma^2_{sf}(f)\, ds dt.$$

Product spaces

$$f:\Omega=\prod_{i=1}^{n}\Omega_i\mapsto f\left(\mathbf{x}\right)=f\left(x_1,...,x_n\right)\in\mathbb{R}.$$

$\mathcal{A}_k=\{$functions in $\mathcal{A}$ which don't depend on the k-th argument$\}$

Substitution operator $S_y^k:\mathcal{A}\to\mathcal{A}_k$

$$S_y^k f\left(\mathbf{x}\right)=f\left(x_1,...,x_{k-1},y,x_{k+1},...,x_n\right)$$

Conditional expectation $E_k:\mathcal{A}\to\mathcal{A}_k$ defined by $E_k\left[g\right]=E_{y\sim\mu_k}\left[S_y^k f\right]$

Conditional quantities

$Z_{k,\beta f} = E_k\left[e^{\beta f}\right]$	conditional partition function
$E_{k,\beta f}[g] = Z_{k,\beta f}^{-1} E_k\left[g e^{\beta f}\right]$	conditional thermal expectation
$\mathsf{Ent}_{k,f}(\beta) = \beta E_{k,\beta f}[g] - \ln Z_{k,\beta f}$	conditional entropy
$\sigma^2_{k,\beta f}[g] = E_{k,\beta f}\left[\left(g - E_{k,\beta f}[g]\right)^2\right]$	conditional thermal variance
$\sigma^2_k[g] = E_k\left[(g - E_k[g])^2\right]$	conditional variance

Lemma: For $g \in \mathcal{A}$, $E_{\beta f}\left[E_{k,\beta f}[g]\right] = E_{\beta f}[g]$.

Thermal subadditivity of entropy

Lemma: If $h, g > 0$ then

$$E[h] \ln \frac{E[h]}{E[g]} \leq E\left[h \ln \frac{h}{g}\right].$$

Lemma: If $\rho \in \mathcal{A}$ and $\rho > 0$ then

$$E\left[\rho \ln \frac{\rho}{E[\rho]}\right] \leq \sum_k E\left[\rho \ln \frac{\rho}{E_k[\rho]}\right].$$

Theorem: For $f \in \mathcal{A}$ and $\beta \in \mathbb{R}$

$$\mathsf{Ent}_f(\beta) \leq E_{\beta f}\left[\sum_{k=1}^{n} \mathsf{Ent}_{k,f}(\beta)\right].$$

Summary of results

For $f \in \mathcal{A}$ and $\beta \in \mathbb{R}$

$$\begin{aligned}
\Pr\{f - Ef > t\} &\leq \inf_{\beta \geq 0} \exp\left(\ln E\left[e^{\beta(f-Ef)}\right] - \beta t\right) \\
\ln E\left[e^{\beta(f-Ef)}\right] &= \beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma \\
\mathrm{Ent}_f(\beta) &\leq E_{\beta f}\left[\sum_{k=1}^n \mathrm{Ent}_{k,f}(\beta)\right] \\
\mathrm{Ent}_f(\beta) &= \int_0^\beta \int_t^\beta \sigma_{sf}^2[f]\, ds\ dt
\end{aligned}$$

Tomorrow

- Efron-Stein inequality
- Bounded Difference inequality
- Bernstein-Bennett inequalities
- Various applications

Summary of results

For $f \in \mathcal{A}$ and $\beta \in \mathbb{R}$

$$\begin{aligned}
\Pr\{f - Ef > t\} &\leq \inf_{\beta \geq 0} \exp\left(\ln E\left[e^{\beta(f-Ef)}\right] - \beta t\right) \\
\ln E\left[e^{\beta(f-Ef)}\right] &= \beta \int_0^\beta \frac{\mathrm{Ent}_f(\gamma)}{\gamma^2} d\gamma \\
\mathrm{Ent}_f(\beta) &\leq E_{\beta f}\left[\sum_{k=1}^n \mathrm{Ent}_{k,f}(\beta)\right] \\
\mathrm{Ent}_f(\beta) &= \int_0^\beta \int_t^\beta \sigma_{sf}^2[f]\, ds\, dt
\end{aligned}$$

where $E_{\beta f}[g] = E\left[ge^{\beta f}\right] / E\left[e^{\beta f}\right]$ and $\sigma_{\beta f}^2(g) = E_{\beta f}\left[\left(g - E_{\beta f}[g]\right)^2\right]$
and $\mathrm{Ent}_{(k)f}(\beta) = \beta E_{(k)\beta f}[f] - \ln E_{(k)}\left[e^{\beta f}\right]$

