Symmetrization for weak interactions

December 8, 2018

1 Introduction

The purpose of this note is to extend some bounds for the expected suprema of empirical processes to a more general, nonlinear setting. Let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent random variables with values in $\mathcal{X}$, $\mathbf{X}'$ iid to $\mathbf{X}$, and let $\mathcal{H}$ be a finite class of functions $h : \mathcal{X} \to [0, 1]$. For $\mathbf{x} \in \mathcal{X}^n$ and $h \in \mathcal{H}$ we use $\mathbf{x}_h$ to denote the vector $\mathbf{x}_h = (h(x_1), ..., h(x_n)) \in [0, 1]^n$ and $\mathcal{H}(\mathbf{x}) = \{\mathbf{x}_h : h \in \mathcal{H}\} \subseteq [0, 1]^n$. Now let $f_0 : [0, 1]^n \to \mathbb{R}$ be the arithmetic mean

$$f_0(s_1, ..., s_n) := \frac{1}{n}\sum_{i=1}^n s_i \text{ for } s_i \in [0, 1].$$

Then it is not hard to show that

$$\mathbb{E}\sup_{h\in\mathcal{H}} \left[\mathbb{E}_{\mathbf{X}'} f_0(\mathbf{X}'_h) - f_0(\mathbf{X}_h)\right] \leq \frac{2}{n}\mathbb{E}R(\mathcal{H}(\mathbf{X})) \leq \frac{\sqrt{2\pi}}{n}\mathbb{E}G(\mathcal{H}(\mathbf{X})), \quad (1)$$

where the Rademacher and Gaussian averages of a subset $Y \subseteq \mathbb{R}^n$ are

$$R(Y) = \mathbb{E}\sup_{\mathbf{y}\in Y}\langle\boldsymbol{\epsilon}, \mathbf{y}\rangle \text{ and } G(Y) = \mathbb{E}\sup_{\mathbf{y}\in Y}\langle\boldsymbol{\gamma}, \mathbf{y}\rangle.$$

Here $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ and $\gamma = (\gamma_1, ..., \gamma_n)$ are vectors of independent Rademacher and standard normal variables respectively.

The symmetrization inequalites (1) have proven very useful in statistical learning theory ([1], [3] and many follow-up references). For a partial extension to functions other than the arithmetic mean we make the following definition.

Definition 1 *Suppose $f : \mathcal{X}^n \to \mathbb{R}$. For $k \in \{1, ..., n\}$ and $y, y' \in \mathcal{X}$, define the k-th partial difference operator as*

$$D^k_{yy'}f(\mathbf{x}) = f(..., x_{k-1}, y, x_{k+1}, ...) - f(..., x_{k-1}, y', x_{k+1}, ...), \textit{ for } \mathbf{x} \in \mathcal{X}^n.$$

If $\mathcal{X} \subseteq \mathbb{R}^m$ define seminorms M_{Lip} and J_{Lip} on the vector space of real functions $f : \mathcal{X}^n \to \mathbb{R}$ by

$$\begin{aligned} M_{Lip}(f) &= \max_k \sup_{\mathbf{x}\in\mathcal{X}^n, y\neq y'\in\mathcal{X}} \frac{D^k_{yy'} f(\mathbf{x})}{\|y-y'\|} \text{ and} \\ J_{Lip}(f) &= n \max_{k\neq l} \sup_{\mathbf{x}\in\mathcal{X}^n, y\neq y', z, z'\in\mathcal{X}} \frac{D^l_{zz'} D^k_{yy'} f(\mathbf{x})}{\|y-y'\|}. \end{aligned}$$

The following is a nonlinear substitute for (1).

Theorem 2 *Let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent random variables with values in $\mathcal{X}$, $\mathbf{X}'$ iid to $\mathbf{X}$, let $\mathcal{H}$ be a finite class of functions $h : \mathcal{X} \to [0,1]$ and let $f : [0,1]^n \to \mathbb{R}$. Then*

$$\mathbb{E} \sup_{h\in\mathcal{H}} \left[\mathbb{E}_{\mathbf{X}'} f(\mathbf{X}'_h) - f(\mathbf{X}_h)\right] \leq \sqrt{J_{Lip}(f)^2 + 2M_{Lip}(f)^2} \left(64\ \mathbb{E}_{\mathbf{X}} G(\mathcal{H}(\mathbf{X})) + 27\sqrt{n}\right) \tag{2}$$

and

$$\mathbb{E} \sup_{h\in\mathcal{H}} \left[\mathbb{E}_{\mathbf{X}'} f(\mathbf{X}'_h) - f(\mathbf{X}_h)\right] \leq 160\sqrt{J_{Lip}(f)^2 + 2M_{Lip}(f)^2}\ \mathbb{E}_{\mathbf{X}} G(\mathcal{H}(\mathbf{X})).$$

Some remarks:

1. For the arithmetic mean f_0 it is easy to see that $M_{Lip}(f_0) = 1/n$ and $J_{Lip}(f_0) = 0$, so the second inequality (which is typically weaker than the first) recovers the Gaussian part of (1) up to a constant.

2. The constants 64, 27 and 160 are certainly not optimal. Instead they are fruits of a laborious struggle to control the constants in Talagrands majorizing measure theorem.

3. We only get the symmetrization inequality for the Gaussian width $G(\mathcal{H}(\mathbf{X}))$. It is however standard to bound this in terms of Rademacher averages with an additional factor of $\sqrt{\ln(n+1)}$.

