On Equivalence of Martingale Tail Bounds and Deterministic Regret Inequalities

Sasha Rakhlin

Department of Statistics, The Wharton School
University of Pennsylvania

Dec 16, 2015

Joint work with K. Sridharan

arXiv:1510.03925

Outline

If $Z_1, \ldots, Z_n$ independent with $\mathbb{E}Z_t = 0$ then

$$\mathbb{E}\left(\sum_{t=1}^{n} Z_t\right)^2 = \sum_{t=1}^{n} \mathbb{E}Z_t^2.$$

If $Z_1, \ldots, Z_n$ independent with $\mathbb{E}Z_t = 0$ then

$$\mathbb{E}\left(\sum_{t=1}^{n} Z_t\right)^2 = \sum_{t=1}^{n} \mathbb{E}Z_t^2.$$

Extends to Hilbert space

$$\mathbb{E}\left\|\sum_{t=1}^{n} Z_i\right\|^2 = \sum_{t=1}^{n} \mathbb{E}\left\|Z_i\right\|^2.$$

(Pinelis '94): Let $Z_1, \ldots, Z_n$ be a martingale difference sequence in a separable 2-smooth Banach space $(\mathfrak{B}, \|\cdot\|)$. For any $u > 0$

$$P\left(\sup_{n\geq 1}\left\|\sum_{t=1}^{n} Z_t\right\| \geq \sigma u\right) \leq 2\exp\left\{-\frac{u^2}{2D^2}\right\},$$

where $\sigma^2 \geq \sum_{t=1}^{\infty} \|Z_t\|_\infty^2$.

(Pinelis '94): Let $Z_1, \ldots, Z_n$ be a martingale difference sequence in a separable 2-smooth Banach space $(\mathfrak{B}, \|\cdot\|)$. For any $u > 0$

$$P\left(\sup_{n\geq 1}\left\|\sum_{t=1}^{n} Z_t\right\| \geq \sigma u\right) \leq 2\exp\left\{-\frac{u^2}{2D^2}\right\},$$

where $\sigma^2 \geq \sum_{t=1}^{\infty} \|Z_t\|_\infty^2$.

Questions:

- replace σ with sequence-dependent version? is it always possible?
- extend beyond linear structure of Banach spaces?

(Pinelis '94): Let $Z_1, \ldots, Z_n$ be a martingale difference sequence in a separable 2-smooth Banach space $(\mathfrak{B}, \|\cdot\|)$. For any $u > 0$

$$P\left(\sup_{n\geq 1}\left\|\sum_{t=1}^{n} Z_t\right\| \geq \sigma u\right) \leq 2\exp\left\{-\frac{u^2}{2D^2}\right\},$$

where $\sigma^2 \geq \sum_{t=1}^{\infty} \|Z_t\|_{\infty}^2$.

Questions:

- replace σ with sequence-dependent version? is it always possible?
- extend beyond linear structure of Banach spaces?

Contributions:

- address these questions
- the actual technique: equivalence of tail bounds and deterministic **pathwise** regret inequalities

Baby version

Unit Euclidean ball $\mathcal{B}$ in $\mathbb{R}^d$. Let $z_1, \ldots, z_n \in \mathcal{B}$ be arbitrary. Define

$$\widehat{y}_{t+1} = \widehat{y}_{t+1}(z_1, \ldots, z_t) = \mathrm{Proj}_{\mathcal{B}}\left(\widehat{y}_t - \frac{1}{\sqrt{n}} z_t\right)$$

with $\widehat{y}_1 = 0$.

Baby version

Unit Euclidean ball $\mathcal{B}$ in $\mathbb{R}^d$. Let $z_1, \dots, z_n \in \mathcal{B}$ be arbitrary. Define

$$\widehat{y}_{t+1} = \widehat{y}_{t+1}(z_1, \dots, z_t) = \mathrm{Proj}_{\mathcal{B}}\left(\widehat{y}_t - \frac{1}{\sqrt{n}} z_t\right)$$

with $\widehat{y}_1 = 0$. Then,

$$\forall y^*, z_1, \dots, z_n \in \mathcal{B}, \qquad \sum_{t=1}^{n} \langle \widehat{y}_t - y^*, z_t \rangle \le \sqrt{n}$$

