Concentration inequalities for smooth random fields [1]

Denis Belomestny and Vladimir Spokoiny
Duisburg-Essen University and WIAS Berlin, Moscow Institute of Physics and Technology

In this note we derive a sharp concentration inequality for the supremum of a smooth random field over a finite dimensional set. It is shown that this supremum can be bounded with high probability by the value of the field at some deterministic point plus an intrinsic dimension of the optimisation problem. As an application we prove the exponential inequality for a function of the maximal eigenvalue of a random matrix is proved.

1 Introduction

Concentration inequalities is a quite active field of research, which is driven by numerous applications, see Ledoux (2001) and Lugosi (2005) for an overview. Concentration inequalities have been used in many fields of both pure and applied mathematics, including stochastic optimization, random matrix theory, geometric functional analysis, randomized algorithms, statistics, machine learning and compressed sensing. A typical situation where the concentration inequalities are useful is the case where one is interested in probabilistic bounds for a random variable which is the solution of a (stochastic) optimization problem. This type of problems appear in statistics and stochastic optimisation. Many statistical estimators (e.g. the maximum-likelihood estimator) are solutions to random optimization problems. There is a substantial statistical literature dealing with concentration in statistics, see Massart (2000) for an overview. On the stochastic optimisation side let us mention the bin packing problem and the travelling salesman problem where the concentration approach leads to rather sharp probabilistic bounds for the quantities of interest. For example, in the bin packing problem we are given n items of sizes in the interval $[0,1]$ and are required to pack them into the fewest number of unit-capacity bins as possible. In the stochastic version, the item sizes are independent random variables in the interval $[0,1]$.

In this note we prove rather general and sharp concentration inequality for smooth random fields. As a simple corollary of the main result we get a sharp exponential inequality for a convex function of the maximal eigenvalue of a random matrix.

[1] This research was partially supported by the Deutsche Forschungsgemeinschaft through the CRC 823 “Statistics of nonlinear dynamical processes”, the CRC 649 “Economic Risk” and by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

2 Main setup

Let $G(x;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ be a family of real valued functions on $\mathbb{R}^n$ and let X be a random vector in $\mathbb{R}^n$. The purpose of this paper is to derive exponential probability bounds for the random variable:

$$\sup_{\boldsymbol{\theta}\in\Theta} G(X,\boldsymbol{\theta}).$$

First we make the following assumptions.

(*GC*) *The function* $G(x,\boldsymbol{\theta})$ *is smooth in* $\boldsymbol{\theta}$ *for any* $x \in \mathbb{R}^n$ *and the mean function* $M(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}G(X,\boldsymbol{\theta})$ *is three times continuously differentiable in* $\boldsymbol{\theta}$. *Denote*

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} M(\boldsymbol{\theta}).$$

There is a positive definite symmetric matrix D^* *and a positive number* $\mathtt{r}_0 > 0$ *such that*

$$\nabla^2 M(\boldsymbol{\theta}) \preceq -D^*, \quad |\boldsymbol{\theta}-\boldsymbol{\theta}^*| = \mathtt{r}, \quad \mathtt{r} > \mathtt{r}_0. \tag{2.1}$$

(*VI*) *There is a symmetric positive definite matrix* V_0 *such that*

$$\operatorname{Var}\{\nabla_{\boldsymbol{\theta}} G(X,\boldsymbol{\theta}^*)\} \preceq V_0^2$$

and

$$V_0^2 \succeq \varepsilon^{-2} I, \quad \mathfrak{a}^2 D_0^2 \succeq V_0^2$$

with $\nabla^2 M(\boldsymbol{\theta}^*) = -D_0^2$, a small parameter $\varepsilon \in (0,1/2)$ and $\mathfrak{a} \in \mathbb{R}$.

Introduce a centred random field:

$$\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} G(X;\boldsymbol{\theta}) - \mathbb{E}G(X;\boldsymbol{\theta})$$

and a local elliptic neighbourhood of $\boldsymbol{\theta}^*$ via

$$\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta : \|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\| \le \mathtt{r}\}.$$

Finally we make two integrability assumptions.

(*ED*) *There exists a constant* ω_0 *such that it holds for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and all $\mathtt{r} \le \mathtt{r}_0$,

$$\sup_{\gamma\in\mathbb{R}^p} \log \mathbb{E}\exp\left\{\lambda \frac{\gamma^\top\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\}}{\omega_0 \varepsilon \mathtt{r} \|V_0\gamma\|}\right\} \le \nu_0^2\lambda^2/2, \qquad |\lambda| \le \mathtt{g}.$$

$(\mathcal{E}\mathtt{r})$ *It holds for any* $\lambda > 0$,

$$\sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp\left\{ \lambda \frac{\gamma^\top \nabla \zeta(\boldsymbol{\theta})}{\|V_0 \gamma\|} \right\} \le \nu_0^2 \lambda^2 / 2.$$

Discussion Under (GC) the second order Taylor expansion of the function $M(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$ gives

$$M(\boldsymbol{\theta}) = M(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla^2 M(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + R(\boldsymbol{\theta})$$

with

$$\frac{R(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2} = O(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|), \quad \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \to 0.$$

Then under (VI)

$$\left| \frac{2(M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}^*))}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} + 1 \right| \le \delta_0 \varepsilon \mathtt{r}, \quad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}) \tag{2.2}$$

for some $\delta_0 > 0$. The condition (2.1) basically means that M is globally concave and together with the Taylor expansion

$$M(\boldsymbol{\theta}) = M(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla^2 M(\boldsymbol{\theta}^* + \alpha(\boldsymbol{\theta} - \boldsymbol{\theta}^*))(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \quad \alpha \in (0,1)$$

gives

$$M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}^*) \le -\frac{\lambda_{\min}(D^*)}{\lambda_{\max}(V_0^2)} \mathtt{r}^2 \stackrel{\text{def}}{=} -\mathtt{b}^* \mathtt{r}^2 \tag{2.3}$$

if $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| = \mathtt{r}$.