The utility of the above depends on how the seminorms M_{Lip} and J_{Lip} can be controlled for the function f in question. If $M_{Lip}(f) \leq a/n$ and $J_{Lip}(f) \leq b/n$ for constants a and b, we call f a *function of weak Lipschitz interaction (WLI)*. The class of WLI-functions is related to the class of weakly interacting functions in [4]. Since by assumption $f : [0,1]^n \to \mathbb{R}$ we have $M(f) \leq M_{Lip}(f)$ and $J(f) \leq J_{Lip}(f)$, where M and J are the seminorms introduced in [4]. If f is WLI the bounded difference inequality immediately yields the following.

Corollary 3 *Under the conditions of Theorem 2 let f be (a,b)-WLI. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$ in the draw of $\mathbf{X}$, we have for every $h \in \mathcal{H}$ that*

$$\mathbb{E}\left[f(\mathbf{X}_h)\right] \leq f(\mathbf{X}_h) + \sqrt{a^2+b^2}\left(\frac{64\mathbb{E}G(\mathcal{H}(\mathbf{X}))}{n} + \frac{27}{\sqrt{n}}\right) + a\sqrt{\frac{\ln(1/\delta)}{n}}.$$

This extends the popular generalization bound in [1] to WLI-functions. For the function f we may take U- or V-statistics with Lipschitz kernels, Lipschitz L-statistics (such as smoothed quantiles or medians) and other weakly interacting functions as in [4].

2 Proof of Theorem 2

The idea of the proof hinges on a result due to Michel Talagrand (see Theorem 15 in Talagrand 1987 or Theorem 2.1.5 in Talagrand 2005). It is a consequence of the celebrated majorizing measure theorem (see e.g. Talagrand 1992).

Theorem 4 *Let X_y be a random process with zero mean, indexed by a finite set $Y \subset \mathbb{R}^m$. Suppose that for any distinct members $y, y' \in Y$ and any $t > 0$*

$$\Pr\left\{X_y - X_{y'} > t\right\} \leq \exp\left(\frac{-t^2}{2\left\|y - y'\right\|^2}\right) \tag{3}$$

Then

$$\mathbb{E}\sup_{y\in Y} X_y \leq c\ G(Y)$$

where c is a universal constant.

Unfortunately the constant c which results from the proof is very large. Nevertheless, as remarked in (Talagrand 1987), if X is a Gaussian process, then Theorem 4 reduces to Slepian's Lemma (Boucheron et al 2013), which inspires the tantalizing conjecture that the optimal c could be in the order of unity.

To get reasonable constants we are forced to use the following constant-conscious variant of Theorem 4, which will be proven in the next section.

Theorem 5 *Suppose that $\{X_y\}_{y\in Y}$ is as in Theorem 4. For $y_0 \in A \subseteq Y$ let Z_{A,y_0}be the random variable*

$$Z_{A,y_0} = \sup_{y\in A} X_y - X_{y_0}.$$

Suppose that for every $A \subseteq Y$ and $y_0 \in A$

$$\Pr\left\{Z_{A,y_0} - \mathbb{E}\left[Z_{A,y_0}\right] > t\right\} \leq \exp\left(\frac{-t^2}{2\sup_{y\in A}\left\|y - y_0\right\|^2}\right). \tag{4}$$

Then

$$\mathbb{E}\left[\sup_{y\in Y} X_y\right] \leq 18G(Y) + 11\ diam(Y).$$

We also use the following concentration inequality (see e.g. [2].

Theorem 6 *Let $(\mathcal{X},\mu)$ be a probability space and $f:\mathcal{X}^n\rightarrow\mathbb{R}$ and define $V^+(f):\mathcal{X}^n\rightarrow\mathbb{R}$ by*

$$V^+(f)(x)=\sum_{k=1}^n\mathbb{E}_{y\sim\mu}\left[\left(f(x)-f\left(S_y^kx\right)\right)_+^2\right]$$

(here S_y^kx replaces the k-th coordinate of x by y).Then

$$\Pr\{f-\mathbb{E}f>t\}\leq\exp\left(\frac{-t^2}{2\sup_{x\in\mathcal{X}^n}V^+(f)(x)}\right).$$

With Theorem 5 and Theorem 6 we can prove the following intermediate result about processes defined on the Bernoulli cube $\{0,1\}^n$.

Theorem 7 *Let μ be the uniform probability measure on $\{0,1\}^n$. Let $Y\subseteq\mathbb{R}^m$ be finite and suppose that for every $y\in Y$ there is a function $X_y:\{0,1\}^n\rightarrow\mathbb{R}$ such that $\mathbb{E}_\mu[X_y]=0$, and for all $y,y'\in Y$ and $\sigma\in\{0,1\}^n$*

$$\sum_{k=1}^n D_{10}^k\left(X_y(\sigma)-X_{y'}(\sigma)\right)^2\leq 2\left\|y-y'\right\|^2. \qquad (5)$$

Then

$$\mathbb{E}_\mu\left[\max_{y\in Y}X_y\right]\leq 18G\ \ (Y)+11\ \mathit{diam}(Y)\leq 46G(Y).$$

Note that (5) and the bounded difference inequality ([2]) immediately give the subgaussian condition (3), so if we don't care about constants the result follows directly from Talagrand's Theorem 4. Also note that the last inequality follows from $\mathrm{diam}(Y)\leq\sqrt{2\pi}\,G(Y)$.