Baby version

Unit Euclidean ball $\mathcal{B}$ in $\mathbb{R}^d$. Let $z_1, \ldots, z_n \in \mathcal{B}$ be arbitrary. Define

$$\widehat{y}_{t+1} = \widehat{y}_{t+1}(z_1, \ldots, z_t) = \mathrm{Proj}_{\mathcal{B}}\left(\widehat{y}_t - \frac{1}{\sqrt{n}} z_t\right)$$

with $\widehat{y}_1 = 0$. Then,

$$\forall y^*, z_1, \ldots, z_n \in \mathcal{B}, \quad \sum_{t=1}^n \langle \widehat{y}_t - y^*, z_t \rangle \leq \sqrt{n}$$

Rewrite as

$$\exists (\widehat{y}_t)\ \forall z_1, \ldots, z_n \in \mathcal{B}, \quad \left\| \sum_{t=1}^n z_t \right\| - \sqrt{n} \leq \sum_{t=1}^n \langle \widehat{y}_t, -z_t \rangle.$$

Deterministic inequality:

$$\exists(\widehat{y}_t)\ \forall z_1,\ldots,z_n \in \mathcal{B}, \qquad \left\|\sum_{t=1}^{n} z_t\right\| - \sqrt{n} \leq \sum_{t=1}^{n} \langle \widehat{y}_t, -z_t \rangle. \tag{1}$$

Deterministic inequality:

$$\exists(\widehat{y}_t)\ \forall z_1, \dots, z_n \in \mathcal{B}, \qquad \left\| \sum_{t=1}^{n} z_t \right\| - \sqrt{n} \leq \sum_{t=1}^{n} \langle \widehat{y}_t, -z_t \rangle. \tag{1}$$

Apply to an MDS $Z_1, \dots, Z_n$ with values in $\mathcal{B}$

$$\mathbf{P}\left(\left\| \sum_{t=1}^{n} Z_t \right\| - \sqrt{n} \geq u \right) \leq \mathbf{P}\left(\sum_{t=1}^{n} \langle \widehat{y}_t, -Z_t \rangle \geq u \right) \tag{2}$$

Deterministic inequality:

$$\exists(\widehat{y}_t)\ \forall z_1, \ldots, z_n \in \mathcal{B}, \qquad \left\| \sum_{t=1}^{n} z_t \right\| - \sqrt{n} \leq \sum_{t=1}^{n} \langle \widehat{y}_t, -z_t \rangle. \tag{1}$$

Apply to an MDS $Z_1, \ldots, Z_n$ with values in $\mathcal{B}$

$$\mathbf{P}\left(\left\| \sum_{t=1}^{n} Z_t \right\| - \sqrt{n} \geq u \right) \leq \mathbf{P}\left(\sum_{t=1}^{n} \langle \widehat{y}_t, -Z_t \rangle \geq u \right) \leq \exp\{-u^2/2n\} \tag{2}$$

by Asuma-Hoeffding.

Deterministic inequality:

$$\exists(\widehat{y}_t)\ \forall z_1, \ldots, z_n \in \mathcal{B}, \qquad \left\| \sum_{t=1}^{n} z_t \right\| - \sqrt{n} \le \sum_{t=1}^{n} \langle \widehat{y}_t, -z_t \rangle. \tag{1}$$

Apply to an MDS $Z_1, \ldots, Z_n$ with values in $\mathcal{B}$

$$\mathbf{P}\left(\left\| \sum_{t=1}^{n} Z_t \right\| - \sqrt{n} \ge u \right) \le \mathbf{P}\left(\sum_{t=1}^{n} \langle \widehat{y}_t, -Z_t \rangle \ge u \right) \le \exp\{-u^2/2n\} \tag{2}$$

by Asuma-Hoeffding. Integrate tails:

$$\mathbb{E} \left\| \sum_{t=1}^{n} Z_t \right\| \le c\sqrt{n} \tag{3}$$

Using von Neumann minimax theorem, it is possible to show

$$\exists(\widehat{y}_t)\ \forall y^*, z_1, \ldots, z_n \in \mathcal{B}, \qquad \sum_{t=1}^{n} \langle \widehat{y}_t - y^*, z_t \rangle \le \sup_{\text{mds}} \mathbb{E} \left\| \sum_{t=1}^{n} W_t \right\|$$

