Universal, non-asymptotic confidence sets for extrinsic and intrinsic means

Thomas Hotz

Institut für Mathematik, TU Ilmenau

joint work with
Matthias Glock, Stefan Heyder & Florian Kelma

Geometric Statistics, Toulouse, 3 September 2019

Abstract

We sketch how to construct confidence sets for extrinsic and intrinsic means using mass concentration inequalities which are non-asymptotic, and universal in the sense that they guarantee coverage of the entire set of means without making distributional assumptions except that the observations are independent and identically distributed.

Contents

1 Extrinsic means

Starting point. Assume we are given a **random variable** X taking values in a **topological space** $\mathcal{S}$. For example, think of data on a circle, a sphere, or some projective space occurring when considering landmark-based shapes. We would like to define a midpoint of the distribution of X, but since $\mathcal{S}$ does not (necessarily) carry the structure of an affine or even linear space the **mean** $\mathbf{E}\,X$ is not defined. So what to do?

Embedding. Often, there is a (natural) **embedding** $\mathcal{S} \hookrightarrow \mathbf{R}^d$ of $\mathcal{S}$ into some **Euclidean space** $\mathbf{R}^d$ (endowed with the standard inner product) of dimension $d \in \mathbf{N}$ so $\mathcal{S}$ may be identified with a subset $\mathcal{S} \subseteq \mathbf{R}^d$, e.g. with a **unit circle** or **unit sphere** (around the origin $0 \in \mathbf{R}^d$). In case of $\mathbf{C}P(k-1)$ which occurs as the space of $k+1$ landmarks in the plane modulo similarity transformations, the so-called **Veronese-Whitney embedding** is given by representing a one-dimensional subspace spanned by a unit vector $y \in \mathbf{C}^k$ by the orthogonal projection $yy^* \in \mathbf{C}^{k\times k} \cong \mathbf{R}^{2k^2}$ onto that subspace.

Extrinsic means. After identfying $\mathcal{S}$ with a subspace of $\mathbf{R}^d$, X takes values in $\mathbf{R}^d$, too, hence $\mathbf{E}\,X$ is well-defined. We will call $a = \mathbf{E}\,X$ the **ambient** (or **Euclidean**) mean of X, since typically, $\mathbf{E}\,X \notin \mathcal{S}$ but in the ambient space $\mathbf{R}^d$, e.g. this will happen for the unit sphere unless the distribution is a Dirac measure. The obvious resolution of this problem is to project the ambient mean a back onto $\mathcal{S}$. To be precise, let

$$\pi : \mathbf{R}^d \ni x \mapsto \pi(x) = \operatorname*{argmin}_{y\in\mathcal{S}} \|y - x\| \subseteq \mathcal{S} \tag{1}$$

be this (orthogonal) **projection** mapping a point x to its set of closest points in $\mathcal{S}$. Then, the **extrinsic mean set** M is defined to be

$$M = \pi(a) = \operatorname*{argmin}_{\mu\in\mathcal{S}} \|\mu - \mathbf{E}\,X\| \subseteq \mathcal{S}\,. \tag{2}$$

Existence and uniqueness. If $\mathcal{S}$ is a **closed** subset of $\mathbf{R}^d$, $\pi(x) \neq \emptyset$ for any $x \in \mathbf{R}^d$ since π is continuous, coercive, and bounded from below. So, under this mild condition, there will **exist** an extrinsic mean. **Uniqueness** of the projection, i.e. whether $\pi(x)$ contains only a single point, will often depend on x. E.g., see (Mardia und Jupp 2000), when $\mathcal{S}$ is the unit sphere, then $\pi(x)$ is unique iff $x \neq 0$ in which case $\pi(x) = \frac{x}{\|x\|}$ whereas $\pi(0) = \mathcal{S}$, in particular uniqueness fails if X is uniformly distributed on the sphere; if $a \neq 0$, $\pi(a)$ is called **spherical mean** or **mean direction** (or **circular mean** for $d = 2$). When $\mathcal{S} = \mathbf{C}P(k-1)$, then $a = \mathbf{E}\, X$ will be a positive semi-definite Hermitian matrix as it is a mean of such (forming a convex cone), and an easy calculation shows that $\pi(a)$ will comprise all one-dimensional subspaces of the eigenspace associated to the largest eigenvalue of a. If $\pi(x)$ is unique, one calls x a **non-focal** point of $\mathcal{S}$, else a **focal point**.