Proof. We just need to show that the X_y satisfy the hypotheses of Theorem 5. Let $A\subseteq Y,y'\in A$ and let $Z:\{0,1\}^n\rightarrow\mathbb{R}$ be defined by $Z(\sigma)=\max_{y\in A}X_y(\sigma)-X_{y'}(\sigma)$. Now fix σ for the moment and let y^* the maximizer in the definition of $Z(\sigma)$. Then, letting $\sigma^{(k)}$ be equal to σ with the k-th bit flipped,

$$\begin{aligned}
V^+(Z)(\sigma) &= \sum_{k=1}^n E_{y\in\{0,1\}}\left[\left(Z(\sigma)-S_y^kZ(\sigma)\right)_+^2\right]\\
&\leq \frac{1}{2}\sum_k\left(Z(\sigma)-Z\left(\sigma^{(k)}\right)\right)_+^2\\
&\leq \frac{1}{2}\sum_k\left(X_{y^*}(\sigma)-X_{y'}(\sigma)-\left(X_{y^*}\left(\sigma^{(k)}\right)-X_{y'}\left(\sigma^{(k)}\right)\right)\right)^2\\
&= \frac{1}{2}\sum_{k=1}^n D_{10}^k\left(X_{y^*}(\sigma)-X_{y'}(\sigma)\right)^2\leq\left\|y-y'\right\|^2.
\end{aligned}$$

It follows from Theorem 6 that

$$\Pr_{\sigma}\left\{\max_{y\in A} X_y(\sigma) - X_{y'}(\sigma) - \mathbb{E}\left[\max_{y\in A} X_y\right] > t\right\} \le \exp\left(\frac{-t^2}{2\left\|y-y'\right\|^2}\right)$$

and Theorem 5 gives the conclusion. ■

Finally we need the following Lemma.

Lemma 8 *Suppose $f:\mathcal{X}^n\to\mathbb{R}$ where $\mathcal{X}\subseteq\mathbb{R}$. Then for $x,y\in\mathcal{X}^n$ and $a,b\in\mathcal{X}$*

$$D^k_{a,b}f(\mathbf{x}) - D^k_{a,b}f(\mathbf{y}) \le \frac{J_{Lip}(f)\left\|\mathbf{x}-\mathbf{y}\right\|}{\sqrt{n}}.$$

Proof. First assume $k=1$ and for $j\in\{2,...,n+1\}$ define $\mathbf{z}^j\in\mathcal{X}^n$ by

$$z^j_i = \left\{\begin{array}{lll} x_i & \text{if} & i<j \\ y_i & \text{if} & i\ge j \end{array}\right.,$$

so that $\mathbf{z}^1=\mathbf{y}$ and $\mathbf{z}^{n+1}=\mathbf{x}$. Then using Cauchy-Schwarz

$$\begin{aligned} D^1_{a,b}f(\mathbf{x}) - D^1_{a,b}f(\mathbf{y}) &= \sum_{j=2}^n D^1_{a,b}f\left(\mathbf{z}^{j+1}\right) - D^1_{a,b}f\left(\mathbf{z}^j\right) = \sum_{j=2}^n D^1_{a,b}D^j_{x_jy_j}f\left(\mathbf{z}^j\right) \\ &\le \frac{J_{Lip}(f)}{n}\sum_{j=2}^n\left|x_j-y_j\right| \\ &\le \frac{J_{Lip}(f)}{\sqrt{n}}\left\|\mathbf{x}-\mathbf{y}\right\|. \end{aligned}$$

If $k\neq 1$ let f_π be the function $f_\pi(x)=f(\pi x)$, where π is the permutation exchanging the first and the k-th argument, and apply the above to f_π. ■

Proof of Theorem 2. In this proof we use $\mathbf{vw}$ to denote the vector

$$\mathbf{vw} = (v_1w_1,...,v_nw_n) \text{ with } \mathbf{v},\mathbf{w}\in\mathbb{R}^n.$$

Let Q be the left hand side of (2). Initially our proof parallels the standard symmetrization argument: we pull the second expectation outside the supremum

$$Q \le \mathbb{E}_{\mathbf{XX}'}\sup_{h\in\mathcal{H}}\left[f(\mathbf{X}_h) - f(\mathbf{X}'_h)\right].$$

Since X_i and X'_i are iid, the last quantity does not change if we exchange X_i and X'_i on an arbirary subset of indices i. If $\boldsymbol{\sigma}\in\{0,1\}^n$ is such that σ_i is zero on this set and one on its complement, we obtain

$$\begin{aligned} Q &\le \mathbb{E}_{\mathbf{XX}'}\sup_{h\in\mathcal{H}}\left[f\left(\boldsymbol{\sigma}\mathbf{X}_h + (\mathbf{1}-\boldsymbol{\sigma})\mathbf{X}'_h\right) - f\left(\boldsymbol{\sigma}\mathbf{X}'_h + (\mathbf{1}-\boldsymbol{\sigma})\mathbf{X}_h\right)\right] \\ &= \mathbb{E}_{\mathbf{XX}'}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{h\in\mathcal{H}}\left[f\left(\boldsymbol{\sigma}\mathbf{X}_h + (\mathbf{1}-\boldsymbol{\sigma})\mathbf{X}'_h\right) - f\left(\boldsymbol{\sigma}\mathbf{X}'_h + (\mathbf{1}-\boldsymbol{\sigma})\mathbf{X}_h\right)\right]. \end{aligned}$$

In the last step we took the expectation over configurations $\boldsymbol{\sigma}$ chosen uniformly from $\{0,1\}^n$. We now condition on the X_i and X_i', which we temporarily replace by lower case letters. For $h \in \mathcal{H}$ denote $\mathbf{h} := (\mathbf{x}_h, \mathbf{x}_h') \in \mathbb{R}^{2n}$ and define the set $T \subseteq \mathbb{R}^{2n}$ by

$$T = \{\mathbf{h} : h \in \mathcal{H}\}.$$

Note that $\operatorname{diam}(T) \leq \sqrt{\operatorname{diam}(\mathcal{H}(\mathbf{x}))^2 + \operatorname{diam}(\mathcal{H}(\mathbf{x}'))^2} \leq \sqrt{2n}$ and that $G(T) \leq G(\mathcal{H}(\mathbf{x})) + G(\mathcal{H}(\mathbf{x}'))$.