Deterministic inequality:

$$\exists(\widehat{y}_t)\ \forall z_1,\dots,z_n \in \mathcal{B}, \qquad \left\|\sum_{t=1}^{n} z_t\right\| - \sqrt{n} \le \sum_{t=1}^{n} \langle \widehat{y}_t, -z_t\rangle. \tag{1}$$

Apply to an MDS $Z_1,\dots,Z_n$ with values in $\mathcal{B}$

$$\mathbf{P}\left(\left\|\sum_{t=1}^{n} Z_t\right\| - \sqrt{n} \ge u\right) \le \mathbf{P}\left(\sum_{t=1}^{n} \langle \widehat{y}_t, -Z_t\rangle \ge u\right) \le \exp\{-u^2/2n\} \tag{2}$$

by Asuma-Hoeffding. Integrate tails:

$$\mathbb{E}\left\|\sum_{t=1}^{n} Z_t\right\| \le c\sqrt{n} \tag{3}$$

Using von Neumann minimax theorem, it is possible to show

$$\exists(\widehat{y}_t)\ \forall y^*, z_1,\dots,z_n \in \mathcal{B}, \qquad \sum_{t=1}^{n} \langle \widehat{y}_t - y^*, z_t\rangle \le \sup_{\text{mds}} \mathbb{E}\left\|\sum_{t=1}^{n} W_t\right\| \le c\sqrt{n}$$

Conclusion: (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1) (up to const)

- $\exists(\widehat{y}_t)\ \forall y^*, z_1, \ldots, z_n \in \mathcal{B}, \quad \sum_{t=1}^n \langle \widehat{y}_t - y^*, z_t \rangle \leq \sqrt{n}$
- $\mathbf{P}\left(\|\sum_{t=1}^n Z_t\| - \sqrt{n} \geq u\right) \leq \exp\{-u^2/2n\}$
- $\mathbb{E}\|\sum_{t=1}^n Z_t\| \leq c\sqrt{n}$

- $\exists(\widehat{y}_t)\ \forall y^*, z_1, \ldots, z_n \in \mathcal{B}, \quad \sum_{t=1}^n \langle \widehat{y}_t - y^*, z_t \rangle \leq \sqrt{n}$
- $\mathbf{P}\left(\|\sum_{t=1}^n Z_t\| - \sqrt{n} \geq u \right) \leq \exp\{-u^2/2n\}$
- $\mathbb{E}\, \|\sum_{t=1}^n Z_t\| \leq c\sqrt{n}$

Curiosities:

- in particular (3) $\Rightarrow$ (2) amplifies in-expectation to high prob.
- improve tail bounds by taking a better gradient descent
- improve gradient descent by finding better tail bounds
- move beyond linear structure of Banach space

- $\exists (\widehat{y}_t)\ \forall y^*, z_1, \ldots, z_n \in \mathcal{B}, \quad \sum_{t=1}^n \langle \widehat{y}_t - y^*, z_t \rangle \le \sqrt{n}$
- $\mathbf{P}\left(\|\sum_{t=1}^n Z_t\| - \sqrt{n} \ge u \right) \le \exp\{-u^2/2n\}$
- $\mathbb{E} \|\sum_{t=1}^n Z_t\| \le c\sqrt{n}$

Curiosities:

- in particular (3) $\Rightarrow$ (2) amplifies in-expectation to high prob.
- improve tail bounds by taking a better gradient descent
- improve gradient descent by finding better tail bounds
- move beyond linear structure of Banach space

< / end of baby version >

Warmup: mirror descent with adaptive step size

$(\mathfrak{B}, \|\cdot\|)$ 2-smooth, $(\mathfrak{B}_*, \|\cdot\|_*)$ denotes dual. $D_{\mathcal{R}} : \mathfrak{B}_* \times \mathfrak{B}_* \to \mathbb{R}$ Bregman divergence w.r.t. $\mathcal{R}$, which is 1-strongly convex on unit ball $\mathcal{B}_* \subset \mathfrak{B}_*$. Denote $R_{\max}^2 \triangleq \sup_{f,g \in \mathcal{B}_*} D_{\mathcal{R}}(f, g)$. *Here z_t's need not be in unit ball.*

Lemma.