Confidence sets. Given i.i.d. random variables $X = X_1, \ldots, X_n$, $n \in \mathbf{N}$, we aim to construct a **confidence set** $K = K(X_1, \ldots, X_n)$ which will cover **all** of M with probability at least $1 - \alpha$ for some predetermined $\alpha \in (0, 1)$, i.e. we want

$$\mathbf{P}(K \supseteq M) \geq 1 - \alpha\,. \tag{3}$$

This can be easily achieved: just construct a $(1-\alpha)$-confidence set B for the ambient mean a, so that $\mathbf{P}(B \ni a) \geq 1 - \alpha$, and define

$$K = \pi(B)\,. \tag{4}$$

One could in fact approximate B using **asymptotics** by the central limit theorem for $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ combined with Slutsky's lemma using a consistent estimator of the covariance of X; note that this differs from the textbook approach which uses the **delta-method** to obtain a central limit theorem for $\pi(\bar{X}_n)$ under the assumption that $\pi(a)$ is a singleton.

Non-asymptotic and universal confidence sets. Asymptotics may be unreliable, however, as one does not know in advance whether the sample size is large enough such that the approximation is sufficiently good. Consider e.g. a distribution composed of two point masses on the same hemisphere, one of which carrying probability $p < 1 - \alpha^{\frac{1}{n}}$. Then, with probability $(1-p)^n > \alpha$ all observations coincide with the other point, leading to a degenerate confidence set composed only of the latter point, i.e. it contains the true spherical mean with probability less than $1 - \alpha$. Therefore, we want the confidence sets to be **non-asymptotic**, guaranteeing coverage for any (fixed) sample size n. Also, we do not want to make any assumptions about the distribution of X, e.g. that $\pi(a)$ is a singleton. In that sense, we want the confidence sets to **universal**. In order to be able to construct such confidence sets, we will assume that $\mathcal{S}$ is bounded; this is fulfilled for the examples above.

2 Mass concentration inequalities

Markov's inequality. The idea is to put a in B if its deviation from the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is not too large, and then set $K = \pi(B)$. To make this precise, we need to determine $t_a > 0$ such that (under the hypothesis $a = \mathbf{E}\, X$)

$$\mathbf{P}(\|\bar{X}_n - a\| > t_a) \leq \alpha\,. \tag{5}$$

For this, recall **Markov's inequality**: if Y is a real-valued, nonnegative random variable, then for any $t > 0$

$$\mathbf{P}(Y > t) \leq \frac{\mathbf{E}\, Y}{t}\,. \tag{6}$$

Choosing $Y = \|\bar{X}_n - a\|^2$ gives **Chebyshev's inequality**:

$$\mathbf{P}(\|\bar{X}_n - a\| > t) = \mathbf{P}\big(\|\bar{X}_n - a\|^2 > t^2\big) \leq \frac{\mathbf{E}\,\|\bar{X}_n - a\|^2}{t^2} = \frac{\mathbf{E}\,\|X - a\|^2}{n^2 t^2}\,. \tag{7}$$

In the special case where $\|X\| = 1$ a.s., which holds in the examples above, in particular for the complex projective space since for a unit vector $x \in \mathbf{C}^k$ we have $\|xx^*\|^2 = \mathbf{tr}\, xx^*(xx^*)^* = \mathbf{tr}(x^*x)(x^*x) = 1$, we get the **total variance**

$$\sigma^2 = \mathbf{E}\,\|X - a\|^2 = \mathbf{E}(X - \mathbf{E}\,X)^*(X - \mathbf{E}\,X) = \mathbf{E}\,X^*X - (\mathbf{E}\,X)^*(\mathbf{E}\,X) = 1 - \|a\|^2 \tag{8}$$

which is small if the so-called **mean resultant length** $\|a\|$ is close to one since the distribution must then be more **concentrated**. We thus obtain the simple formula

$$\mathbf{P}(\|\bar{X}_n - a\| > t) \leq \frac{\sigma^2}{n^2 t^2}\,. \tag{9}$$

Rosenthal type inequalities. Chebyshev's inequality is often quite conservative; it can be improved by going to higher moments which, however, are more difficult to calculate. Using the bound in equation (10) of (Pinelis 2015) we obtain

$$\mathbf{P}(\|\bar{X}_n - a\| > t) \leq \frac{(b\sigma + \sqrt{3}\sigma)^2}{n^4 t^4} \tag{10}$$

which decreases faster in t and n; here, $b = 1 + \|a\|$ is the a.s. bound for the centred random variable $X - a$.