3 Main result

Define for $\mathbb{B} \stackrel{\text{def}}{=} D_0^{-1} V_0^2 D_0^{-1}$

$$\mathtt{p} \stackrel{\text{def}}{=} \operatorname{tr}(\mathbb{B}), \qquad \mathtt{v}^2 \stackrel{\text{def}}{=} 2\operatorname{tr}(\mathbb{B}^2), \qquad \lambda_0 \stackrel{\text{def}}{=} \|\mathbb{B}\|_\infty = \lambda_{\max}(\mathbb{B}).$$

Theorem 3.1. *Under assumptions* (GC), (VI), (ED) *and* $(\mathcal{E}\mathtt{r})$

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta} \in \Theta} G(X, \boldsymbol{\theta}) > G(X, \boldsymbol{\theta}^*) + \lambda_0 \mathtt{p}/2 + c\lambda_0(\mathtt{v}\sqrt{x} + x) \right) \le e^{-x},$$

for any $x > 0$ *satisfying* $\varepsilon\sqrt{(x + 3p)} < 1$ *and some constant* c *depending on* ν_0, $\mathtt{b}^*$, δ_0 *and* ω_0 *only.*

Applications (maximal eigenvalue) Let $A = (a_{ij})_{i,j=1}^{p}$ be a Hermitian random matrix with a positive definite symmetric mean matrix $\mathbb{E}A$ and let

$$G(A,\boldsymbol{\theta}) \stackrel{\text{def}}{=} \boldsymbol{\theta}^\top A \boldsymbol{\theta} - f(\|\boldsymbol{\theta}\|^2), \quad \boldsymbol{\theta} \in \Theta$$

with $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : |\boldsymbol{\theta}| < R\}$ for some large enough $R > 0$ and a nonnegative monotone increasing smooth function f. Let f^* be the Legendre transform of f, then

$$\sup_{\boldsymbol{\theta}\in\Theta} G(A,\boldsymbol{\theta}) = f^*(\lambda_{\max}(A)).$$

Since $M(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}G(A,\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbb{E}A\boldsymbol{\theta} - f(\|\boldsymbol{\theta}\|^2)$,

$$\sup_{\boldsymbol{\theta}\in\Theta} M(\boldsymbol{\theta}) = f^*(\lambda_{\max}(\mathbb{E}A))$$

and the maximum is attained in the point $\boldsymbol{\theta}^* = \sqrt{r^*}\boldsymbol{e}_p$, where $\boldsymbol{e}_p$ is the eigenvector of the matrix $\mathbb{E}A$ corresponding to its largest eigenvalue and $r^* > 0$ solves the equation $f'(r^*) = \lambda_{\max}(\mathbb{E}A)$. Moreover

$$\nabla^2 M(\boldsymbol{\theta}) = \mathbb{E}A - f'(\|\boldsymbol{\theta}\|^2)\, I - f''(\|\boldsymbol{\theta}\|^2)\, \boldsymbol{\theta}\boldsymbol{\theta}^\top$$

and as a result

$$\nabla^2 M(\boldsymbol{\theta}^*) = \mathbb{E}A - \lambda_{\max}(\mathbb{E}A)\, I - f''(r^*)\, r^* \boldsymbol{e}_p \boldsymbol{e}_p^\top \stackrel{\text{def}}{=} -D_0^2$$

for some positive definite matrix D_0, provided $f''(r^*) > 0$. Hence the assumption (GC) is fulfilled if f is globally convex. Assume

$$\sup_{\|\boldsymbol{\theta}\|=r^*} \sup_{\boldsymbol{\gamma}\in\mathbb{R}^p} \log \mathbb{E} \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top (A - \mathbb{E}A)\boldsymbol{\theta}}{\|V_0\boldsymbol{\gamma}\|} \right\} \le \nu_0^2 \lambda^2/2, \tag{3.1}$$

where $V_0 = \operatorname{Var}(A\boldsymbol{\theta}^*)$. Our main result implies

$$\mathbb{P}\Bigl(f^*(\lambda_{\max}(A)) \, - \, f^*(\lambda_{\max}(\mathbb{E}A)) \ge \\ \frac{\lambda_0 \mathtt{p}}{2} + (\boldsymbol{\theta}^*)^\top (A - \mathbb{E}A)\boldsymbol{\theta}^* + c\lambda_0(\mathtt{v}\sqrt{x} + x) \Bigr) \le e^{-x} \tag{3.2}$$

with $\mathtt{p} = \operatorname{tr}\bigl(D_0^{-2}V_0^2\bigr)$ and $\mathtt{v}^2 = \operatorname{tr}\bigl(D_0^{-4}V_0^4\bigr)$. Furthermore it follows from (3.1)

$$\mathbb{P}\Bigl((\boldsymbol{\theta}^*)^\top (A - \mathbb{E}A)\boldsymbol{\theta}^* > \sqrt{x}\|V_0\boldsymbol{\theta}^*\| \Bigr) \le e^{\nu_0^2/2} e^{-x}. \tag{3.3}$$

Combining (3.2) with (3.3), we get

$$\mathbb{P}\Big(f^*(\lambda_{\max}(A)) - f^*(\lambda_{\max}(\mathbb{E}A)) \ge \frac{\lambda_0 \mathrm{p}}{2} + \sqrt{x}\|V_0\boldsymbol{\theta}^*\| + c\lambda_0(\mathrm{v}\sqrt{x}+x)\Big) \le (1+e^{\nu_0^2/2})e^{-x} \tag{3.4}$$

Let us compare the above inequality with the known results on the maximal eigenvalue of a random Hermitian matrix. For example, in Mackey at el. (2012) an exponential inequality for the spectral norm of a bounded Hermitian random matrix A is derived via the method of exchangeable pairs. In particular, it is shown that if $A = X_1 + \ldots + X_n$, where $X_1, \ldots, X_n$ are independent identically distributed Hermitian $p \times p$ matrices satisfying

$$X_k^2 \preceq B^2, \quad k = 1, \ldots, n, \tag{3.5}$$

then

$$\mathbb{P}(\lambda_{\max}(A - \mathbb{E}A) > t) \le p \cdot \exp\left(-t^2/2\sigma^2\right) \tag{3.6}$$

with $\sigma^2 = \frac{n}{2}\left\|B^2 + \mathrm{Var}(X_1)\right\|$. The inequality (3.6) is in fact equivalent to the following one

$$\mathbb{P}\left(\lambda_{\max}(A - \mathbb{E}A) > \sqrt{2(x+\log p)}\sigma\right) \le \exp(-x) \tag{3.7}$$

