International Journal for Uncertainty Quantification, 2 (1): 21–38 (2012)

DISTANCES AND DIAMETERS IN CONCENTRATION INEQUALITIES: FROM GEOMETRY TO OPTIMAL ASSIGNMENT OF SAMPLING RESOURCES

***T. J. Sullivan*[1,2,*] *& H. Owhadi*[1,3]**

[1] *Department of Applied and Computational Mathematics, California Institute of Technology, Pasadena, California 91125, USA*

[2] *Graduate Aerospace Laboratories, California Institute of Technology, Pasadena, California 91125, USA*

[3] *Department of Control and Dynamical Systems, California Institute of Technology, Pasadena, California 91125, USA*

Original Manuscript Submitted: 5/3/2011; Final Draft Received: 10/15/2011

This note reviews, compares and contrasts three notions of "distance" or "size" that arise often in concentration-of-measure inequalities. We review Talagrand's convex distance and McDiarmid's diameter, and consider in particular the normal distance *on a topological vector space $\mathcal{X}$, which corresponds to the method of Chernoff bounds, and is in some sense "natural" with respect to the duality structure on $\mathcal{X}$. We show that, notably, with respect to this distance, concentration inequalities on the tails of linear, convex, quasiconvex and measurable functions on $\mathcal{X}$ are mutually equivalent. We calculate the normal distances that correspond to families of Gaussian and of bounded random variables in $\mathbb{R}^N$, and to functions of N empirical means. As an application, we consider the problem of estimating the confidence that one can have in a quantity of interest that depends upon many empirical—as opposed to exact—means and show how the normal distance leads to a formula for the optimal assignment of sampling resources.*

KEY WORDS: *concentration of measure, large deviations, normal distance, optimal sampling, Talagrand distance, uncertainty quantification*

1. INTRODUCTION

It is by now almost classical that smooth enough convex functions enjoy good concentration properties; see, e.g., [1–4] for surveys of the literature. It also is known that convexity can be neglected in the Gaussian case and that the smoothness assumptions are not essential and can be replaced, for instance, with bounded martingale differences; see e.g., [5, 6] and also [7]. Concentration inequalities have found many applications beyond pure mathematics; e.g., in fields such as uncertainty quantification [8], machine learning [9] and distributed computing [10].

Concentration of measure is based on a simple but non-trivial observation originally due to Lévy [11]: in a high-dimensional probability space; "nearly all" the probability mass lies close to any set with measure at least $1/2$. Put another way, functions of many independent variables with small sensitivity to each individual input are very nearly constant. A typical concentration (or deviation) inequality on a space $\mathcal{X}$ is of the form

$$\mathbb{P}[|f(X) - m| \geq r] \leq C_1 \exp(-C_2 r^2), \tag{1}$$

where $f \colon \mathcal{X} \to \mathbb{R}$ is a suitably well-behaved function; X is an $\mathcal{X}$-valued random variable, such that the push-forward measure $(f \circ X)_* \mathbb{P}$ has some concentration property; and m is either the mean value $\mathbb{E}[f(X)]$ or median value $\mathbb{M}[f(X)]$. Some times the control is one sided and the absolute value in Eq. (1) is omitted.

*Correspond to T. J. Sullivan, E-mail: tjs@caltech.edu, URL: http://www.its.caltech.edu/∼tjs/

2152–5080/12/$35.00

A common feature of many concentration results is that an appropriate notion of size or distance is needed; e.g., the McDiarmid diameter [5] or Talagrand's convex distance [12]. This paper reviews both the McDiarmid diameter and Talagrand's distance and discusses, in particular, the distance associated with the method of Chernoff bounds [13], which we term "normal distance." Chernoff bounding is a technique often used in large deviations theory [14–16], in which the measure of a set is estimated using a containing half-space. Although simple, this method leads to a notion of normal distance that is in some sense "natural" with respect to the duality structure on $\mathcal{X}$. Notably, with respect to this distance, concentration inequalities on the tails of linear, convex, quasi-convex, and non-linear functions on $\mathcal{X}$ are mutually equivalent (see Theorem 1).

In Section 5, we identify the normal distance in several commonly encountered cases. In particular, Proposition 3 identifies the normal distance that corresponds to the concentration of a vector, the entries of which are the empirical (sampled) means of functions of independent random variables. In the example that follows it, we consider the problem of estimating the confidence that one can have in a quantity of interest that depends upon such a vector of $N \in \mathbb{N}$ empirical, as opposed to exact, means. In particular, we show how the Chernoff method and normal distance lead to a formula for the optimal assignment of sampling resources to the N to-be-estimated means.