$\mathcal{F} \subset \mathfrak{B}_*$ convex. Define,

$$\widehat{y}_{t+1} = \widehat{y}_{t+1}(z_1, \ldots, z_t) = \operatorname*{argmin}_{y \in \mathcal{F}} \left\{ \eta_t \langle y, z_t \rangle + D_{\mathcal{R}}(y, \widehat{y}_t) \right\}$$

and

$$\eta_t \triangleq R_{\max} \min \left\{ 1, \left(\sqrt{\textstyle\sum_{s=1}^{t} \|z_s\|^2} + \sqrt{\textstyle\sum_{s=1}^{t-1} \|z_s\|^2} \right)^{-1} \right\}.$$

Then for any $y^* \in \mathcal{F}$ and any $z_1, \ldots, z_n \in \mathfrak{B}$,

$$\sum_{t=1}^{n} \left\langle \widehat{y}_t - y^*, z_t \right\rangle \leq 2.5 R_{\max} \left(\sqrt{\sum_{t=1}^{n} \|z_t\|^2} + 1 \right).$$

Warmup: mirror descent with adaptive step size

Let $\mathbb{E}_t$ be conditional expectation.

Theorem.

Let $Z_1, \ldots, Z_n$ be a $\mathfrak{B}$-valued MDS. For any $u > 0$,

$$P\left(\frac{\|\sum_{t=1}^n Z_t\| - 2.5R_{\max}\left(\sqrt{V_n} + 1\right)}{\sqrt{V_n + W_n + \left(\mathbb{E}\sqrt{V_n + W_n}\right)^2}} > u \right) \leq \sqrt{2}\exp\left\{-u^2/16\right\},$$

where

$$V_n = \sum_{t=1}^n \|Z_t\|^2 \qquad \text{and} \qquad W_n = \sum_{t=1}^n \mathbb{E}_{t-1}\|Z_t\|^2 .$$

Holds with $W_n \equiv 0$ if MDS conditionally symmetric.

- n-independent, self-normalized, can be extended to p-smooth

summary so far

connection between first-order convex optimization methods and one-sided probabilistic tail bounds

Outline

Interpret as supremum of stochastic process

$$\left\| \sum_{t=1}^{n} Z_t \right\| = \sup_{\|y\|_* \le 1} \sum_{t=1}^{n} \langle y, Z_t \rangle$$

Generalization (after centering): take any stochastic process Z_t and

$$\sup_{g \in \mathcal{G}} \sum_{t=1}^{n} g(Z_t) - \mathbb{E}_{t-1}[g(Z_t)]$$

Interpret as supremum of stochastic process

$$\left\| \sum_{t=1}^{n} Z_t \right\| = \sup_{\|y\|_* \le 1} \sum_{t=1}^{n} \langle y, Z_t \rangle$$

Generalization (after centering): take any stochastic process Z_t and

$$\sup_{g \in \mathcal{G}} \sum_{t=1}^{n} g(Z_t) - \mathbb{E}_{t-1}[g(Z_t)]$$

Enough to consider $\mathcal{D}_t = \sigma(\epsilon_1, \ldots, \epsilon_t)$ generated by i.i.d. Rademacher:

$$\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t)$$

where $\mathbf{x}_t$ is $\mathcal{D}_{t-1}$-measurable. (extend Panchenko's symmetrization technique to martingales)

$$f(\mathbf{x}_t(\epsilon_{1:t-1})) = g(Z_t(\epsilon_{1:t-1}, +1)) - g(Z_t(\epsilon_{1:t-1}, -1))$$

Deterministic regret inequalities

Let

$$y_1, \ldots, y_n \in \{\pm 1\}, \quad x_1, \ldots, x_n \in \mathcal{X}, \quad \mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$$

For a given function $B : \mathcal{F} \times \mathcal{X}^n \to \mathbb{R}$, want a prediction strategy

$$\widehat{y}_t = \widehat{y}_t(x_1, \ldots, x_t, y_1, \ldots, y_{t-1})$$

such that

$$\forall\ (x_t, y_t)_{t=1}^n, \quad \sum_{t=1}^n -\widehat{y}_t y_t \le \inf_{f \in \mathcal{F}} \left\{ \sum_{t=1}^n -y_t f(x_t) + 2B(f; x_1, \ldots, x_n) \right\}.$$