In our setting with $f(x) = nx^2$ we get $r^* = \lambda_{\max}(\mathbb{E}A)/(2n) = \lambda_{\max}(\mathbb{E}X_1)/2$,

$$\begin{aligned} V_0^2 &= \lambda_{\max}(\mathbb{E}A)\,\mathrm{Var}(A\boldsymbol{e}_p)/(2n) \\ &= n \cdot \lambda_{\max}(\mathbb{E}X_1)\,\mathrm{Var}(X_1\boldsymbol{e}_p)/2 \\ D_0^2 &= n \cdot \lambda_{\max}(\mathbb{E}A)(I + 2\boldsymbol{e}_p\boldsymbol{e}_p^\top - \mathbb{E}X_1/\lambda_{\max}(\mathbb{E}X_1)) \end{aligned}$$

and

$$D_0^{-2}V_0^2 = (I + 2\boldsymbol{e}_p\boldsymbol{e}_p^\top - \mathbb{E}X_1/\lambda_{\max}(\mathbb{E}X_1))^{-1}\,\mathrm{Var}(X_1\boldsymbol{e}_p)/2.$$

Hence $\mathrm{p} = c_1 \cdot p$ and $\mathrm{v} = c_2 \cdot p$ for some constants c_1 and c_2 not depending on n and p. Furthermore,

$$\|V_0\boldsymbol{\theta}^*\| = \sqrt{n} \cdot \lambda_{\max}(\mathbb{E}X_1)\left\|\boldsymbol{e}_p^\top\,\mathrm{Var}^{1/2}(X_1\boldsymbol{e}_p)\right\|.$$

and the inequality (3.4) transforms to

$$\mathbb{P}\left(\lambda_{\max}^2(A) - \lambda_{\max}^2(\mathbb{E}A) \ge c(\sqrt{nx} + x + p)\right) \le (1+e^{\nu_0^2/2})e^{-x} \tag{3.8}$$

with some constant $c > 0$ not depending on p and n.

Note that in the domain $\lambda_{\max}(A+\mathbb{E}A) > 1$, $p/n < 1$ the inequality (3.8) is more accurate than (3.7). Moreover, while the condition (3.5) basically means that A is bounded with probability 1, our assumption (3.1) only requires a sub-gaussian behaviour of $A - \mathbb{E}A$.

4 Proof of the main result

Proof. Denote $Z(\boldsymbol{\theta},\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} G(X,\boldsymbol{\theta}) - G(X,\boldsymbol{\theta}^*)$. We get from Proposition 4.2, Lemma 4.3 and Lemma 4.4

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} Z(\boldsymbol{\theta},\boldsymbol{\theta}^*) &\le \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{Z}_\epsilon(\boldsymbol{\theta},\boldsymbol{\theta}^*) + \diamondsuit_\epsilon(\mathtt{r}) \\
&\le \|\xi_\epsilon\|^2/2 + \diamondsuit_\epsilon(\mathtt{r}) \\
&= \|\xi\|^2/2 + \{\|\xi_\epsilon\|^2 - \|\xi\|^2\}/2 + \diamondsuit_\epsilon(\mathtt{r}) \\
&\le \|\xi\|^2/2 + \frac{\tau_\epsilon}{2(1-\tau_\epsilon)}\|\xi\|^2 + \diamondsuit_\epsilon(\mathtt{r}) \\
&= \frac{\|\xi\|^2}{2(1-\tau_\epsilon)} + \diamondsuit_\epsilon(\mathtt{r})
\end{aligned}$$

Now Proposition 4.5 implies

$$\mathbb{P}\Bigl(\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r}_0)} Z(\boldsymbol{\theta},\boldsymbol{\theta}^*) > \frac{\lambda_0\cdot\mathfrak{z}(\mathtt{x},\mathbb{B})}{2(1-\tau_\epsilon)} + 6\nu_0\omega_0\varepsilon\mathtt{r}_0\bigl(1+\sqrt{\mathtt{x}+3p}\bigr)^2\Bigr) \le 4e^{-\mathtt{x}},$$

where $\mathfrak{z}(\mathtt{x},\mathbb{B})$ is given by (4.5). Next, we shall prove that there is $\mathtt{r}_0 > 0$ and a deterministic upper function $\mathfrak{u}(\boldsymbol{\theta}) \ge 0$ such that

$$\mathbb{P}\Bigl(\sup_{\boldsymbol{\theta}\in\Theta\setminus\Theta_0(\mathtt{r})} \{Z(\boldsymbol{\theta},\boldsymbol{\theta}^*) + \mathfrak{u}(\boldsymbol{\theta})\} \ge 0\Bigr) \le \mathrm{e}^{-\mathtt{x}} \tag{4.1}$$

for $\mathtt{r} > \mathtt{r}_0$ and $\mathtt{x} > 0$. The inequality (4.1) then implies

$$\mathbb{P}\Bigl(\sup_{\boldsymbol{\theta}\notin\Theta_0(\mathtt{r}_0)} Z(\boldsymbol{\theta},\boldsymbol{\theta}^*) \ge 0\Bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

A possible way of checking the condition (4.1) is based on a lower quadratic bound for the negative expectation $M(\boldsymbol{\theta})$ in the sense of condition (2.3).

Lemma 4.1. *Suppose* (GC) *and* $(E\mathtt{r})$. *Let, for* $\mathtt{r} \ge \mathtt{r}_0$,

$$6\nu_0\sqrt{\mathtt{x}+3p} \le \mathtt{r}\mathtt{b}^*,$$

with $\mathtt{x}+3p \ge 2.5$. *Then*

$$\mathbb{P}\Bigl(\sup_{\boldsymbol{\theta}\notin\Theta_0(\mathtt{r})} Z(\boldsymbol{\theta},\boldsymbol{\theta}^*) \ge 0\Bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

Proof. The result follows from Theorem 5.8 with $\mu = \frac{\mathtt{b}^*}{3\nu_0}$, $\mathfrak{t}(\mu) \equiv 0$, $\mathscr{U}(\boldsymbol{\theta}) = Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{E} Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and $M(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}^*) \geq \frac{\mathtt{b}^*}{2} \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2$. □

It follows now from Lemma 4.1 that the inequality

$$\sup_{\boldsymbol{\theta} \in \Theta} Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \leq \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)} Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$$

holds with probability at least $1 - \mathrm{e}^{-\mathtt{x}}$. As a result we get the desired inequality. □

4.1 Auxiliary results

Let δ, ϱ be nonnegative constants. Introduce for a vector $\boldsymbol{\epsilon} = (\delta, \varrho)$ the following notation:

$$\begin{aligned} \mathbb{Z}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2 \\ &= \xi_{\boldsymbol{\epsilon}}^\top D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2, \end{aligned} \tag{4.2}$$

where $\nabla \zeta(\boldsymbol{\theta}^*) = \nabla_{\boldsymbol{\theta}} G(X, \boldsymbol{\theta}^*)$ by $\nabla M(\boldsymbol{\theta}^*) = 0$ and

$$D_{\boldsymbol{\epsilon}}^2 = D_0^2(1 - \delta) - \varrho V_0^2, \qquad \xi_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} D_{\boldsymbol{\epsilon}}^{-1} \nabla G(X, \boldsymbol{\theta}^*).$$