The Cramér-Chernoff method. If $Z = Z_1, \ldots, Z_n$ are i.i.d. real-valued random variables with $|Z| \leq b$ a.s., one can employ Markov's inequality for $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ and $h > 0$ to get

$$\mathbf{P}(\bar{Z}_n - \mathbf{E}\, Z \geq t) = \mathbf{P}\big(e^{h(\bar{Z}_n - \mathbf{E}\, Z)} \geq e^{ht}\big) \leq e^{-ht}\, \mathbf{E}\, e^{h(\bar{Z}_n - \mathbf{E}\, Z)} = e^{-ht}\big(\mathbf{E}\, e^{\frac{h}{n}(Z - \mathbf{E}\, Z)}\big)^n \tag{11}$$

by independence. Estimating $z \mapsto e^{h(z - \mathbf{E}\, Z)}$ by a linear function over $[-b, b]$ and optimising over $h > 0$ leads to **Hoeffding's inequality**, (Hoeffding 1963),

$$\mathbf{P}(\bar{Z}_n - \mathbf{E}\, Z \geq t) \leq \exp\left(-\frac{nt^2}{2b^2}\right) \tag{12}$$

which decreases exponentially fast in t and n; estimating it by a quadratic function gives rise to the **Bennett-Hoeffding inequality**

$$\mathbf{P}(\bar{Z}_n - \mathbf{E}\, Z \geq t) \leq \exp\left(-\frac{n\tau^2}{b^2}\phi\left(\frac{bt}{\tau^2}\right)\right) \tag{13}$$

where $\tau^2 \geq \mathbf{Var}\, Z$ and $\phi(s) = (1+s)\log(1+s) - s$ for $s > 0$. Extending this to higher dimensions is non-trivial, since $e^{\|\bar{X}_n - a\|}$ no longer factorises but Pinelis und Sakhanenko (1986) have nontheless derived the **multivariate Bennett-Hoeffding inequality**

$$\mathbf{P}(\|\bar{X}_n - a\| > t) \leq 2\exp\left(-\frac{n\tau^2}{b^2}\phi\left(\frac{bt}{\tau^2}\right)\right) \tag{14}$$

for $\tau^2 \geq \mathbf{E}\, \|X - a\|^2$.

Critical values. Note that none of these inequalities (Chebyshev, 4th moment Rosenthal, Bennett-Hoeffding) is strictly superior to any of the others. Setting the right hand side equal to α and (numerically, if necessary) solving for t in all cases yields an upper bound to the critical value t_a which we ideally seek; we thus may work with the smallest of these critical values which we will denote by t_a; it will only depend on the mean resultant length $\|a\|$. We then get indeed

$$\mathbf{P}(\|\bar{X}_n - a\| > t_a) \leq \alpha\,. \tag{15}$$

3 Confidence sets for extrinsic means

Assumptions. We assume $X = X_1, \ldots, X_n$ are i.i.d. random variables taking values in a closed set $\mathcal{S} \subseteq \mathbf{R}^d$ which is bounded by 1, i.e. $\mathcal{S}$ is contained in the unit ball of $\mathbf{R}^d$. These assumptions are fulfilled in the examples above.

Construction of the confidence set. For any a in the convex hull C of $\mathcal{S}$ (i.e. a is a possible ambient mean), (numerically) determine t_a using the above inequalities such that, if a were $\mathbf{E}\, X$, $\mathbf{P}(\|\bar{X}_n - a\| > t_a) \leq \alpha$. Now set

$$B = \{a \in C \,:\, \|\bar{X}_n - a\| \leq t_a\}\,, \qquad K = \pi(B)\,. \tag{16}$$

Then $\mathbf{P}(K \supseteq M) \geq 1 - \alpha$, i.e. K is a $(1-\alpha)$-confidence set for the entire set M. Note that, if $\mathcal{S}$ is a subset of the unit sphere, it will be enough to calculate the maximal angle formed by $a \in B$ and $\bar{X}_n$ to get K as the intersection of $\mathcal{S}$ with the corresponding ball around $\pi(\bar{X}_n)$. Details on this construction using Chebyshev's inequality may be found in (Hotz und Kelma 2016).