Deterministic regret inequalities

Let

$$y_1, \ldots, y_n \in \{\pm 1\}, \quad x_1, \ldots, x_n \in \mathcal{X}, \quad \mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$$

For a given function $B : \mathcal{F} \times \mathcal{X}^n \to \mathbb{R}$, want a prediction strategy

$$\widehat{y}_t = \widehat{y}_t(x_1, \ldots, x_t, y_1, \ldots, y_{t-1})$$

such that

$$\forall \ (x_t, y_t)_{t=1}^n, \quad \sum_{t=1}^n -\widehat{y}_t y_t \le \inf_{f \in \mathcal{F}} \left\{ \sum_{t=1}^n -y_t f(x_t) + 2B(f; x_1, \ldots, x_n) \right\}.$$

If existence of $(\widehat{y}_t)$ is certified, apply to $y_t = \epsilon_t$ and $x_t = \mathbf{x}_t(\epsilon)$:

$$\mathbf{P}\left(\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^n \epsilon_t f(\mathbf{x}_t) - 2B(f; \mathbf{x}_1, \ldots, \mathbf{x}_n) \right\} \ge u \right) \le \mathbf{P}\left(\sum_{t=1}^n \epsilon_t \widehat{y}_t \ge u \right) \le \exp\{\ldots\}.$$

Lemma.

If for any predictable process $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t) - 2B(f; \mathbf{x}_1, \dots, \mathbf{x}_n)\right] \le 0 \ ,$$

then **there exists** a strategy $(\widehat{y}_t)$ with values $|\widehat{y}_t| \le \sup_{f\in\mathcal{F}} |f(x_t)|$ such that the deterministic inequality holds for all sequences.

Lemma.

If for any predictable process $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t) - 2B(f; \mathbf{x}_1, \ldots, \mathbf{x}_n)\right] \le 0 \ ,$$

then **there exists** a strategy $(\widehat{y}_t)$ with values $|\widehat{y}_t| \le \sup_{f\in\mathcal{F}} |f(x_t)|$ such that the deterministic inequality holds for all sequences.

- automatic amplification to high probability
- existential – no explicit prediction strategy $(\widehat{y}_t)$
- an offset version of sequential Rademacher complexity

$$\mathscr{R}_n(\mathcal{F}; \mathbf{x}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t)\right]$$

- $(\epsilon_1, \ldots, \epsilon_n) \mapsto \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t)$ is not Lipschitz; concentration methods fail

Definition.

Let $r \in (1,2]$. We say that sequential Rademacher complexity of $\mathcal{F}$ exhibits an $n^{1/r}$ growth if

$$\forall n \geq 1, \quad \forall \mathbf{x}, \quad \mathscr{R}_n(\mathcal{F};\mathbf{x}) \leq Cn^{1/r} \cdot \sup_{f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, t \leq n} |f(\mathbf{x}_t(\epsilon))| \, .$$

Using amplification and reverse Hölder (due to Burkholder/Pisier):

Lemma.

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. Suppose sequential Rademacher complexity exhibits $n^{1/r}$ growth, $r \in (1, 2]$. For any $p < r$,

$$\mathbb{E}\left|\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t)\right| \leq C_{r,p}\ \mathbb{E}\left[\left(\sum_{t=1}^{n} \sup_{f\in\mathcal{F}} |f(\mathbf{x}_t)|^p\right)^{1/p}\right].$$

Further, if $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$, then

$$\mathbb{E}\left|\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t)\right| \leq C \log n \cdot \mathbb{E}\left[\left(\sum_{t=1}^{n} \sup_{f\in\mathcal{F}} |f(\mathbf{x}_t)|^r\right)^{1/r}\right]$$

In spirit of:

if can prove $\mathbb{E}\left\|\sum Z_t\right\| \leq \sqrt{n}$ then $\mathbb{E}\left\|\sum Z_t\right\| \leq \sqrt{\sum \mathbb{E}\left\|Z_t\right\|^2}$

Definition.

We say $\mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$ has martingale type p if $\exists$ C such that

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \sum_{t=1}^{n} (g(Z_t) - \mathbb{E}_{t-1}[g(Z_t)])\right] \le C\,\mathbb{E}\left(\sum_{t=1}^{n} \mathbb{E}'_{t-1} \sup_{g\in\mathcal{G}} |g(Z_t) - g(Z'_t)|^p\right)^{1/p}$$

Theorem.