Some operators on $\mathcal{A}$

For $k \in \{1, ..., n\}$, $y, y' \in \Omega_k$ and $f \in \mathcal{A}$

$$
\begin{array}{rrcl}
\text{partial difference operator} & D^k_{y,y'} f & = & S^k_y f - S^k_{y'} f. \\
\text{conditional variance} & \sigma^2_k(f) & = & \frac{1}{2} E_{(y,y') \sim \mu^2_k} \left[\left(D^k_{y,y'} f \right)^2 \right] \\
\text{conditional range} & r_k(f) & = & \sup_{y,y' \in \Omega_k} D^k_{y,y'} f \\
\text{sum of conditional variances} & \Sigma^2(f) & = & \sum_{k=1}^{n} \sigma^2_k(f) \\
\text{sum of conditional squared ranges} & R^2(f) & = & \sum_{k=1}^{n} r^2_k(f)
\end{array}
$$

The Efron-Stein Inequality

Theorem (Efron-Stein inequality)**:** For $f \in \mathcal{A}$

$$\sigma^2(f) \leq E\left[\Sigma^2(f)\right]$$

The bounded difference inequality

Lemma: If $a \leq f \leq b$ then

$$\sigma^2(f) \leq \frac{(b-a)^2}{4}$$

Theorem: For $f \in \mathcal{A}$ and $t > 0$

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-2t^2}{\sup_{\mathbf{x} \in \Omega} R^2(f)(\mathbf{x})}\right).$$

Recall that $R^2(f) = \sum_{k=1}^{n} r_k^2(f) = \sum_{k=1}^{n} \sup_{y,y' \in \Omega_k} \left(D^k_{y,y'} f\right)^2$

Corollaries of the bounded difference inequality

Corollary 1 *(Hoeffding's inequality)*:
Let X_k be real random variables $a_k \leq X_k \leq b_k$. Then

$$\Pr\left\{\sum_k (X_k - E[X_k]) > t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right).$$

Corollary 2 *(Little bounded difference inequality)*: For $f \in \mathcal{A}$ and $t > 0$

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-2t^2}{nc^2}\right),$$

$$\text{where } c = \max_k \sup_{\mathbf{x}\in\Omega, y, y' \in \Omega_k} D^k_{y,y'} f(\mathbf{x}).$$

Application: vector valued concentration

Theorem: X_i independent r.v. with values in a normed space $\mathcal{B}$
$E[X_i] = 0$ and $\|X_i\| \leq c_i$.
Then for $\delta > 0$ with probability at least $1 - \delta$

$$\left\|\sum_i X_i\right\| \leq E\left\|\sum_i X_i\right\| + \sqrt{2\sum_i c_i^2 \ln(1/\delta)}.$$

If $\mathcal{B}$ is a Hilbert-space and the X_i are iid then

$$\left\|\frac{1}{n}\sum_i X_i\right\| \leq \sqrt{\frac{E\left[\|X_1\|^2\right]}{n}} + c_1\sqrt{\frac{2\ln(1/\delta)}{n}}$$

Application: Rademacher complexities

Theorem: $\mathcal{F}$ a class of functions $f : \mathcal{X} \rightarrow [0, 1]$
$\mathbf{X} = (X_1, ..., X_n)$ be a vector of iid r.v. with values in $\mathcal{X}$.
Define for Rademacher variables $\epsilon_1, ..., \epsilon_n$

$$\text{Rad}(\mathcal{F}, \mathbf{x}) = \frac{2}{n} E_{\boldsymbol{\epsilon}} \sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i f(x_i) \right|.$$

Then with probability at least $1 - \delta$ in $\mathbf{X}$

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i f(X_i) - E[f(X_i)] \right| \leq \text{Rad}(\mathcal{F}, \mathbf{X}) + 2\sqrt{\frac{2 \ln (2/\delta)}{n}}.$$

Bennett and Bernstein inequalities

Lemma: If $f - Ef \leq 1$. Then for $\beta > 0$ we have $\sigma^2_{\beta f}(f) \leq e^{\beta}\sigma^2(f)$.