Now consider the random process indexed by T

$$Y_{\mathbf{h}}(\boldsymbol{\sigma}) = f\left(\boldsymbol{\sigma}\mathbf{x}_h + (\mathbf{1}-\boldsymbol{\sigma})\,\mathbf{x}_h'\right) - f\left(\boldsymbol{\sigma}\mathbf{x}_h' + (\mathbf{1}-\boldsymbol{\sigma})\,\mathbf{x}_h\right).$$

Clearly $\mathbb{E}_{\boldsymbol{\sigma}} Y_{\mathbf{h}}(\boldsymbol{\sigma}) = 0$ for all $h \in \mathcal{H}$. Now we want to apply Theorem 7 and seek to prove, for fixed $h, g \in \mathcal{H}$ a bounded difference condition as in (5) on the random variable $Y_{\mathbf{h}} - Y_{\mathbf{g}}$.

Let $Z(\sigma) := Y_{\mathbf{h}}(\boldsymbol{\sigma}) - Y_{\mathbf{g}}(\boldsymbol{\sigma})$ and fix a configuration $\boldsymbol{\sigma} \in \{0,1\}^n$. We define vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in [0,1]^n$ by

$$\begin{aligned}
\mathbf{a} &= \boldsymbol{\sigma}\mathbf{x}_h + (\mathbf{1}-\boldsymbol{\sigma})\,\mathbf{x}_h' \\
\mathbf{b} &= \boldsymbol{\sigma}\mathbf{x}_g + (\mathbf{1}-\boldsymbol{\sigma})\,\mathbf{x}_g' \\
\mathbf{c} &= \boldsymbol{\sigma}\mathbf{x}_h' + (\mathbf{1}-\boldsymbol{\sigma})\,\mathbf{x}_h \\
\mathbf{d} &= \boldsymbol{\sigma}\mathbf{x}_g' + (\mathbf{1}-\boldsymbol{\sigma})\,\mathbf{x}_g.
\end{aligned}$$

Then $Y_{\mathbf{h}}(\boldsymbol{\sigma}) = f(\mathbf{a}) - f(\mathbf{c})$, $Y_{\mathbf{g}}(\boldsymbol{\sigma}) = f(\mathbf{b}) - f(\mathbf{d})$. Also note that

$$\|\mathbf{a}-\mathbf{b}\|^2 + \|\mathbf{c}-\mathbf{d}\|^2 = \|\mathbf{h}-\mathbf{g}\|^2. \tag{6}$$

For any $k \in \{1, ..., n\}$

$$\begin{aligned}
& D_{1,0}^k Z(\boldsymbol{\sigma}) \\
= \; & D^k_{h(x_k),h(x_k')} f(\mathbf{a}) - D^k_{g(x_k),g(x_k')} f(\mathbf{b}) + D^k_{h(x_k),h(x_k')} f(\mathbf{c}) - D^k_{g(x_k),g(x_k')} f(\mathbf{d}) \\
= \; & D^k_{h(x_k),h(x_k')} \left(f(\mathbf{a}) - f(\mathbf{b})\right) + D^k_{h(x_k),h(x_k')} \left(f(\mathbf{c}) - f(\mathbf{d})\right) \\
& + D^k_{h(x_k),h(x_k')} f(\mathbf{b}) - D^k_{g(x_k),g(x_k')} f(\mathbf{b}) + D^k_{h(x_k),h(x_k')} f(\mathbf{d}) - D^k_{g(x_k),g(x_k')} f(\mathbf{d}) \\
= \; & D^k_{h(x_k),h(x_k')} \left(f(\mathbf{a}) - f(\mathbf{b})\right) + D^k_{h(x_k),h(x_k')} \left(f(\mathbf{c}) - f(\mathbf{d})\right) \\
& + D^k_{h(x_k),g(x_k)} f(\mathbf{b}) - D^k_{h(x_k'),g(x_k')} f(\mathbf{b}) + D^k_{h(x_k),g(x_k)} f(\mathbf{d}) - D^k_{h(x_k'),g(x_k')} f(\mathbf{d}),
\end{aligned}$$

where the identity $D^k_{y,y'} f(\mathbf{x}) - D^k_{z,z'} f(\mathbf{x}) = D^k_{y,z} f(\mathbf{x}) - D^k_{y',z'} f(\mathbf{x})$ was used in the last equality. Using Jensens inequality (which is responsible for the factor

1/6) we get

$$
\begin{aligned}
& \frac{1}{6}\left(D^k_{1,0} Z\left(\boldsymbol{\sigma}\right)\right)^2 \\
\leq\ & \left(D^k_{h(x_k),h\left(x'_k\right)}\left(f\left(\mathbf{a}\right)-f\left(\mathbf{b}\right)\right)\right)^2 + \left(D^k_{h(x_k),h\left(x'_k\right)}\left(f\left(\mathbf{c}\right)-f\left(\mathbf{d}\right)\right)\right)^2 + \\
& + \left(D^k_{h(x_k),g(x_k)} f\left(\mathbf{b}\right)\right)^2 + \left(D^k_{h\left(x'_k\right),g\left(x'_k\right)} f\left(\mathbf{b}\right)\right)^2 + \\
& + \left(D^k_{h(x_k),g(x_k)} f\left(\mathbf{d}\right)\right)^2 + \left(D^k_{h\left(x'_k\right),g\left(x'_k\right)} f\left(\mathbf{d}\right)\right)^2 \\
\leq\ & \frac{J_{Lip}\left(f\right)^2 \left\|\mathbf{h}-\mathbf{g}\right\|^2}{n} + 2M_{Lip}\left(f\right)^2\left(\left(h\left(x_k\right)-g\left(x_k\right)\right)^2 + \left(h\left(x'_k\right)-g\left(x'_k\right)\right)^2\right),
\end{aligned}
\tag{7}
$$