Question. As opposed to the asymptotic approach were B is an ellipsoid, here it is a ball – does that matter? are ellipsoids "better" than balls? Note that we did take the trace of the covariance into account, but not the individual eigenvalues – should we rather do that?

4 Intrinsic means

Assumptions. We again assume that $X = X_1, \ldots, X_n$ are i.i.d. random variables taking values in some set $\mathcal{S}$ but now $\mathcal{S}$ is a (connected) **compact Riemannian manifold**. The examples above all consitute embedded manifolds, and a (natural) Riemannian metric is given by the restriction of the standard Riemannian metric on $\mathbf{R}^d$ to the tangent spaces of the regular submanifold. The Riemannian metric then induces a distance $d : \mathcal{S} \times \mathcal{S} \to [0, D]$, the so-called **intrinsic metric**, where $D = \max d(\mathcal{S}, \mathcal{S}) < \infty$ is the **diameter** of the manifold.

Intrinsic means. Now, there no longer is a notion of averaging points on the manifold. The idea of Fréchet (1948) is to generalise the characterisation of a Euclidean mean as the minimiser of expected squared distances, and define the **set of intrinsic means** as

$$M = \operatorname*{argmin}_{y \in \mathcal{S}} \mathbf{E}\, d(X, y)^2 \,. \tag{17}$$

This definition is in fact reasonable for any metric, leading to so-called **Fréchet means**. In fact, for the Euclidean distance restricted to an embedded manifold, the **bias-variance decomposition** implies that the corresponding Fréchet means are the extrinsic means considered earlier. But the distance was then measured across the ambient space, possibly outside of $\mathcal{S}$, which is why they are called **extrinsic**, while the intrinsic distance is defined as the minimal length of paths within the manifold, hence leading to **intrinsic** means.

Existence and uniqueness. By compactness of $\mathcal{S}$ and continuity of d, intrinsic means do exist, i.e. $M \neq \emptyset$. The question of uniqueness is non-trivial but it definitely does not hold in general: consider e.g. the uniform distribution on a Riemannian homogeneous space like the examples above.

5 Confidence sets for intrinsic means

Aim. Again, we would like to construct a **confidence set** K covering **all** of M with at least probability $1 - \alpha$. Since asymptotic distributions for intrinsic sample means are difficult to derive and to estimate, and even less reliable for finite sample sizes, we again would like our confidence sets to be **non-asymptotic** and **universal**.

Empirical process approach. Let

$$F : \mathcal{S} \to [0, D^2] \,, \quad y \mapsto F(y) = \mathbf{E}\, d(X, y)^2 \tag{18}$$

be the **Fréchet functional**, so that M is the set of minimisers of F. Then F can be estimated by its empirical analogue, the **empirical Fréchet functional**

$$\hat{F} : \mathcal{S} \to [0, D^2] \,, \quad y \mapsto \hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} d(X_i, y)^2 \,. \tag{19}$$

Then, for any **fixed** $y \in \mathcal{S}$, $\mathbf{E}\, \hat{F}(y) = F(y)$, and the **deviations** $\hat{F}(y) - F(y)$ can be controlled by the **mass concentration inequalities** discussed above since F is bounded. If we now could control the deviation between $\hat{F}$ from F in the supremum norm, we could draw conclusions about the minimisers of F from $\hat{F}$. Indeed, assume $\tau > 0$ has been determined such that

$$\mathbf{P}\Big(\sup_{y \in \mathcal{S}} |\hat{F}(y) - F(y)| > \tau \Big) \leq \alpha \,, \tag{20}$$

then on the event $\{|\hat{F}(y) - F(y)| \leq \tau\}$

$$\max_{y \in M} \hat{F}(y) \leq \max_{y \in M} F(y) + \tau = \min_{y \in \mathcal{S}} F(y) + \tau \leq \min_{y \in \mathcal{S}} \hat{F}(y) + 2\tau \,, \tag{21}$$

so

$$K = \Big\{ y \in \mathcal{S} \,:\, \hat{F}(y) \leq \min_{z \in \mathcal{S}} \hat{F}(z) + 2\tau \Big\} \tag{22}$$

is a $(1 - \alpha)$-confidence set for M.