The notation and setting of the paper are covered in Section 2. Section 3 reviews the inequalities and distances of Talagrand and McDiarmid. Normal distance is introduced and its main properties (including Theorem 1) are examined in Section 4. In Section 5, the normal distance is determined explicitly in several cases, thereby connecting Theorem 1 with classical concentration results. In Section 6, it is shown that the equivalent inequalities of Theorem 1 are asymptotically sharp (in the sense used in large deviations theory) in the high-dimensional limit, provided that the sets of interest are convex and "sufficiently round" at those points that are closest to the center of mass $\mathbb{E}[X]$.

2. NOTATION AND BACKGROUND

Throughout, $\mathcal{X}$ will denote a real topological vector space with continuous dual space $\mathcal{X}^*$; $\langle \ell, x \rangle$ denotes the dual pairing between $\ell \in \mathcal{X}^*$ and $x \in \mathcal{X}$; $\langle v, \ell \rangle$ also will denote the dual pairing between $v \in \mathcal{X}^{**}$ and $\ell \in \mathcal{X}^*$. It is not strictly necessary to assume that $\mathcal{X}$ is locally convex, but the results of this paper may be trivially true if $\mathcal{X}^*$ does not contain enough linear functionals.

2.1 Half-Spaces

Given $p \in \mathcal{X}$ and $\nu \in \mathcal{X}^*$, $\mathbb{H}_{p,\nu}$ will denote the closed half-space of $\mathcal{X}$ that has p in its frontier and outward-pointing normal ν; i.e.,

$$\mathbb{H}_{p,\nu} := \{x \in \mathcal{X} \,|\, \langle \nu, x \rangle \leq \langle \nu, p \rangle\}. \tag{2}$$

Note well the degenerate case $\mathbb{H}_{p,0} = \mathcal{X}$. Every $(p, \nu) \in \mathcal{X} \times \mathcal{X}^*$ defines a unique closed half-space of $\mathcal{X}$, whereas a given closed half-space can have multiple distinct representations: $\mathbb{H}_{p,\nu} = \mathbb{H}_{p',\nu'}$ if, and only if, ν is a positive multiple of ν' and $\langle \nu, p - p' \rangle = \langle \nu', p - p' \rangle = 0$.

2.2 Convex Analysis

The closed convex hull of $A \subseteq \mathcal{X}$ will be denoted by $\overline{\text{co}}(A)$. Given a closed convex set $K \subseteq \mathcal{X}$ and $p \in K$, $\mathrm{N}_p^* K$ denotes the outward normal cone to K at p, and $\mathrm{N}^* K$ denotes the outward normal bundle of K:

$$\mathrm{N}_p^* K := \{\nu \in \mathcal{X}^* \,|\, K \subseteq \mathbb{H}_{p,\nu}\}, \tag{3}$$

$$\mathrm{N}^* K := \left\{(p, \nu) \in \mathcal{X} \times \mathcal{X}^* \,\middle|\, p \in K, \nu \in \mathrm{N}_p^* K\right\}. \tag{4}$$

The outward normal cone $\mathrm{N}_p^* K$ is a pointed convex cone: it contains 0, is convex, and $s_1 \nu_1 + s_2 \nu_2 \in \mathrm{N}_p^* K$ for all $s_1, s_2 \geq 0$ and all $\nu_1, \nu_2 \in \mathrm{N}_p^* K$. Also, $\mathrm{N}_p^* K = \{0\}$ if p is an interior point of K. Note that $\mathrm{N}^* K \subseteq \mathcal{X} \times \mathcal{X}^*$ is not necessarily a convex set (see Fig. 1 for an illustration).

For $A \subseteq \mathcal{X}$, χ_A denotes the characteristic function of A, which is convex when A is a convex set:

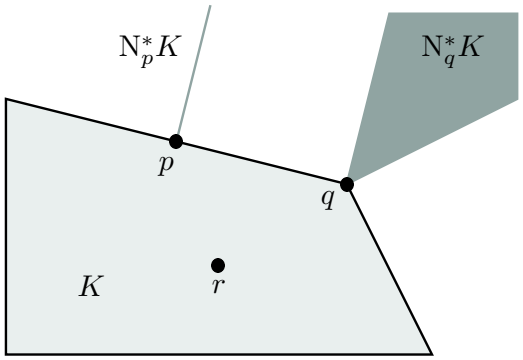

FIG. 1: A convex set K and its outward normal cones at $p, q, r \in K$. ∂K is smooth at $p \in \partial K$, so $\mathrm{N}_p^* K$ is a half-line; ∂K has a vertex at q, so $\mathrm{N}_q^* K$ has non-empty interior; at the interior point r, $\mathrm{N}_r^* K$ is empty.