Where we bounded the first two terms using Lemma 8 and (6), and used the definition of M_{Lip} for the remaining terms. Summing over k we get

$$\sum_{k=1}^{n}\left(D^k_{1,0}Z\left(\boldsymbol{\sigma}\right)\right)^2 \leq 3\left(J_{Lip}\left(f\right)^2 + 2M_{Lip}\left(f\right)^2\right)\times 2\left\|\mathbf{h}-\mathbf{g}\right\|^2.$$

Rescaling and applying Theorem 7 gives

$$
\begin{aligned}
\mathbb{E}_\sigma \sup_{h\in\mathcal{H}} Y_{\mathbf{h}} &\leq \sqrt{3\left(J_{Lip}\left(f\right)^2 + 2M_{Lip}\left(f\right)^2\right)}\left(18G\ \ (T) + 11\ \mathrm{diam}\left(T\right)\right) \\
&\leq \sqrt{J_{Lip}\left(f\right)^2 + 2M_{Lip}\left(f\right)^2}\left(32\left(G\left(\mathcal{H}\left(\mathbf{x}\right)\right) + G\left(\mathcal{H}\left(\mathbf{x}'\right)\right)\right) + 11\sqrt{6n}\right).
\end{aligned}
$$

We now remove the conditioning and return to the X_i-variables, to get

$$
\begin{aligned}
Q &\leq \mathbb{E}_{XX'}\mathbb{E}_\sigma \sup_{h\in\mathcal{H}} Y_{\mathbf{h}} \\
&\leq \sqrt{J_{Lip}\left(f\right)^2 + 2M_{Lip}\left(f\right)^2}\,\mathbb{E}_{\mathbf{XX}'}\left(32\left(G\left(\mathcal{H}\left(\mathbf{X}\right)\right) + G\left(\mathcal{H}\left(\mathbf{X}'\right)\right)\right) + 11\sqrt{6n}\right) \\
&= \sqrt{J_{Lip}\left(f\right)^2 + 2M_{Lip}\left(f\right)^2}\left(64\ \mathbb{E}_{\mathbf{X}} G\left(\mathcal{H}\left(\mathbf{X}\right)\right) + 27\sqrt{n}\right).
\end{aligned}
$$

The second inequality is obtained by using the second inequality of Theorem 7.
■

3 Proof of the subgaussian-gaussian comparison, Theorem 5

In this section we give a proof of Theorem 5. All the relevant ideas are taken from Talagrand's proof [7] of the majorizing measure theorem. The only contribution here is to keep the constants small, by simultaneously considering the subgaussian upper, and the Gaussian lower bound. First we need some standard minorization results for Gaussian processes.

Lemma 9 *Let $\gamma_1,...,\gamma_n$ be independent standard normal variables. Then*

$$\mathbb{E}\max_i \gamma_i \geq \sqrt{2\ln n} - 2.$$

Proof. We may assume $n > 7$, the conclusion being immediate otherwise. We have

$$\begin{aligned}
\mathbb{E}\max_i \gamma_i &\geq \int_0^\delta \Pr\left\{\max_i \gamma_i > t\right\} dt \\
&= \int_0^\delta \left(1 - \left(1 - \Pr\left\{\gamma_i > t\right\}\right)^n\right) dt \\
&\geq \int_0^\delta \left(1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{e^{-\delta^2/2}}{\delta + 1/\delta}\right)^n\right) dt \\
&\geq \delta\left(1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{e^{-\delta^2/2}}{\delta + 1/\delta}\right)^n\right) \\
&\geq \delta\left(1 - \exp\left(-n\frac{1}{\sqrt{2\pi}} \frac{e^{-\delta^2/2}}{\delta + 1/\delta}\right)\right).
\end{aligned}$$

In the third line we used a standard approximation and in the last line we used $1 - x \leq e^x$. Now set $\delta = \sqrt{2\ln n} - 1$. Since $n \geq 8$ we have $1 \leq \sqrt{2\ln n} - 1$ so we obtain

$$\begin{aligned}
\mathbb{E}\max_i \gamma_i &\geq \left(\sqrt{2\ln n} - 1\right)\left(1 - \exp\left(-\frac{1}{\sqrt{2e\pi}} \frac{e^{\sqrt{2\ln n}}}{\sqrt{2\ln n}}\right)\right) \\
&\geq \sqrt{2\ln n} - 1 - \max_{x>0}\,(x-1)\exp\left(-\frac{1}{\sqrt{2e\pi}}\frac{e^x}{x}\right) \\
&\geq \sqrt{2\ln n} - 1 - \max_{x>0}\,(x-1)\exp\left(-\frac{1}{\sqrt{e\pi}}\frac{x^2}{2}\right)
\end{aligned}$$