Lemma: Assume that $f - E_k f \leq 1$ for all $k \in \{1, ..., n\}$. Then for $\beta > 0$

$$\text{Ent}_f(\beta) \leq \left(\beta e^{\beta} - e^{\beta} + 1\right) \; E_{\beta f}\left[\Sigma^2(f)\right].$$

Lemma: For $x \geq 0$ we get $(1+x)\ln(1+x) - x \geq 3x^2/(6+2x)$.

Theorem *(Bennett/Bernstein inequalities)*: Assume $f - E_k f \leq 1, \forall k$. Let $t > 0$ and denote $V = \sup_{\mathbf{x}\in\Omega}\Sigma^2(f)(\mathbf{x})$. Then

$$\begin{aligned}\Pr\{f - E[f] > t\} &\leq \exp\left(-V\left(\left(1 + tV^{-1}\right)\ln\left(1 + tV^{-1}\right) - tV^{-1}\right)\right) \\ &\leq \exp\left(\frac{-t^2}{2V + 2t/3}\right).\end{aligned}$$

Application: vector valued concentration revisited

Theorem: X_i iid r.v. with values in a Hilbert space $\mathcal{H}$
$E[X_i] = 0$ and $\|X_i\| \leq c$.

Then for $\delta > 0$ with probability at least $1 - \delta$

$$\left\| \frac{1}{n} \sum_i X_i \right\| \leq \sqrt{\frac{E\left[\|X_1\|^2\right]}{n}} \left(1 + \sqrt{2 \ln(1/\delta)}\right) + \frac{4c \ln(1/\delta)}{2n}.$$

Tomorrow

- Gaussian concentration
- Exponential Efron-Stein inequalities
- Application to convex Lipschitz functions
- Application to random matrices

Gaussian concentration

Theorem (*Ibragimov, Tsirelson, Sudakov*):
$f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is L-Lipschitz (possibly unbounded)
$\mathbf{X} = (X_1, ..., X_n)$ an iid vector $X_i \sim N(0, 1)$.

Then for $t > 0$

$$\Pr\{f(\mathbf{X}) > \mathbb{E}[f(\mathbf{X})] + s\} \leq e^{-t^2/(2L^2)}.$$

A monotonicity bound on the thermal variance

Lemma (Chebychev's association inequality):
$g, h : \mathbb{R} \to \mathbb{R}$, X a real random variable.
If g and h are either both nondecreasing or both nonincreasing then

$$E\left[g\left(X\right)h\left(X\right)\right] \geq E\left[g\left(X\right)\right]E\left[h\left(X\right)\right].$$

If either one of g or h is nondecreasing and the other nonincreasing then

$$E\left[g\left(X\right)h\left(X\right)\right] \leq E\left[g\left(X\right)\right]E\left[h\left(X\right)\right].$$

Lemma (monotonicity bound): If $0 \leq s \leq \beta$. Then

$$\sigma^2_{sf}\left(f\right) \leq E_{x\sim\mu_{\beta f}}\left[E_{x'\sim\mu}\left[\left(f\left(x\right) - f\left(x'\right)\right)^2_+\right]\right].$$

More operators on $\mathcal{A}$

Define two operators $D^2 : \mathcal{A} \to \mathcal{A}$ and $V_+^2 : \mathcal{A} \to \mathcal{A}$ by

$$
\begin{aligned}
D^2 f &= \sum_k \left(f - \inf_{y \in \Omega_k} S_y^k f \right)^2 \\
\text{and } V_+^2 f &= \sum_k E_{y \sim \mu_k} \left[\left(\left(f - S_y^k f \right)_+ \right)^2 \right].
\end{aligned}
$$

Exponential Efron-Stein inequality, upper tail

Lemma : For $\beta > 0$ and $f \in \mathcal{A}$

$$\begin{aligned} \mathsf{Ent}_f(\beta) &\leq \frac{\beta^2}{2} E_{\beta f}\left[V_+^2(f)\right] \\ \text{and } \ln E\left[e^{\beta(f - E[f])}\right] &\leq \frac{\beta}{2}\int_0^\beta E_{\gamma f}\left[V_+^2 f\right] d\gamma. \end{aligned}$$

Theorem: With $t > 0$

$$\Pr\{f - E[f] > t\} \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\mathbf{\Omega}} V_+^2 f(\mathbf{x})}\right) \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\mathbf{\Omega}} D^2 f(\mathbf{x})}\right).$$