$$\chi_A(x) := \begin{cases} 0, & \text{if } x \in A, \\ +\infty, & \text{if } x \notin A. \end{cases} \tag{5}$$

For $f\colon \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$, $f^\star\colon \mathcal{X}^* \to \mathbb{R} \cup \{\pm\infty\}$ denotes the Legendre–Fenchel transform or convex conjugate of f, defined by

$$f^\star(\ell) := \sup_{x \in \mathcal{X}} \langle \ell, x \rangle - f(x). \tag{6}$$

If $K \subseteq \mathcal{X}$ is a convex set, then a function $f\colon K \to \mathbb{R} \cup \{\pm\infty\}$ is said to be quasi-convex if, for every $\theta \in \mathbb{R} \cup \{\pm\infty\}$, the sublevel set

$$f^{-1}([-\infty, \theta]) := \{x \in K \mid -\infty \leq f(x) \leq \theta\} \tag{7}$$

is a convex set; equivalently, f is quasi-convex if, for all $x, y \in K$ and $t \in [0, 1]$,

$$f((1-t)x + ty) \leq \max\{f(x), f(y)\}. \tag{8}$$

f is said to be quasi-concave if $-f$ is quasi-convex, and f is said to be quasi-linear if it is both quasi-convex and quasi-concave. Every convex (respectively, concave, linear) function is quasi-convex (respectively, quasi-concave, quasi-linear), but not vice versa. In particular, a function $f\colon \mathbb{R}^N \to \mathbb{R}$ is quasi-linear if, and only if, it is the composition of a monotone function with a linear functional on $\mathbb{R}^N$ [17, p. 122].

2.3 Probabilistic Notions

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to \mathcal{X}$ be an $\mathcal{X}$-valued random variable. $\mathbb{E}[\cdot]$ denotes the expectation operator with respect to the probability measure $\mathbb{P}$: $\mathbb{E}[X]$ is defined to be any $m \in \mathcal{X}$ such that

$$\mathbb{E}[\langle \ell, X - m \rangle] \equiv \int_\Omega \langle \ell, X(\omega) - m \rangle \, \mathrm{d}\mathbb{P}(\omega) = 0 \text{ for all } \ell \in \mathcal{X}^*; \tag{9}$$

if $\mathcal{X}^*$ separates the points of $\mathcal{X}$ (e.g., if $\mathcal{X}$ is a Banach space), then $\mathbb{E}[X]$ is unique. For $Y\colon \Omega \to \mathbb{R}$, any $m \in \mathbb{R}$ that satisfies

$$\sup\left\{v \in \mathbb{R} \,\middle|\, \mathbb{P}[Y \leq v] \leq \frac{1}{2}\right\} \leq m \leq \inf\left\{v \in \mathbb{R} \,\middle|\, \mathbb{P}[Y \leq v] \geq \frac{1}{2}\right\} \tag{10}$$

will be called a median of Y and denoted $\mathbb{M}[Y]$. $M_X\colon \mathcal{X}^* \to [0, +\infty]$ denotes the moment-generating function of X, defined by

$$M_X(\ell) := \mathbb{E}\left[\exp\langle \ell, X \rangle\right] \text{ for all } \ell \in \mathcal{X}^*, \tag{11}$$

and $\Lambda_X(\ell) := \log M_X(\ell)$ denotes the cumulant-generating function (or logarithmic moment-generating function) of X. By Hölder's inequality, Λ_X is a convex function.

3. TALAGRAND'S AND MCDIARMID'S INEQUALITIES

3.1 Talagrand's Inequalities

It has been known for some time that convex sets and functions enjoy good concentration properties; moreover, to get good concentration results, it is necessary to measure distances in the right way.