Where we used $e^x/x \geq x^2/\sqrt{2}$. From calculus we find for $(x-1)\,e^{-ax^2/2}$ the maximizer $x = \frac{1}{2}\left(1 + \sqrt{1 + 4/a}\right)$. Resubstitution and using $a = 1/\sqrt{e\pi}$ we get

$$(x-1)\exp\left(-\frac{1}{\sqrt{e\pi}}\frac{x^2}{2}\right) \leq 0.526 \leq 1,$$

so $\mathbb{E}\max_i \gamma_i \geq \sqrt{2\ln n} - 2$. ■

The following is our version of Sudakov Minoration

Lemma 10 *Let $T = \{t_1, ..., t_N\} \subseteq \mathbb{R}^d$ be finite with $\|t_k - t_l\| \geq r$ for $k \neq l$. Then*

$$G(T) \geq r\sqrt{\ln N} - \sqrt{2}r$$

Proof. Let $T' = \{e_1, ..., e_N\}$ be an orthonormal basis for $\mathbb{R}^N$. Then

$$\mathbb{E} \left(\langle \gamma, e_k \rangle - \langle \gamma, e_l \rangle \right)^2 \leq \frac{2}{r^2} \mathbb{E} \left(\langle \gamma, t_k \rangle - \langle \gamma, t_l \rangle \right)^2 ,$$

whence by Slepian's Lemma

$$G(T) \geq \frac{r}{\sqrt{2}} G(T') \geq r\sqrt{\ln n} - \sqrt{2}r,$$

where Lemma 9 was used in the second inequality. ■

Lemma 11 *Suppose $t_k \in T \subseteq \mathbb{R}^d$ for $1 \leq k \leq N$, that $\|t_k - t_l\| \geq \sqrt{8}r$ for $t \neq l$ and that $A_k \subseteq B(t_k, r)$. Then*

$$G\left(\bigcup_l A_l \right) \geq r\sqrt{2 \ln N} + \min_k G(A_k) - 4r.$$

Proof. For $k \in \{1, ..., N\}$ define a random variable $Y_k = \sup_{t \in A_k} \langle \gamma, t - t_k \rangle$. The map $z \in \mathbb{R}^d \mapsto \sup_{t \in A_k} \langle z, t - t_k \rangle$ is Lipschitz with Lipschitz constant $\sup_{t \in A_k} \|t - t_k\| \leq r$. So by Gaussian concentration (Theorem 5.5 in [2]) for every $\lambda > 0$

$$\mathbb{E} \exp\left(\lambda \left(\mathbb{E}Y_k - Y_k \right) \right) \leq \exp\left(\frac{\lambda^2 r^2}{2} \right).$$

So, from Jensen's inequality,

$$\exp\left(\lambda \mathbb{E} \max_k \left(\mathbb{E}Y_k - Y_k \right) \right) \leq \sum_k \mathbb{E} \exp\left(\lambda \left(\mathbb{E}Y_k - Y_k \right) \right) \leq N \exp\left(\frac{\lambda^2 r^2}{2} \right).$$

Taking the logarithm and dividing by λ and substituting $\lambda = r^{-1}\sqrt{2 \ln N}$ we obtain

$$\mathbb{E} \max_k \left(\mathbb{E}Y_k - Y_k \right) \leq \frac{\lambda r^2}{2} + \frac{1}{\lambda} \ln N = r\sqrt{2 \ln N}. \tag{8}$$

Since $\|t_k - t_l\| \geq \sqrt{8}r$, Sudakov Minoration (Lemma 10) gives

$$\mathbb{E} \max_k \langle \gamma, t_k \rangle = G(\{t_1, ..., t_N\})) \geq 2r\sqrt{2 \ln N} - 4r. \tag{9}$$

Thus

$$\begin{aligned} G\left(\bigcup_k A_k \right) &= \mathbb{E} \max_k \sup_{t \in A_k} \langle \gamma, t \rangle \\ &= \mathbb{E} \max_k \left[\langle \gamma, t_k \rangle - \left(\mathbb{E}Y_k - Y_k \right) + \mathbb{E}Y_k \right] \\ &\geq \mathbb{E} \max_k \left[\langle \gamma, t_k \rangle - \left(\mathbb{E}Y_k - Y_k \right) \right] + \min_l G(A_l) \\ &\geq \mathbb{E} \max_k \langle \gamma, t_k \rangle - \mathbb{E} \max_k \left(\mathbb{E}Y_k - Y_k \right) + \min_l G(A_l) \\ &\geq \left(2r\sqrt{2 \ln N} - 4r \right) - r\sqrt{2 \ln N} + \min_l G(A_l) \\ &= r\sqrt{2 \ln N} + \min_k G(A_k) - 3r. \end{aligned}$$

The second identity follows from the definition of the random variable Y_k. The first inequality follows from $\mathbb{E}Y_k = G(A_k) \geq \min_l G(A_l)$. The next is the triangle inequality $\max(a-b) \geq \max a - \max b$ and linearity of the expectation. Finally we used the estimates (9) and (8). ■

Next we need a lemma exploiting subgaussian concentration

Lemma 12 *: Suppose that $X_1, ..., X_n$ are random variables satisfying*

$$\Pr\{X_i - \mathbb{E}X_i > t\} \leq e^{-t^2/2b^2}$$

for all $1 \leq i \leq n$. Then

$$\mathbb{E}\max_i X_i \leq \max_{k \leq n}\left(\mathbb{E}X_k + (9/8)\, b\sqrt{2\ln k}\right) + 5b/2$$

Proof. We can assume $b = 1$. Let

$$Z = \max_{1 \leq i \leq n} X_i - \max_{1 \leq k \leq n}\left(\mathbb{E}X_k + (9/8)\sqrt{2\ln(2+k)}\right)$$