Here we implicitly assume that with the proposed choice of the constants δ and ϱ, the matrix $D_{\boldsymbol{\epsilon}}^2$ is non-negative: $D_{\boldsymbol{\epsilon}}^2 \geq 0$. The representation (4.2) indicates that the process $\mathbb{Z}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ has the quadratic local structure. Now, given $\mathtt{r}$, fix some $\delta \geq \delta_0 \varepsilon \mathtt{r}$ and $\varrho \geq 3\nu_0 \omega_0 \varepsilon \mathtt{r}$ with the value δ_0 from (2.2) and ω_0 from condition (ED). Finally set $\underline{\boldsymbol{\epsilon}} = -\boldsymbol{\epsilon}$, so that $D_{\underline{\boldsymbol{\epsilon}}}^2 = D_0^2(1 + \delta) + \varrho V_0^2$.

Proposition 4.2. *Assume* (ED) *and* (VI). *Let for some* $\mathtt{r}$, *the values* $\varrho \geq 3\nu_0 \omega_0 \varepsilon \mathtt{r}$ *and* $\delta \geq \delta_0 \varepsilon \mathtt{r}$ *be such that* $D_0^2(1 - \delta) - \varrho V_0^2 \geq 0$. *Then*

$$\mathbb{Z}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \leq Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \leq \mathbb{Z}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \quad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}),$$

with $\mathbb{Z}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*), \mathbb{Z}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ *defined by* (4.2). *The error terms* $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ *and* $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ *satisfy*

$$\mathbb{P}\bigl(\varrho^{-1} \max\{\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})\} \geq \bigl(1 + \sqrt{\mathtt{x} + 3p}\bigr)^2\bigr) \leq \exp(-\mathtt{x}). \tag{4.3}$$

Proof. Consider for fixed $\mathtt{r}$ and $\boldsymbol{\epsilon} = (\delta, \varrho)$ the quantity

$$\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \Bigl\{ Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{E} Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \frac{\varrho}{2} \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \Bigr\}.$$

As $\delta \ge \delta_0 \varepsilon \mathtt{r}$, it holds $-M(\boldsymbol{\theta}) \ge (1-\delta) D_0^2$ and $Z(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{Z}_{\epsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \diamondsuit_{\epsilon}(\mathtt{r})$. Moreover, in view of $\nabla M(\boldsymbol{\theta}^*) = 0$, the definition of $\diamondsuit_{\epsilon}(\mathtt{r})$ can be rewritten as

$$\diamondsuit_{\epsilon}(\mathtt{r}) = \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl\{ \zeta(\boldsymbol{\theta}) - \zeta(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top} \nabla \zeta(\boldsymbol{\theta}^*) - \frac{\varrho}{2} \| V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \|^2 \bigr\}.$$

Now the claim of the theorem can be easily reduced to an exponential bound for the quantity $\diamondsuit_{\epsilon}(\mathtt{r})$. We apply Theorem 5.6 to the process

$$\mathscr{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \frac{1}{\omega_0 \varepsilon \mathtt{r}} \bigl\{ \zeta(\boldsymbol{\theta}) - \zeta(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top} \nabla \zeta(\boldsymbol{\theta}^*) \bigr\}, \qquad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}),$$

and $H_0 = V_0$. Condition $(\mathscr{E}D)$ follows from (ED) with the same ν_0 and $\mathtt{g}$ in view of $\nabla \mathscr{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \bigl\{ \nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}^*) \bigr\} / \omega \varepsilon \mathtt{r}$. So, the conditions of Theorem 5.6 are fulfilled yielding (4.3) in view of $\varrho \ge 3 \nu_0 \, \omega_0 \varepsilon \mathtt{r}$. ☐

Lemma 4.3. *It holds*

$$\sup_{\boldsymbol{\theta} \in \Theta} \mathbb{Z}_{\epsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \sup_{\boldsymbol{\theta} \in I\!\!R^p} \mathbb{Z}_{\epsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \| \xi_{\epsilon} \|^2 / 2$$

Lemma 4.4. *Define* $\xi \stackrel{\text{def}}{=} D_0^{-1} \nabla \zeta(\boldsymbol{\theta}^*)$. *Suppose* (VI) *and let* $\tau_{\epsilon} \stackrel{\text{def}}{=} \varepsilon r_0 (\delta_0 + 3 \nu_0 \omega_0 \mathfrak{a}^2) < 1$. *Then*

$$\| \xi_{\epsilon} \|^2 - \| \xi \|^2 \le \frac{\tau_{\epsilon}}{1 - \tau_{\epsilon}} \| \xi \|^2, \quad \| \xi \|^2 - \| \xi_{\underline{\epsilon}} \|^2 \le \frac{\tau_{\epsilon}}{1 + \tau_{\epsilon}} \| \xi \|^2.$$

Proposition 4.5. *Let* (ED) *hold with* $\nu_0 = 1$. *Then* $I\!\!E \| \xi \|^2 \le \mathtt{p}$, *and for each* $\mathtt{x} > 0$

$$I\!\!P \bigl(\| \xi \|^2 \ge \lambda_0 \cdot \mathfrak{z}(\mathtt{x}, I\!\!B) \bigr) \le 2 \mathrm{e}^{-\mathtt{x}}, \tag{4.4}$$

where $\mathfrak{z}(\mathtt{x}, I\!\!B)$ *is defined by*

$$\mathfrak{z}(\mathtt{x}, I\!\!B) \stackrel{\text{def}}{=} \begin{cases} \mathtt{p} + 2 \mathtt{v} \mathtt{x}^{1/2}, & \mathtt{x} \le \mathtt{v}/18, \\ \mathtt{p} + 6 \mathtt{x} & \mathtt{v}/18 < \mathtt{x}. \end{cases} \tag{4.5}$$

Proof. It holds

$$\begin{aligned} I\!\!E \| \xi \|^2 &= I\!\!E \operatorname{tr} \xi \xi^{\top} \\ &= \operatorname{tr} D_0^{-1} \bigl[I\!\!E \nabla \zeta(\boldsymbol{\theta}^*) \{ \nabla \zeta(\boldsymbol{\theta}^*) \}^{\top} \bigr] D_0^{-1} = \operatorname{tr} \bigl[D_0^{-2} \operatorname{Var} \{ \nabla \zeta(\boldsymbol{\theta}^*) \} \bigr] \end{aligned}$$

and $\boldsymbol{\gamma}^{\top} \operatorname{Var} \{ \nabla \zeta(\boldsymbol{\theta}^*) \} \boldsymbol{\gamma} \le \boldsymbol{\gamma}^{\top} V_0^2 \boldsymbol{\gamma}$ and thus, $I\!\!E \| \xi \|^2 \le \mathtt{p}$. The deviation bound (4.4) is proved in Corollary 5.2. ☐

5 Appendix

The proofs of the results below can be found in Appendix A and Appendix B of Spokoiny (2011).