Exponential Efron-Stein inequality, lower tail

Lemma : If $f \in \mathcal{A}$ and $f - \inf_k f \leq 1, \forall k$ then for $\beta > 0$

$$\begin{aligned} \mathrm{Ent}_{-f}(\beta) &\leq \psi(\beta) E_{-\beta f}\left[D^2 f\right], \\ \text{where } \psi(\beta) &= e^{\beta} - \beta - 1 \\ \text{and } \ln E\left[e^{\beta(E[f]-f)}\right] &\leq \frac{\psi(\beta)}{\beta} \int_0^{\beta} E_{-\gamma f}\left[D^2 f\right] d\gamma \end{aligned}$$

Theorem : If $f - \inf_k f \leq 1$ for all k and with $\Delta := \sup_{\mathbf{x} \in \mathbf{\Omega}} D^2 f(\mathbf{x})$, then for $t > 0$

$$\begin{aligned} \Pr\{E[f] - f > t\} &\leq \exp\left(-\Delta\left(\left(1 + \frac{t}{\Delta}\right) \ln\left(1 + \frac{t}{\Delta}\right) - \frac{t}{\Delta}\right)\right) \\ &\leq \exp\left(\frac{-t^2}{2 \sup_{\mathbf{x} \in \mathbf{\Omega}} D^2 f(\mathbf{x}) + 2t/3}\right). \end{aligned}$$

Application: convex Lipschitz functions

Theorem: $f : [0,1]^n \longrightarrow \mathbb{R}$
f is L-Lipschitz
f is separately convex (i.e. $y \in [0,1] \mapsto S_y^k f(\mathbf{x})$ is convex for all k and all $\mathbf{x}$)
$X_1, ..., X_n$ are independent with values in $[0,1]$

Then

$$\Pr\{f(\mathbf{X}) > Ef(\mathbf{X}) + s\} \leq e^{-s^2/2L^2}.$$

We wait with the lower tail...

Application: operator norm of a random matrix

Recall: If M is an $m \times n$ matrix its operator norm is

$$\|M\|_\infty = \sup_{v \in \mathbb{R}^n, w \in \mathbb{R}^m, \|w\|, \|v\| = 1} \langle Mv, w \rangle$$

Theorem:
$\mathbf{X} = \left(X_{ij}\right) \in [-1, 1]^{mn}$ a random $m \times n$ matrix with independent X_{ij}.

Then

$$\begin{aligned} \Pr\left\{\|\mathbf{X}\|_\infty - E\left[\|\mathbf{X}\|_\infty\right] \geq t\right\} &\leq \exp\left(\frac{-t^2}{8}\right) \\ \text{and} \ \Pr\left\{E\left[\|\mathbf{X}\|_\infty\right] - \|\mathbf{X}\|_\infty \geq t\right\} &\leq \exp\left(\frac{-t^2}{8 + 4t/3}\right). \end{aligned}$$

Tomorrow

- Concentration of self bounding functions
- Application to convex Lipschitz functions
- Decoupling
- Concentration of the supremum of an empirical process

Beyond uniform bounds

Previous strategy:

1. Bound $\text{Ent}_f(\gamma) \le \xi(\gamma) E_{\gamma f}[G(f)]$
with $G : \mathcal{A} \to \mathcal{A}$, $\xi : \mathbb{R} \to \mathbb{R}_+$
(e.g. $\xi(\gamma) = \gamma^2/8$ and $G = R^2$ for bded difference,
$\xi(\gamma) = \gamma e^\gamma - e^\gamma + 1$ and $G = \Sigma^2$ for Bennet, etc)

2. Bound mgf as

$$\ln E e^{\beta(f-Ef)} \le \beta \int_0^\beta \frac{\xi(\gamma)}{\gamma^2} E_{\gamma f}[G(f)]\, d\gamma \le \beta \sup_{\mathbf{x}} G(f)(\mathbf{x}) \int_0^\beta \frac{\xi(\gamma)\, d\gamma}{\gamma^2}.$$

Now we want to avoid the uniform estimate on $G(f)$ of the last step!