Then for $t > 0$

$$\begin{aligned}
\Pr\{Z > t\} &= \Pr\left\{\exists i : X_i > \max_{1 \leq k \leq n}\left(\mathbb{E}X_k + (9/8)\sqrt{2\ln(2+k)}\right) + t\right\} \\
&\leq \sum_i \Pr\left\{X_i > \mathbb{E}X_i + \sqrt{2\ln(2+i)^{(9/8)^2}} + t\right\} \\
&\leq \sum_i \exp\left(\frac{-\left(\sqrt{2\ln(2+i)^{(9/8)^2}} + t\right)^2}{2}\right)
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}Z_i &\leq \int_0^\infty \Pr\{Z_i > t\}\, dt \\
&\leq \int_0^\infty \sum_{i\geq 1} \exp\left(\frac{-\left(\sqrt{2\ln(2+i)^{(9/8)^2}} + t\right)^2}{2} \right) dt \\
&= \sum_{i\geq 1} \int_{\sqrt{2\ln(2+i)^{(9/8)^2}}}^\infty \exp\left(\frac{-t^2}{2}\right) dt \\
&\leq \sum_{i\geq 1} \frac{1}{\sqrt{4\pi\ln(2+i)^{(9/8)^2}}} \exp\left(\frac{-\sqrt{2\ln(2+i)^{(9/8)^2}}^2}{2} \right) \\
&\leq \frac{1}{\sqrt{4\pi\ln(2+1)^{(9/8)^2}}} \sum_{i>1} (2+i)^{-(9/8)^2} \\
&\leq \frac{4}{9\sqrt{\pi\ln 3}} \int_0^\infty (2+s)^{-(9/8)^2}\, ds \\
&\leq 0.8,
\end{aligned}$$

where we used a standard estimate for the tail of the normal distribution in the third inequality. The result follows then by rescaling and since $\sqrt{\ln(2+k)} \leq \sqrt{\ln k} + \sqrt{\ln 3}$ and $0.8 + (9/8)\sqrt{2\ln 3} \leq 5/2$. ■

Recall the definition of the random variable

$$Z_{A,y_0} := \sup_{y\in A} X_y - X_{y_0},$$

for any $A \subseteq \mathbb{R}^n$ and $y_0 \in \mathbb{R}^n$, and note that by zero mean $\mathbb{E}Z_{A,y_1} = \mathbb{E}Z_{A,y_2}$ for any $y_1, y_2 \in \mathbb{R}^n$. If $A \subseteq \bigcup_k D_k$ and $y_k \in D_k$, then it follows from the concentration property (4) and Lemma 12 that

$$\mathbb{E}Z_{A,y_0} \leq \max_k \left(\mathbb{E}Z_{D_k,y_k} + (9/8) \sup_{y\in A} \|y - y_0\| \sqrt{2\ln k} \right) + (5/2) \sup_{y\in A} \|y - y_0\|. \tag{10}$$

Proof of Theorem 5. Let $\delta := \mathrm{diam}(Y) > 0$ and for $t \in \mathbb{R}^n$ and $s > 0$ use $B(t,s)$ to denote the scaled ball

$$B(t,s) = \{y \in \mathbb{R}^n : \|y - t\| \leq \delta s\}.$$

Throughout this proof we abbreviate $r := \sqrt{8}$. Define for $A \subseteq \mathbb{R}^n$ and $j \in \mathbb{Z}$

$$b_j(A) := \sup_{t\in A} G\left(A \cap B\left(t, r^{-j-1}\right)\right).$$

For $j \in \mathbb{Z}$ let $P(j)$ be the statement

$$\begin{aligned} \forall A &\subseteq Y, \forall t_0 \in A, \\ \left(A \subseteq B\left(t_0, r^{-j}\right)\right) &\implies \mathbb{E}Z_{A,y_0} \leq 9\left(G(A) + b_j(A)\right) + \frac{7\delta r^{-j}}{1 - r^{-1}}. \end{aligned}$$

We claim that this statement is true for all $j \in \mathbb{N}_0$.

Since Y is finite, there exists j_{sep} such that for all $j \geq j_{sep}, A \subseteq Y$ and $t_0 \in A$ the inclusion $\left(A \subseteq B\left(t_0, r^{-j}\right)\right)$ implies that $A = \{t_0\}$, so that $P(j)$ holds, because the LHS in the inequality is zero. We can thus prove $P(j)$ for smaller values of j by downward induction in j. To this end we assume $P(j+1)$ to hold and set about to prove $P(j)$.

Let $A \subseteq Y$, $t_0 \in A$ and assume that $A \subseteq B\left(t_0, r^{-j}\right)$.

Now we construct for $1 \leq k \leq p$ the points $t_k \in A$, the partition D_k and the sets $A_k \subseteq D_k$ just as in Talagrand's paper: t_1 is chosen such that $G\left(A \cap B\left(t_1, r^{-j-2}\right)\right)$ is maximal, $A_1 = A \cap B\left(t_1, r^{-j-2}\right)$ and $D_1 = A \cap B\left(t_1, r^{-j-1}\right)$. Having chosen t_l, A_l, and D_l for $l \in \{1, ..., k-1\}$ we let $H_k = \bigcup_{1 \leq l \leq k-1} D_k$. If $A \subseteq H_k$ we stop, otherwise we choose $t_k \in A \backslash H_k$ such that $G\left(A \cap B\left(t_k, r^{-j-2}\right)\right)$ is maximal, $A_k = A \cap B\left(t_k, r^{-j-2}\right)$ and $D_k = \left(A \backslash H_k\right) \cap B\left(t_k, r^{-j-1}\right)$.