5.1 Deviation probability for quadratic forms

Assume that

$$\log \mathbb{E} \exp\bigl(\gamma^\top \xi\bigr) \le \|\gamma\|^2/2, \qquad \gamma \in \mathbb{R}^p,\ \|\gamma\| \le \mathrm{g}. \tag{5.1}$$

This section presents a general exponential bound for the probability $\mathbb{P}\bigl(\|\mathbb{B}\xi\| > \mathrm{y}\bigr)$ with a given matrix $\mathbb{B}$ and a vector ξ obeying the condition (5.1). We assume that $\mathbb{B}$ is symmetric. Define important characteristics of $\mathbb{B}$

$$\mathrm{p} = \operatorname{tr}(\mathbb{B}^2), \qquad \mathrm{v}^2 = 2\operatorname{tr}(\mathbb{B}^4), \qquad \lambda^* \overset{\text{def}}{=} \|\mathbb{B}^2\|_\infty \overset{\text{def}}{=} \lambda_{\max}(\mathbb{B}^2).$$

For simplicity of formulation we suppose that $\lambda^* = 1$, otherwise one has to replace p and v^2 with p/λ^* and v^2/λ^*. Let g be given in (5.1). Define w_c by the equation

$$\frac{w_c(1+w_c)}{(1+w_c^2)^{1/2}} = \mathrm{g}\mathrm{p}^{-1/2}.$$

Define also $\mu_c = w_c^2/(1+w_c^2) \wedge 2/3$. Note that $w_c^2 \ge 2$ implies $\mu_c = 2/3$. Further define

$$\mathrm{y}_c^2 = (1+w_c^2)\mathrm{p}, \qquad 2\mathrm{x}_c = \mu_c \mathrm{y}_c^2 + \log\det\{I_p - \mu_c \mathbb{B}^2\}. \tag{5.2}$$

Theorem 5.1. *Let a random vector* ξ *in* $\mathbb{R}^p$ *fulfill* (5.1). *Then for each* $\mathrm{x} < \mathrm{x}_c$

$$\mathbb{P}\bigl(\|\mathbb{B}\xi\|^2 - \mathbb{E}\|\mathbb{B}\xi\|^2 > (2\mathrm{v}\mathrm{x}^{1/2}) \vee (6\mathrm{x}), \|\mathbb{B}\xi\| \le \mathrm{y}_c\bigr) \le 2\exp(-\mathrm{x}).$$

Moreover, for $\mathrm{y} \ge \mathrm{y}_c$, *with* $\mathrm{g}_c = \mathrm{g} - \sqrt{\mu_c \mathrm{p}} = \mathrm{g} w_c/(1+w_c)$, *it holds*

$$\mathbb{P}\bigl(\|\mathbb{B}\xi\| > \mathrm{y}\bigr) \le 8.4\exp\bigl(-\mathrm{x}_c - \mathrm{g}_c(\mathrm{y}-\mathrm{y}_c)/2\bigr).$$

Let us now describe the value $\mathfrak{z}(\mathrm{x},\mathbb{B})$ ensuring a small value for the large deviation probability $\mathbb{P}\bigl(\|\mathbb{B}\xi\|^2 > \mathfrak{z}(\mathrm{x},\mathbb{B})\bigr)$. For ease of formulation, we suppose that $\mathrm{g}^2 \ge 2\mathrm{p}$ yielding $\mu_c^{-1} \le 3/2$. The other case can be easily adjusted.

Corollary 5.2. *Let* ξ *fulfill* (5.1) *with* $\mathrm{g}^2 \ge 2\mathrm{p}$. *Then it holds for* $\mathrm{x} \le \mathrm{x}_c$ *with* x_c *from* (5.2)*:*

$$\mathbb{P}\bigl(\|\mathbb{B}\xi\|^2 - \mathbb{E}\|\mathbb{B}\xi\|^2 \ge \mathfrak{z}(\mathrm{x},\mathbb{B})\bigr) \le 2e^{-\mathrm{x}} + 8.4e^{-\mathrm{x}_c},$$

$$\mathfrak{z}(\mathrm{x},\mathbb{B}) \overset{\text{def}}{=} \begin{cases} 2\mathrm{v}\mathrm{x}^{1/2}, & \mathrm{x} \le \mathrm{v}/18, \\ 6\mathrm{x} & \mathrm{v}/18 < \mathrm{x} \le \mathrm{x}_c. \end{cases}$$

For $\mathtt{x} > \mathtt{x}_c$

$$\mathbb{P}\bigl(\|\mathbb{B}\xi\|^2 \ge \mathfrak{z}_c(\mathtt{x},\mathbb{B})\bigr) \le 8.4\mathrm{e}^{-\mathtt{x}}, \qquad \mathfrak{z}_c(\mathtt{x},\mathbb{B}) \stackrel{\mathrm{def}}{=} \bigl|\mathtt{y}_c + 2(\mathtt{x}-\mathtt{x}_c)/\mathtt{g}_c\bigr|^2.$$

5.2 Some results for empirical processes

This chapter presents some general results of the theory of empirical processes. We assume some exponential moment conditions on the increments of the process which allows to apply the well developed chaining arguments in Orlicz spaces; We, however, follow the more recent approach inspired by the notions of generic chaining and majorizing measures due to M. Talagrand; see e.g. Talagrand (2005). The results are close to that of Bednorz (2006). We state the results in a slightly different form and present an independent and self-contained proof. The first result states a bound for local fluctuations of the process $\mathcal{U}(\boldsymbol{v})$ given on a metric space Υ. Then this result will be used for bounding the maximum of the negatively drifted process $\mathcal{U}(\boldsymbol{v}) - \mathcal{U}(\boldsymbol{v}_0) - \rho d^2(\boldsymbol{v},\boldsymbol{v}_0)$ over a vicinity $\Upsilon_\circ(\mathtt{r})$ of the central point $\boldsymbol{v}_0$. The behavior of $\mathcal{U}(\boldsymbol{v})$ outside of the local central set $\Upsilon_\circ(\mathtt{r})$ is described using the *upper function* method. Namely, we construct a multiscale deterministic function $\mathfrak{u}(\mu,\boldsymbol{v})$ ensuring that with probability at least $1-\mathrm{e}^{-\mathtt{x}}$ it holds $\mu\mathcal{U}(\boldsymbol{v}) + \mathfrak{u}(\mu,\boldsymbol{v}) \le \mathfrak{z}(\mathtt{x})$ for all $\boldsymbol{v} \notin \Upsilon_\circ(\mathtt{r})$ and $\mu \in \mathbb{M}$, where $\mathfrak{z}(\mathtt{x})$ grows linearly in $\mathtt{x}$.