First trick: self-boundedness

Idea: Suppose $G(f) \leq f$. Then

$$\beta \int_0^\beta \frac{\xi(\gamma)}{\gamma^2} E_{\gamma f}[G(f)]\, d\gamma \leq \beta \int_0^\beta \frac{\xi(\gamma)}{\gamma^2} E_{\gamma f}[f]\, d\gamma = \beta \int_0^\beta \frac{\xi(\gamma)}{\gamma^2} \left(\frac{d}{d\gamma} \ln Z_{\gamma f} \right) d\gamma$$

Theorem: If $V_+^2 f \leq af + b$ then

$$\ln E\left[e^{\beta(f-E[f])}\right] \leq \frac{\beta^2 (aE[f] + b)}{2 - a\beta} \text{ and } \ln E\left[e^{\beta f}\right] \leq \frac{2\beta E[f] + \beta^2 b}{2 - a\beta}$$

If $D^2 f \leq af + b$ and $f - \inf_k f \leq 1$ then

$$\ln E\left[e^{\beta(E[f]-f)}\right] \leq \frac{\beta^2 (aE[f] + b)}{2}.$$

Self-boundedness - tail bounds

Lemma: If C, b, $t > 0$, then

$$\inf_{\beta \in [0,1/b)} \left(-\beta t + \frac{C\beta^2}{1 - b\beta} \right) \leq \frac{-t^2}{2\left(2C + bt\right)}.$$

Theorem: If $V_+^2 f \leq af + b$ then

$$\Pr\left\{f - E\left[f\right] > t\right\} \leq \exp\left(\frac{-t^2}{2\left(aE\left[f\right] + b + at/2\right)} \right).$$

If $D^2 f \leq af + b$ and $f - \inf_k f \leq 1$ then

$$\Pr\left\{E\left[f\right] - f > t\right\} \leq \exp\left(\frac{-t^2}{2\left(aE\left[f\right] + b\right)} \right).$$

Convex Lipschitz functions revisited

Theorem: $f : [0, 1]^n \longrightarrow \mathbb{R}$
f is L-Lipschitz
f is separately convex (i.e. $y \in [0,1] \mapsto S_y^k f(\mathbf{x})$ is convex for all k and all $\mathbf{x}$)
f^2 takes values in an interval of length ≤ 1
$X_1, ..., X_n$ are independent with values in $[0, 1]$

Then for $t \in [0, E[f(\mathbf{X})]]$

$$\Pr\{E[f(\mathbf{X})] > f(\mathbf{X}) + s\} \leq e^{-s^2/8L^2}.$$

Second trick: decoupling

Recall Fenchel-Young inequality: $\forall$p.-density ρ, $\forall g$

$$E_{\beta f}\left[g\right] \leq \mathsf{Ent}_f\left(\beta\right) + E\left[\ln e^g\right].$$

With $g = \theta G\left(f\right)$

$$\begin{aligned}\mathsf{Ent}_f\left(\beta\right) &\leq \xi\left(\beta\right) E_{\beta f}\left[G\left(f\right)\right] = \xi\left(\beta\right)\theta^{-1} E_{\beta f}\left[\theta G\left(f\right)\right] \\ &\leq \xi\left(\beta\right)\theta^{-1}\left(\mathsf{Ent}_f\left(\beta\right) + \ln E\left[\exp\left(\theta G\left(f\right)\right)\right]\right).\end{aligned}$$

Rearranging we get for $\theta > \xi\left(\beta\right)$

$$\mathsf{Ent}_f\left(\beta\right) \leq \frac{\xi\left(\beta\right)}{\theta - \xi\left(\beta\right)} \ln E\left[\exp\left(\theta G\left(f\right)\right)\right].$$

The supremum of an empirical process

Theorem: $X_1, ..., X_n$ independent with values in $\mathcal{X}$
$\mathcal{F}$ be a ctble class of functions $f : \mathcal{X} \to [-1, 1]$
$E[f(X_i)] = 0, \forall i \in \{1, ..., n\}$
Define $F : \mathcal{X}^n \to \mathbb{R}$ and $W : \mathcal{X}^n \to \mathbb{R}$ by

$$\begin{aligned} F(\mathbf{x}) &= \sup_{f \in \mathcal{F}} \sum_i f(x_i) \text{ and} \\ W(\mathbf{x}) &= \sup_{f \in \mathcal{F}} \sum_i \left(f^2(x_i) + E\left[f^2(X_i) \right] \right). \end{aligned}$$

Then for $t > 0$

$$\Pr\{F - E[F] > t\} \leq \exp\left(\frac{-t^2}{2E[W] + t} \right).$$