For any $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$,

1. If sequential Rademacher exhibits $n^{1/r}$ growth, $r \in (1,2]$, then $\mathcal{G}$ has martingale type p for every $p < r$.
2. If $\mathcal{G}$ has martingale type p, then sequential Rademacher exhibits $n^{1/p}$ growth.

Finer analysis for type 2

Define

$$\mathbf{Var} = \sum_{t=1}^{n} \sup_{f \in \mathcal{F}} f(\mathbf{x}_t)^2, \quad \mathbf{Var}(f) = \sum_{t=1}^{n} f(\mathbf{x}_t)^2$$

Whenever $\log \mathcal{N}^{\text{seq}}(\alpha) \leq \alpha^{-q}$, $q \in [0, 2]$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t) - C \, (\mathbf{Var}^{1/2})^{\frac{q}{4}} (\mathbf{Var}^{1/2}(f))^{\frac{2-q}{4}}\right] \leq 0$$

High probability via amplification.

Finer analysis for type 2

Define

$$\mathbf{Var} = \sum_{t=1}^{n} \sup_{f \in \mathcal{F}} f(\mathbf{x}_t)^2, \quad \mathbf{Var}(f) = \sum_{t=1}^{n} f(\mathbf{x}_t)^2$$

Whenever $\log \mathcal{N}^{\mathrm{seq}}(\alpha) \le \alpha^{-q}$, $q \in [0,2]$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t) - C \left(\mathbf{Var}^{1/2}\right)^{\frac{q}{4}} \left(\mathbf{Var}^{1/2}(f)\right)^{\frac{2-q}{4}}\right] \le 0$$

High probability via amplification.

Compare to (Massart, Rossignol '13): *weak variance* improvement of Nemirovskii inequality: for i.i.d. zero mean $Z_1, \ldots, Z_n \in \mathbb{R}^d$:

$$\mathbb{E}\left[\max_{j \le d} \left|\sum_{t=1}^{n} \epsilon_t Z_{t,j}\right|\right] \le \sqrt{2 \ln(2d) \mathbb{E} \max_{j \le d} \sum_{t=1}^{n} Z_{t,j}^2}.$$

We match index j on both sides; extend to martingales beyond finite case.

Conclusions

- Equivalence of deterministic regret inequalities and martingale tail bounds
  - gives a way of proving tail bounds (for martingales or i.i.d.) by exhibiting a method or certifying its existence
  - amplification to high probability
- Use it to extend notion of martingale type to general classes
- Not in this talk: *data-dependent bounds for online learning*

Open questions

- What is behind the equivalence?
- Replace

$$\mathbb{E}\left(\sum_{t=1}^{n} \mathbb{E}'_{t-1} \sup_{g\in\mathcal{G}} |g(Z_t) - g(Z'_t)|^p\right)^{1/p}$$

with

$$\mathbb{E}\left(\sum_{t=1}^{n} \sup_{g\in\mathcal{G}} |g(Z_t) - \mathbb{E}'_{t-1} g(Z'_t)|^p\right)^{1/p}$$

- If sequential Rademacher complexity exhibits $n^{1/r}$ growth rate, then does $\mathcal{G}$ have martingale type r? We only prove martingale type p for any $p < r$.

Outline

Reverse Hölder principle

For $p \in (0, \infty)$, define $\|Z\|_{p,\infty} = (\sup_{t>0} t^p \mathbf{P}(Z > t))^{1/p}$

Lemma (Pisier)**.**

For any $\delta \in (0, 1)$ and any R there exists $C_p(\delta, R)$ s.t. the following holds. For i.i.d. $(Z_i)_{i \geq 0}$, if

$$\sup_{N \geq 1} \mathbf{P}\left(\sup_{i \leq N} N^{-1/p} Z_i > R\right) \leq \delta$$

then

$$\|Z\|_{p,\infty} \leq C_p(\delta, R)$$

Corollary: For any $0 < q < p < \infty$ there exists $C_{p,q}$ such that

$$\|Z\|_{p,\infty} \leq C_{p,q} \sup_{N \geq 1} \left\| N^{-1/p} \sup_{i \leq N} Z_i \right\|_q$$