Let $d(\boldsymbol{v},\boldsymbol{v}')$ be a semi-distance on Υ. We suppose the following condition to hold:

$(\mathcal{E}d)$ *There exist* $\mathtt{g} > 0$, $\mathtt{r}_0 > 0$, $\nu_0 \ge 1$, *such that for any* $\lambda \le \mathtt{g}$ *and* $\boldsymbol{v},\boldsymbol{v}' \in \Upsilon$ *with* $d(\boldsymbol{v},\boldsymbol{v}') \le \mathtt{r}_0$

$$\log \mathbb{E}\exp\Bigl\{\lambda \frac{\mathcal{U}(\boldsymbol{v}) - \mathcal{U}(\boldsymbol{v}')}{d(\boldsymbol{v},\boldsymbol{v}')}\Bigr\} \le \nu_0^2\lambda^2/2. \tag{5.3}$$

Formulation of the result involves a sigma-finite measure π on the space Υ which is often called the *majorizing measure* and used in the *generic chaining* device; A typical example of choosing π is the Lebesgue measure on $\mathbb{R}^p$. Let Υ° be a subset of Υ, a sequence $\mathtt{r}_k$ be fixed with $\mathtt{r}_0 = \mathrm{diam}(\Upsilon^\circ)$ and $\mathtt{r}_k = \mathtt{r}_0 2^{-k}$. Let also $\mathcal{B}_k(\boldsymbol{v}) \stackrel{\mathrm{def}}{=} \{\boldsymbol{v}' \in \Upsilon^\circ : d(\boldsymbol{v},\boldsymbol{v}') \le \mathtt{r}_k\}$ be the d-ball centered at $\boldsymbol{v}$ of radius $\mathtt{r}_k$ and $\pi_k(\boldsymbol{v})$ denote its π-measure:

$$\pi_k(\boldsymbol{v}) \stackrel{\mathrm{def}}{=} \int_{\mathcal{B}_k(\boldsymbol{v})} \pi(d\boldsymbol{v}') = \int_{\Upsilon^\circ} \mathbb{I}\bigl(d(\boldsymbol{v},\boldsymbol{v}') \le \mathtt{r}_k\bigr)\pi(d\boldsymbol{v}').$$

Denote also

$$M_k \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{v}\in\Upsilon^\circ} \frac{\pi(\Upsilon^\circ)}{\pi_k(\boldsymbol{v})} \qquad k \ge 1.$$

Finally set $c_1 = 1/3$, $c_k = 2^{-k+2}/3$ for $k \ge 2$, and define the value $\mathbb{Q}(\Upsilon^\circ)$ by

$$\mathbb{Q}(\Upsilon^\circ) \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} c_k \log(2M_k) = \frac{1}{3}\log(2M_1) + \frac{4}{3}\sum_{k=2}^{\infty} 2^{-k}\log(2M_k).$$

Theorem 5.3. *Let $\mathscr{U}$ be a separable process following to $(\mathscr{E}d)$. If Υ° is a d-ball in Υ with the center $\boldsymbol{v}^\circ$ and the radius $\mathtt{r}_0$, i.e. $d(\boldsymbol{v},\boldsymbol{v}^\circ) \le \mathtt{r}_0$ for all $\boldsymbol{v} \in \Upsilon^\circ$, then for $\lambda \le \mathrm{g}_0 \stackrel{\text{def}}{=} \nu_0 \mathrm{g}$*

$$\log \mathbb{E}\exp\Bigl\{\frac{\lambda}{3\nu_0\mathtt{r}_0}\sup_{\boldsymbol{v}\in\Upsilon^\circ}\bigl|\mathscr{U}(\boldsymbol{v}) - \mathscr{U}(\boldsymbol{v}^\circ)\bigr|\Bigr\} \le \lambda^2/2 + \mathbb{Q}(\Upsilon^\circ).$$

Due to the result of Theorem 5.3, the bound for the maximum of $\mathscr{U}(\boldsymbol{v},\boldsymbol{v}_0)$ over $\boldsymbol{v} \in \mathscr{B}_{\mathtt{r}}(\boldsymbol{v}_0)$ grows quadratically in $\mathtt{r}$. So, its applications to situations with $\mathtt{r}^2 \gg \mathbb{Q}(\Upsilon^\circ)$ are limited. The next result shows that introducing a negative quadratic drift helps to state a uniform in $\mathtt{r}$ local probability bound. Namely, the bound for the process $\mathscr{U}(\boldsymbol{v},\boldsymbol{v}_0) - \rho d^2(\boldsymbol{v},\boldsymbol{v}_0)/2$ with some positive ρ over a ball $\mathscr{B}_{\mathtt{r}}(\boldsymbol{v}_0)$ around the point $\boldsymbol{v}_0$ only depends on the drift coefficient ρ but not on $\mathtt{r}$.