Covers. For finitely many points $y_1, \dots, y_N \in \mathcal{S}$, $N \in \mathbf{N}$, we can use the **union bound** (or Bonferroni approach) to get

$$\mathbf{P}\Big(\sup_{j=1,\dots,N} |\hat{F}(y_j) - F(y_j)| > t\Big) \le \sum_{j=1}^{N} \mathbf{P}(|\hat{F}(y_j) - F(y_j)| > t) \le 2N \exp\Big(-\frac{nt^2}{2D^4}\Big) \tag{23}$$

where we used Hoeffding's inequality (12) (and the union bound for the two signs) in the last step. Note that N enters the exponential as $\log N$ so that using many points is not too problematic but N needs of course to be finite. If for some given $\varepsilon > 0$ the points $y_1, \dots, y_N$ give an **ε-cover** of $\mathcal{S}$, i.e. the (closed) balls of radius ε around the points cover $\mathcal{S}$, then any point in $\mathcal{S}$ is at most ε from one of the points where we control $\hat{F} - F$. Now both F and $\hat{F}$ are Lipschitz-continuous with constant $2D$, so their difference $\hat{F} - F$ is Lipschitz-continuous with constant $4D$, giving

$$\sup_{y \in \mathcal{S}} |\hat{F}(y) - F(y)| \le \sup_{j=1,\dots,N} |\hat{F}(y_j) - F(y_j)| + 4D\varepsilon\,. \tag{24}$$

The smallest number $N_\varepsilon \in \mathbf{N}$ such that there exists cover of $\mathcal{S}$ with N_ε balls of radius ε is called the **ε-covering number**. Hence, setting

$$t = \sqrt{-\frac{2D^4}{n} \log \frac{\alpha}{2N_\varepsilon}}\,, \qquad \text{and} \qquad \tau = t + 4D\varepsilon \tag{25}$$

in (22) gives the desired confidence set.

Chaining. One might (in advance) optimise τ in (24) over $\varepsilon > 0$, thus ensuring a good estimate on each ball without using too many balls. However, one can do much better by using the fact that due to its Lipschitz continuity, $\hat{F} - F$ will differ very little at nearby points, i.e. for $y, z \in \mathcal{S}$,

$$\big|\hat{F}(z) - F(z) - \big(\hat{F}(y) - F(z)\big)\big| \le 4D\, d(y, z), \tag{26}$$

leading to better estimates in Hoeffding's inequality, and thus a better control of

$$|\hat{F}(z) - F(z)| \le \big|\hat{F}(z) - F(z) - \big(\hat{F}(y) - F(z)\big)\big| + |\hat{F}(y) - F(y)|\,. \tag{27}$$

This suggests to start with a coarse cover, using the Lipschitz-estimate on a finer cover, and again on an even finer cover, continuing *ad infinitum*, or using the crude estimate (25) on the last, small balls. In any case, one obtains confidence sets which will converge to M when n tends to infinity (question: in what sense?), i.e. they are consistent.

Controlling the derivative. Unfortunately, t, and thus also τ, will be at best of order $n^{-\frac{1}{2}}$, while F will increase away from M at best quadratically, so the convergence rate will at best be $n^{-\frac{1}{4}}$! We address this by looking at the derivative which should increase linearly. The problem is that $y \mapsto d(x, y)^2$ is not differentiable if x is at the cut-locus of y. However, if $y \in M$ is an intrinsic mean then F is differentiable at y with vanishing gradient, and furthermore the cut-locus of y carries no mass, so $y \mapsto d(X, y)^2$ is a.s. differentiable at y. But we cannot use chaining since neither do the centres of the balls forming an ε-cover have to be intrinsic means, nor does the derivative have to be continuous outside M. Nonetheless, we may assume that some ε-balls around $y_1, \dots, y_N$ cover M, and every such ball contains some $z_j \in M$, i.e. $d(z_j, y_j) \le \varepsilon$, $j = 1, \dots, N$. Then, setting $\|\nabla \hat{F}(y)\| = \infty$ if $\hat{F}$ is not differentiable at y,

$$\mathbf{P}\Big(\exists j \,:\, \inf_{x \in \mathcal{S}\,:\, d(x, y_j) \le \varepsilon} \|\nabla \hat{F}(x)\| > t\Big) \le \mathbf{P}(\exists j \,:\, \|\nabla \hat{F}(z_j)\| > t) \le \sum_{j=1}^{N} \mathbf{P}(\|\nabla \hat{F}(z_j)\| > t)\,. \tag{28}$$