For example, a theorem of Talagrand shows that if a convex set $K \subseteq \mathbb{R}^N$ occupies a "significant" portion of the Hamming cube $\{-1,+1\}^N$ and $t \gg 1$, then nearly all of the points of the Hamming cube lie within Euclidean distance t of K. More precisely, define the Euclidean Hausdorff distance from $x \in \mathbb{R}^N$ to $A \subseteq \mathbb{R}^N$ by

$$d_{\mathrm{H}}(x, A) := \inf\{\|x - a\|_2 \mid a \in A\}. \tag{12}$$

Talagrand [18] showed that if X is uniformly distributed in $\{-1,+1\}^N$ then, for any $A \subseteq \mathbb{R}^N$, $\mathbb{E}[\exp(d_{\mathrm{H}}(X, \overline{\mathrm{co}}(A))^2/8)] \leq \mathbb{P}[X \in A]^{-1}$; hence, Chebyshev's inequality implies that, for any $t \geq 0$,

$$\mathbb{P}[X \in A]\mathbb{P}[d_{\mathrm{H}}(X, \overline{\mathrm{co}}(A)) \geq t] \leq \exp\left(-\frac{t^2}{8}\right). \tag{13}$$

More interesting results can be obtained if one uses not the Euclidean distance but the Hamming distance—or, more accurately, a supremum over weighted Hamming distances. For $w = (w_1, \ldots, w_N) \in [0, +\infty)^N$, define the w-weighted Hamming distance d_w on a product of sets $\mathcal{X} = \prod_{n=1}^N \mathcal{X}_n$ by

$$d_w(x, y) := \sum_{n=1}^N w_n \mathbf{1}[x_n \neq y_n]; \tag{14}$$

that is, $d_w(x, y)$ is the w-weighted sum of the number of components in which $x, y \in \mathcal{X}$ differ. For $x \in \mathcal{X}$ and $A \subseteq \mathcal{X}$, set $d_w(x, A) := \inf_{a \in A} d_w(x, a)$. Define Talagrand's convex distance from $x \in \mathcal{X}$ to $A \subseteq \mathcal{X}$ by

$$d_{\mathrm{T}}(x, A) := \sup\left\{ d_w(x, A) \,\middle|\, w \in [0,1]^N \text{ and } \sum_{n=1}^N w_n^2 = 1 \right\}, \tag{15}$$

and, for $A, B \subseteq \mathcal{X}$, let $d_{\mathrm{T}}(A, B) := \inf_{a \in A} d_{\mathrm{T}}(a, B)$. Talagrand [12, §4.1] showed that if $X = (X_1, \ldots, X_N)$ is any $\mathcal{X}$-valued random variable with independent components, then

$$\mathbb{P}[X \in A]\mathbb{P}[X \in B] \leq \exp\left(-\frac{d_{\mathrm{T}}(A, B)^2}{4}\right). \tag{16}$$

These bounds on the probabilities of sets lead to deviation inequalities for convex Lipschitz functions. For example (cf. [18, 19]), let X be any random variable in the unit cube in $\mathbb{R}^N$ with independent components, and let $f\colon [0,1]^N \to \mathbb{R}$ be convex and Lipschitz with $\|f\|_{\mathrm{Lip}} \leq 1$; then, for any $t \geq 0$,

$$\mathbb{P}[f(X) \geq \mathbb{M}[f(X)] + t] \leq 2\exp\left(-\frac{t^2}{4}\right). \tag{17}$$

Note, however, that these results use not only the convexity of the function of interest, but also require Lipschitz continuity. What concentration inequalities can be shown to hold without smoothness assumptions?

3.2 McDiarmid's Inequality

One smoothness-free concentration inequality is McDiarmid's inequality [5], also known as the bounded differences inequality, which itself generalizes an earlier inequality of Hoeffding [20]. McDiarmid's inequality is by no means the strongest concentration-of-measure inequality in the literature, but is useful because of its simple hypotheses

and proof. McDiarmid's inequality and its variants have been used for uncertainty quantification in the context of certification [8, 21, 22].

Define the McDiarmid diameter of f, denoted $\mathcal{D}[f]$, by

$$\mathcal{D}[f] := \left(\sum_{n=1}^{N} \mathcal{D}_n[f]^2 \right)^{1/2}, \tag{18}$$

where the nth McDiarmid subdiameter $\mathcal{D}_n[f]$ is defined by

$$\mathcal{D}_n[f] := \sup\{|f(x) - f(y)| \mid x_j = y_j \text{ for } j \neq n\}. \tag{19}$$

When $\mathbb{E}[|f(X)|]$ is finite and $X_1, \ldots, X_N$ are independent, McDiarmid's inequality bounds the deviations of $f(X)$ from $\mathbb{E}[f(X)]$ in terms of the McDiarmid diameter of f: for any $r > 0$,