Theorem 5.4. *Let $\mathtt{r}^*$ be such that $(\mathscr{E}d)$ holds on $\mathscr{B}_{\mathtt{r}^*}(\boldsymbol{v}_0)$. Let also $\mathbb{Q}(\Upsilon^\circ) \le \mathbb{Q}$ for $\Upsilon^\circ = \mathscr{B}_{\mathtt{r}}(\boldsymbol{v}_0)$ with $\mathtt{r} \le \mathtt{r}^*$. If $\rho > 0$ and $\mathfrak{z}$ are fixed to ensure $\sqrt{2\rho\mathfrak{z}} \le \mathrm{g}_0 = \nu_0\mathrm{g}$ and $\rho(\mathfrak{z}-1) \ge 2$, then it holds*

$$\log \mathbb{P}\Bigl(\sup_{\boldsymbol{v}\in\mathscr{B}_{\mathtt{r}^*}(\boldsymbol{v}_0)}\Bigl\{\frac{1}{3\nu_0}\mathscr{U}(\boldsymbol{v},\boldsymbol{v}_0) - \frac{\rho}{2}d^2(\boldsymbol{v},\boldsymbol{v}_0)\Bigr\} > \mathfrak{z}\Bigr) \\ \le -\rho(\mathfrak{z}-1) + \log(4\mathfrak{z}) + \mathbb{Q}.$$

Moreover, if $\sqrt{2\rho\mathfrak{z}} > \mathrm{g}_0$, then

$$\log \mathbb{P}\Bigl(\sup_{\boldsymbol{v}\in\mathscr{B}_{\mathtt{r}^*}(\boldsymbol{v}_0)}\Bigl\{\frac{1}{3\nu_0}\mathscr{U}(\boldsymbol{v},\boldsymbol{v}_0) - \frac{\rho}{2}d^2(\boldsymbol{v},\boldsymbol{v}_0)\Bigr\} > \mathfrak{z}\Bigr) \\ \le -\mathrm{g}_0\sqrt{\rho(\mathfrak{z}-1)} + \mathrm{g}_0^2/2 + \log(4\mathfrak{z}) + \mathbb{Q}.$$

This result can be used for describing the concentration bound for the maximum of $(3\nu_0)^{-1}\mathscr{U}(\boldsymbol{v},\boldsymbol{v}_0) - \rho d^2(\boldsymbol{v},\boldsymbol{v}_0)/2$. Namely, it suffices to find $\mathfrak{z}$ ensuring the prescribed deviation probability. We state the result for a special case with $\rho = 1$ and $\mathrm{g}_0 \ge 3$ which simplifies the notation.

Corollary 5.5. *Under the conditions of Theorem 5.4, for any $\mathrm{x} \ge 0$ with $\mathrm{x} + \mathbb{Q} \ge 4$*

$$\mathbb{P}\Bigl(\sup_{\boldsymbol{v}\in\mathscr{B}_{\mathtt{r}^*}(\boldsymbol{v}_0)}\Bigl\{\frac{1}{3\nu_0}\mathscr{U}(\boldsymbol{v},\boldsymbol{v}_0) - \frac{1}{2}d^2(\boldsymbol{v},\boldsymbol{v}_0)\Bigr\} > \mathfrak{z}_0(\mathrm{x},\mathbb{Q})\Bigr) \le \exp(-\mathrm{x}),$$

where with $\mathtt{g}_0 = \nu_0 \mathtt{g} \geq 2$

$$\mathfrak{z}_0(\mathtt{x}, \mathbb{Q}) \stackrel{\text{def}}{=} \begin{cases} \bigl(1 + \sqrt{\mathtt{x} + \mathbb{Q}}\bigr)^2 & \text{if } 1 + \sqrt{\mathtt{x} + \mathbb{Q}} \leq \mathtt{g}_0, \\ 1 + \bigl\{2\mathtt{g}_0^{-1}(\mathtt{x} + \mathbb{Q}) + \mathtt{g}_0\bigr\}^2 & \text{otherwise}. \end{cases} \tag{5.4}$$

Let us now discuss the special case when Υ is an open subset in $\mathbb{R}^p$, the stochastic process $\mathcal{U}(\boldsymbol{\upsilon})$ is absolutely continuous and its gradient $\nabla \mathcal{U}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} d\mathcal{U}(\boldsymbol{\upsilon})/d\boldsymbol{\upsilon}$ has bounded exponential moments.

$(\mathcal{E}D)$ *There exist* $\mathtt{g} > 0$, $\nu_0 \geq 1$, *and for each* $\boldsymbol{\upsilon} \in \Upsilon$, *a symmetric non-negative matrix* $H(\boldsymbol{\upsilon})$ *such that for any* $\lambda \leq \mathtt{g}$ *and any unit vector* $\boldsymbol{\gamma} \in \mathbb{R}^p$, *it holds*

$$\log \mathbb{E} \exp\Bigl\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla \mathcal{U}(\boldsymbol{\upsilon})}{\|H(\boldsymbol{\upsilon})\gamma\|}\Bigr\} \leq \nu_0^2 \lambda^2/2.$$

Consider the local sets of the elliptic form $\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} : \|H_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0)\| \leq \mathtt{r}\}$, where H_0 dominates $H(\boldsymbol{\upsilon})$ on this set: $H(\boldsymbol{\upsilon}) \preceq H_0$.

Theorem 5.6. *Let* $(\mathcal{E}D)$ *hold with some* $\mathtt{g}$ *and a matrix* $H(\boldsymbol{\upsilon})$. *Suppose that* $H(\boldsymbol{\upsilon}) \preceq H_0$ *for all* $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$. *Then*

$$\mathbb{P}\biggl(\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \Bigl\{\frac{1}{3\nu_0} \mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) - \frac{1}{2}\|H_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0)\|^2\Bigr\} \geq \mathfrak{z}_0(\mathtt{x}, p)\biggr) \leq \exp(-\mathtt{x}),$$

where $\mathfrak{z}_0(\mathtt{x}, p)$ *coincides with* $\mathfrak{z}_0(\mathtt{x}, \mathbb{Q})$ *from* (5.4) *for* $\mathbb{Q} = \mathfrak{c}_1 p$.

The previous result can be explained as a local upper function for the process $\mathcal{U}(\cdot)$. Indeed, in a vicinity $\mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)$ of the central point $\boldsymbol{\upsilon}_0$, it holds $(3\nu_0)^{-1}\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) \leq d^2(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0)/2 + \mathfrak{z}$ with a probability exponentially small in $\mathfrak{z}$. Now we extend this local result to the whole set Υ using multiscaling arguments. For simplifying the notations assume that $\mathcal{U}(\boldsymbol{\upsilon}_0) \equiv 0$. Then $\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) = \mathcal{U}(\boldsymbol{\upsilon})$. We say that $\mathfrak{u}(\mu, \boldsymbol{\upsilon})$ is a *multiscale upper function* for $\mu\mathcal{U}(\cdot)$ on a subset Υ° of Υ if

$$\mathbb{P}\Bigl(\sup_{\mu \in \mathbb{M}} \sup_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \{\mu\mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{u}(\mu, \boldsymbol{\upsilon})\} \geq \mathfrak{z}(\mathtt{x})\Bigr) \leq e^{-\mathtt{x}},$$

for some fixed function $\mathfrak{z}(\mathtt{x})$. An upper function can be used for describing the concentration sets of the point of maximum $\tilde{\boldsymbol{\upsilon}} = \operatorname{argmax}_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \mathcal{U}(\boldsymbol{\upsilon})$; see Theorem 5.8 below.