The latter probabilities can now be estimated using the multivariate Bennett-Hoeffding inequality (14), since $z_j \in M$ we have $\mathbf{E}\nabla \hat{F}(z_j) = \nabla F(z_j) = 0$, $\nabla d(\cdot, z_j)^2$ is a.e. bounded by $2D$, and $\mathbf{E}\,\|\nabla \hat{F}(z_j)\|^2 = F(z_j)$. By choosing t appropriately, we thus can ensure the right hand side of (28) to be at most α. For such t, and ε-balls around $y_1, \dots, y_{N_\varepsilon}$ providing an ε-cover of $\mathcal{S}$,

$$K = \Big\{z \,:\, \exists j = 1, \dots, N_\varepsilon \,:\, d(z, y_j) \le \varepsilon \wedge \inf_{x \in \mathcal{S}\,:\, d(x, y_j) \le \varepsilon} \|\nabla \hat{F}(x)\| > t\Big\} \tag{29}$$

is an $(1 - \alpha)$-confidence set for M.

Combining and iterating. In fact, we need to control both $\hat{F} - F$ and its derivative: the former to ensure consistency by excluding local minima and other critical points, the latter to obtain a good convergence rate. This can be achieved by constructing the two as $\left(1 - \frac{\alpha}{2}\right)$-confidence sets, and then calculating their intersection. In fact, a little thought shows that one needs to control those empirical processes only on M, so after constructing K one only has to consider ε-covers of K, and one can iterate the procedure until convergence. This is admissible because one obtains upper estimates for the correct critical values of the suprema of the empirical processes as long as K covers M – which will be the case unless at least one of the suprema is larger than the correct critical value corresponding to $\frac{\alpha}{2}$.

Notes and questions. Note that in order to construct such confidence sets, one needs to be able to construct the necessary ε-covers, and the confidence set itself will be described by such an ε-cover as well. Since M does not have to be connected (e.g. if $\mathcal{S}$ is a circle and X is equal to any of the three points forming some equilateral triangle with probability $\frac{1}{3}$, M comprises those three points), it is unclear whether one can in general obtain a confidence set with a description that is easier to interpret. Also, there is room for generalisations, e.g. the Fréchet mean with respect to the extrinsic metric, i.e. the extrinsic mean could be treated like this as well, although the F is differentiable in that case, so chaining is possible for ∇F as well, but the result is worse than the one obtained by the direct, geometrical approach. Hence, there is a need for better concentration equalities, and for some kind of chaining for the derivative in the general case. Details on this construction on the circle without chaining may be found in Glock und Hotz (2017).

References

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié, *Annales de l'Institut Henri Poincaré* **10**(4): 215–310.

Glock, M. and Hotz, T. (2017). Constructing universal, non-asymptotic confidence sets for intrinsic means on the circle, *in* F. Nielsen and F. Barbaresco (eds), *Geometric Science of Information - Third International Conference, GSI 2017, Paris, France, November 7-9, 2017, Proceedings*, Vol. 10589 of *Lecture Notes in Computer Science*, Springer, pp. 477–485.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* **58**(301): 13–30.

Hotz, T. and Kelma, F. (2016). Non-asymptotic Confidence Sets for Extrinsic Means on Spheres and Projective Spaces, *arXiv e-prints*. 1602.04117.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*, Wiley, New York.

Pinelis, I. (2015). Rosenthal-type inequalities for martingales in 2-smooth Banach spaces, *Theory of Probability & Its Applications* **59**(4): 699–706.

Pinelis, I. and Sakhanenko, A. (1986). Remarks on inequalities for large deviation probabilities, *Theory of Probability & Its Applications* **30**(1): 143–148.