The relevant properties of this construction are

$$\begin{aligned} A &\subseteq B\left(t_0, r^{-j}\right) && (11) \\ A &= \bigcup_k D_k && (12) \\ D_k &\subseteq B\left(t_k, r^{-j-1}\right) && (13) \\ A_k &\subseteq B\left(t_k, r^{-j-2}\right) && (14) \\ \|t_k - t_l\| &\geq \delta r^{-j-1} \text{ for } j \neq l && (15) \\ G\left(A_{k+1}\right) &\leq G\left(A_k\right) && (16) \\ b_{j+1}\left(D_k\right) &\leq G\left(A_k\right) && (17) \\ G\left(D_k\right) &\leq b_j(A). && (18) \end{aligned}$$

Of these all but the last two are immediate. For any $t \in D_k$

$$\begin{aligned} G\left(D_k \cap B\left(t, r^{-j-2}\right)\right) &\leq G\left(\left(A \backslash H_k\right) \cap B\left(t, r^{-j-2}\right)\right) \\ &\leq \max_{t \in (A \backslash H_k)} G\left(\left(A \backslash H_k\right) \cap B\left(t, r^{-j-2}\right)\right) \\ &\leq G\left(A_k\right). \end{aligned}$$

This gives (17). Property (18) follows from $D_k \subseteq B\left(t_k, r^{-j-1}\right)$, so $G\left(D_k\right) \leq b_j(A)$.

We now use inequality (10) and properties (11) and (12) above to get

$$\mathbb{E}Z_{A,t_0} \leq \max_k \left((9/8)\, \delta r^{-j} \sqrt{2 \ln k} + \mathbb{E}Z_{D_k, t_k}\right) + (5/2)\, \delta r^{-j} \tag{19}$$

Since $D_k \subseteq B\left(t_k, r^{-(j+1)}\right)$ by (13) we can use $P(j+1)$ with D_k and t_k. This reads

$$\begin{aligned}
\mathbb{E}Z_{D_k,t_k} &\leq 9\left(G\left(D_k\right)+b_{j+1}\left(D_k\right)\right)+\frac{7\delta r^{-j-1}}{1-r^{-1}} \\
&\leq 9\left(G\left(A_k\right)+b_j\left(A\right)\right)+\frac{(5/2)\,\delta r^{-j-1}}{1-r^{-1}}+\frac{(9/2)\,\delta r^{-j-1}}{1-r^{-1}},
\end{aligned}$$

where we used the properties (17) and (18) in the second inequality. Note that

$$\frac{r^{-j-1}}{1-r^{-1}}+r^{-j}=\frac{r^{-j}}{1-r^{-1}}, \tag{20}$$

so substitution in (19) gives

$$\begin{aligned}
&\mathbb{E}Z_{A,t_0} \\
\leq\ & \max_k\left((9/8)\,\delta r^{-j}\sqrt{2\ln k}+9\left(G\left(A_k\right)+b_j\left(A\right)\right)\right)+\frac{(5/2)\,\delta r^{-j}}{1-r^{-1}}+\frac{(9/2)\,\delta r^{-j-1}}{1-r^{-1}} \\
=\ & (9/8)\,\delta r^{-j}\sqrt{2\ln k^*}+9\left(G\left(A_{k^*}\right)+b_j\left(A\right)\right)+\frac{(5/2)\,\delta r^{-j}}{1-r^{-1}}+\frac{(9/2)\,\delta r^{-j-1}}{1-r^{-1}}, \quad (21)
\end{aligned}$$

where we passed to a maximizer k^*. By (16) the $G\left(A_k\right)$ are nonincreasing so that $G\left(A_{k^*}\right)=\min_{k\leq k^*}G\left(A_k\right)$. By properties (14) and (15) and $r=\sqrt{8}$ we can use the minoration Lemma 11 with δr^{-j-2} and k^* in place of r and N to obtain

$$G\left(A\right)\geq G\left(\bigcup_{k\leq k^*}A_k\right)\geq \delta r^{-j-2}\sqrt{2\ln k^*}+G\left(A_{k^*}\right)-4\delta r^{-j-2}$$

or

$$G\left(A_{k^*}\right)\leq G\left(A\right)-\delta r^{-j-2}\sqrt{2\ln k^*}+\frac{1}{2}\delta r^{-j}.$$

Since $(9/8)\,r^{-j}-9r^{-j-2}=0$, substitution in (21) above and using (20) gives that

$$\mathbb{E}Z_{A,t_0}\leq 9\left(G\left(A\right)+b_j\left(A\right)\right)+\frac{7\delta r^{-j}}{1-r^{-1}},$$

which is the conclusion of $P(j)$ and completes the induction.

Since $Y\subseteq B\left(y_0,1\right)$ we have that y_0 and $j=0$ satisfy the induction hypothesis and using $b_0\left(Y\right)\leq G\left(Y\right)$ we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{y\in Y}X_y\right] &= \mathbb{E}Z_{Y,y_0}\leq 18G\left(Y\right)+\frac{7\delta}{1-r^{-1}} \\
&\leq 18G\left(Y\right)+11\ \mathrm{diam}\left(Y\right).
\end{aligned}$$

■

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3: 463–482, 2002.

[2] S. Boucheron, G. Lugosi, P. Massart. *Concentration Inequalities*, Oxford University Press, 2013

[3] V. I. Koltchinskii and D. Panchenko. *Rademacher processes and bounding the risk of function learning.* In E. Gine, D. Mason, and J. Wellner, editors, *High Dimensional Probability II*, pages 443–459. 2000.

[4] Maurer, A., & Pontil, M. (2018). Empirical bounds for functions with weak interactions, COLT 20018.

[5] C.McDiarmid, *Concentration*, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, (1998) 195-248. Springer, Berlin

[6] M. Talagrand. Regularity of Gaussian processes. Acta Mathematica. 159: 99–149, 1987.

[7] M. Talagrand. A simple proof of the majorizing measure theorem. *Geometric and Functional Analysis*. Vol 2, No.1: 118–125, 1992.

[8] M. Talagrand. *The Generic Chaining. Upper and Lower Bounds for Stochastic Processes.* Springer, Berlin, 2005.