The desired global bound requires an extension of the local exponential moment condition $(\mathcal{E}d)$. Below we suppose that the pseudo-metric $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}')$ is given on the whole set Υ. For each $\mathtt{r}$ this metric defines the ball $\Upsilon_\circ(\mathtt{r})$ by the constraint $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) \leq \mathtt{r}$. Below the condition $(\mathcal{E}d)$ is assumed to be fulfilled for any $\mathtt{r}$, however the constant $\mathtt{g}$ may be dependent of the radius $\mathtt{r}$.

($\mathscr{E}\mathtt{r}$) *For any* $\mathtt{r}$*, there exists* $\mathrm{g}(\mathtt{r}) > 0$ *such that* (5.3) *holds for all* $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_{\circ}(\mathtt{r})$ *and all* $\lambda \le \mathrm{g}(\mathtt{r})$.

Condition ($\mathscr{E}\mathtt{r}$) implies a similar condition for the scaled process $\mu \mathcal{U}(\boldsymbol{\upsilon})$ with $\mathrm{g} = \mu^{-1}\mathrm{g}(\mathtt{r})$ and $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}')$ replaced by $\mu d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}')$. Corollary 5.5 implies for any $\mathtt{x}$ with $1 + \sqrt{\mathtt{x} + \mathbb{Q}} \le \mathrm{g}_0(\mathtt{r}) \stackrel{\text{def}}{=} \nu_0 \mathrm{g}(\mathtt{r})/\mu$

$$\mathbb{P}\biggl(\sup_{\boldsymbol{\upsilon} \in \mathscr{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)} \Bigl\{ \frac{\mu}{3\nu_0} \mathcal{U}(\boldsymbol{\upsilon}) - \frac{1}{2}\mu^2 \mathtt{r}^2 \Bigr\} > \mathfrak{z}_0(\mathtt{x}, \mathbb{Q}) \biggr) \le \exp(-\mathtt{x}). \tag{5.5}$$

Let now a finite or separable set $\mathbb{M}$ and a function $\mathfrak{t}(\mu) \ge 1$ be fixed such that

$$\sum_{\mu \in \mathbb{M}} \mathrm{e}^{-\mathfrak{t}(\mu)} \le 2. \tag{5.6}$$

One possible choice of the set $\mathbb{M}$ and the function $\mathfrak{t}(\mu)$ is to take a geometric sequence $\mu_k = \mu_0 2^{-k}$ with any fixed μ_0 and define $\mathfrak{t}(\mu_k) = k = -\log_2(\mu_k/\mu_0)$ for $k \ge 0$.

Putting together the bounds (5.5) for different $\mu \in \mathbb{M}$ yields the following result.

Theorem 5.7. *Suppose* ($\mathscr{E}\mathtt{r}$) *and* (5.6). *Then for any* $\mathtt{x} \ge 2$*, there exists a random set* $A(\mathtt{x})$ *of a total probability at least* $1 - 2\mathrm{e}^{-\mathtt{x}}$*, such that it holds on* $A(\mathtt{x})$ *for any* $\mathtt{r}$

$$\sup_{\boldsymbol{\upsilon} \in \mathscr{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)} \sup_{\mu \in \mathbb{M}(\mathtt{r},\mathtt{x})} \Bigl[\frac{\mu}{3\nu_0} \mathcal{U}(\boldsymbol{\upsilon}) - \frac{1}{2}\mu^2 \mathtt{r}^2 - \bigl\{ 1 + \sqrt{\mathtt{x} + \mathbb{Q} + \mathfrak{t}(\mu)} \bigr\}^2 \Bigr] < 0,$$

where

$$\mathbb{M}(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} \bigl\{ \mu \in \mathbb{M} :\, 1 + \sqrt{\mathtt{x} + \mathbb{Q} + \mathfrak{t}(\mu)} \le \nu_0 \mathrm{g}(\mathtt{r})/\mu \bigr\}.$$

Let $M(\boldsymbol{\upsilon})$ be a deterministic *boundary* function. We aim at bounding the probability that a process $\mathcal{U}(\boldsymbol{\upsilon})$ hits this boundary on the set Υ. This precisely means the probability that $\sup_{\boldsymbol{\upsilon} \in \Upsilon} \bigl\{ \mathcal{U}(\boldsymbol{\upsilon}) - M(\boldsymbol{\upsilon}) \bigr\} \ge 0$. An important observation here is that multiplication by any positive factor μ does not change the relation. This allows to apply the multiscale result from Theorem 5.7. For any fixed $\mathtt{x}$ and any $\boldsymbol{\upsilon} \in \mathscr{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)$, define

$$\mathfrak{M}^*(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \sup_{\mu \in \mathbb{M}(\mathtt{r},\mathtt{x})} \Bigl\{ \frac{1}{3\nu_0} \mu M(\boldsymbol{\upsilon}) - \frac{1}{2}\mu^2 \mathtt{r}^2 - 2\mathfrak{t}(\mu) \Bigr\}.$$

Theorem 5.8. *Suppose* ($\mathscr{E}\mathtt{r}$)*,* (5.6)*, and* $\mathtt{x} + \mathbb{Q} \ge 2.5$. *Let, given* $\mathtt{x}$*, it hold*

$$\mathfrak{M}^*(\boldsymbol{\upsilon}) \ge 2(\mathtt{x} + \mathbb{Q}), \qquad \boldsymbol{\upsilon} \in \Upsilon.$$

Then

$$\mathbb{P}\Big(\sup_{\boldsymbol{v}\in\Upsilon}\{\mathcal{U}(\boldsymbol{v})-M(\boldsymbol{v})\}\geq 0\Big)\leq 2e^{-\mathtt{x}}.$$

References

W. Bednorz. A theorem on majorizing measures. Ann. Probab., 34(5):1771-1781, 2006.

M. Ledoux. The Concentration of Measure Phenomenon. American Mathematical Society, 2001.

G. Lugosi. Concentration of Measure Inequalities. lecture Notes, 2005.

L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp, Matrix Concentration Inequalities via the Method of Exchangeable Pairs, ArXiv e-prints (January 2012), available at 1201.6002.

P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245-303, 2000.

V. Spokoiny. Parametric estimation. Finite sample theory http://arxiv.org/abs/1111.3029v4, 2011.

M. Talagrand. The generic chaining. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005.