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \leq -r] \leq \exp\left(-\frac{2r^2}{\mathcal{D}[f]^2} \right), \tag{20a}$$

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq r] \leq \exp\left(-\frac{2r^2}{\mathcal{D}[f]^2} \right). \tag{20b}$$

McDiarmid's inequality implies that, for any $\theta \in \mathbb{R} \cup \{\pm\infty\}$,

$$\mathbb{P}[f(X) \leq \theta] \leq \exp\left(-\frac{2(\mathbb{E}[f(X)] - \theta)_+^2}{\mathcal{D}[f]^2} \right), \tag{21a}$$

$$\mathbb{P}[f(X) \geq \theta] \leq \exp\left(-\frac{2(\theta - \mathbb{E}[f(X)])_+^2}{\mathcal{D}[f]^2} \right), \tag{21b}$$

where, for $t \in \mathbb{R}$, $t_+ := \max\{0, t\}$ and $t_- = -(-t)_+$. McDiarmid's inequality (and similar inequalities such as martingale inequalities) have the advantage that a bound on the tails of $f(X)$ is obtained solely in terms of the mean output $\mathbb{E}[f(X)]$ and the McDiarmid diameter $\mathcal{D}[f]$. However, McDiarmid's inequality cannot take advantage of any other properties of f such as convexity or monotonicity; furthermore, if f has a infinite McDiarmid diameter on the essential range of X, then the trivial upper bound 1 is obtained.

3.3 Other Concentration Inequalities

There is a large body of literature on other sources of concentration-of-measure inequalities: these include logarithmic Sobolev inequalities and the Herbst argument [23–25], the entropy method [26–28], and information-theoretic methods [29, 30]. Of particular interest are those concentration results that apply to infinite-dimensional settings [31].

4. NORMAL DISTANCE

As noted above, efficient presentation of many concentration-of-measure inequalities relies on having an appropriate notion of function variation (e.g., the Lipschitz norm or McDiarmid diameter) or distance (e.g., Talagrand's convex distance). The inequalities that will be established in Section 5 will be phrased in terms of a normal distance, which will be introduced in this section, and is the distance that corresponds to the method of Chernoff bounds.

4.1 Definitions

Fix a function $\Psi\colon \mathcal{X}^* \to [0, +\infty]$ that is positively homogeneous of degree 1; i.e., such that $\Psi(\alpha\ell) = \alpha\Psi(\ell)$ for all $\alpha \geq 0$ and all $\ell \in \mathcal{X}^*$. By analogy with the situation in finite-dimensional Euclidean space, in which $\Psi = \|\cdot\|_2$ on $(\mathbb{R}^N)^*$, define the distance from a point $x \in \mathcal{X}$ to a half-space $\mathbb{H}_{p,\nu} \subseteq \mathcal{X}$ by

$$d_{\perp,\Psi}(x, \mathbb{H}_{p,\nu}) := \frac{\langle \nu, x - p \rangle_+}{\Psi(\nu)}, \tag{22}$$

with the convention that $0/0 = 0$, since the distance from $x \in \mathcal{X}$ to the trivial half-space $\mathbb{H}_{p,0} = \mathcal{X}$ ought to be zero. Note that $d_{\perp,\Psi}(x, \mathbb{H}_{p,\nu}) = 0$ whenever $x \in \mathbb{H}_{p,\nu}$; note also that the homogeneity assumption on Ψ ensures that Eq. (22) is an unambiguous definition. We now generalize Eq. (22) to more general subsets of $\mathcal{X}$ than half-spaces. The heuristic is that the distance from x to $A \subseteq \mathcal{X}$ should be the greatest possible distance [in the sense of Eq. (22)] from x to any half-space that contains A; the existence of the degenerate half-space $\mathbb{H}_{p,0}$ ensures that the normal distance is zero if there are no proper half-spaces that contain A.

4.1.1 Definitions 1

Let $x \in \mathcal{X}$ and $A \subseteq \mathcal{X}$. The Ψ-normal distance from x to A, denoted $d_{\perp,\Psi}(x, A)$, is defined (with the same convention that $0/0 = 0$) by

$$d_{\perp,\Psi}(x, A) := \sup \left\{ \frac{\langle \nu, x - p \rangle_+}{\Psi(\nu)} \,\middle|\, \begin{array}{l} p \in \mathcal{X} \text{ and } \nu \in \mathcal{X}^* \\ \text{such that } A \subseteq \mathbb{H}_{p,\nu} \end{array} \right\}. \tag{23}$$

The Ψ-normal distance from $A \subseteq \mathcal{X}$ to $B \subseteq \mathcal{X}$ is defined by $d_{\perp,\Psi}(A, B) := \inf_{a \in A} d_{\perp,\Psi}(a, B)$. In the special case $\mathcal{X} = \mathbb{R}^N$ and $\Psi = \| \cdot \|_2$ on $(\mathbb{R}^N)^*$, we shall simply write $d_\perp$ for $d_{\perp,\Psi}$; i.e.,

$$d_\perp(x, A) := \sup \left\{ \frac{[\nu \cdot (x - p)]_+}{\|\nu\|_2} \,\middle|\, \begin{array}{l} p \in \mathbb{R}^N \text{ and } \nu \in (\mathbb{R}^N)^* \\ \text{such that } A \subseteq \mathbb{H}_{p,\nu} \end{array} \right\}. \tag{24}$$

Note well that the definition of the normal distance $d_{\perp,\Psi}(x, A)$ does not require $\mathcal{X}$ to be normed; even when $\mathcal{X}$ is equipped with a norm $\| \cdot \|_{\mathcal{X}}$ and Ψ is the corresponding operator norm, the normal distance $d_{\perp,\Psi}(x, A)$ is not the same as the Hausdorff distance from x to A defined by

$$d_{\mathrm{H}}(x, A) := \inf\{\|x - a\|_{\mathcal{X}} \mid a \in A\}; \tag{25}$$

(see Fig. 2 for an illustration). Note, also, that it is not generally true that $d_{\perp,\Psi}(A, B) = d_{\perp,\Psi}(B, A)$: consider, e.g., $B := \{(0, 1)\}$ and A as in Fig. 2, in which case

$$d_{\perp,\Psi}(A, B) = \inf_{a \in A} d_{\perp,\Psi}(a, B) = 1 \neq 0 = d_{\perp,\Psi}(B, A).$$

For any $x \in \mathcal{X}$ and $A \subseteq B \subseteq \mathcal{X}$, it holds that $d_{\perp,\Psi}(x, B) \leq d_{\perp,\Psi}(x, A)$. Furthermore, since a closed half-space $\mathbb{H}_{p,\nu}$ contains A if, and only if, it contains the closed convex hull $\overline{\mathrm{co}}(A)$ of A, the following equality holds:

$$d_{\perp,\Psi}(x, A) = d_{\perp,\Psi}(x, \overline{\mathrm{co}}(A)) \text{ for all } x \in \mathcal{X} \text{ and all } A \subseteq \mathcal{X}. \tag{26}$$

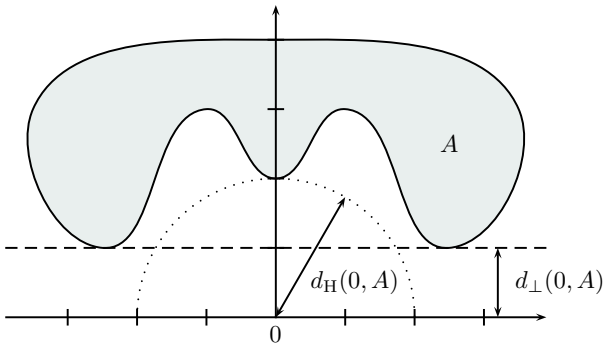

FIG. 2: An example of a subset A of the Euclidean plane $\mathbb{R}^2$ for which the normal distance $d_\perp(0, A) = 1$ unit (*cf.* the dashed line), as opposed to the Euclidean Hausdorff distance $d_{\mathrm{H}}(0, A) = 2$ units (*cf.* the dotted arc). Also, $d_{\mathrm{T}}(0, A) = 1$, with the supremum in (15) being attained by $w = (0, 1)$.

4.2 Comparison of Normal and Talagrand Distances

A full comparison of the normal distance and Talagrand's convex distance is not possible, since each belongs to a different setting: Talagrand's distance is defined on a product of sets, whereas the normal distance is defined on a topological vector space that might not be a product space.

On $\mathbb{R}^N$ with its usual (product) Euclidean structure, the two distances can be compared. It is immediately apparent that the two distances measure different quantities: in some sense, $d_{\mathrm{T}}(x, A)$ measures how many of the coordinates of x are covered by A, but does not measure the geometric distance between them; on the other hand, $d_{\perp,\Psi}(x, A)$ is a much more geometric measure of how far x is from A in terms of linear functionals on $\mathcal{X}$, and the "size" of those linear functionals is measured by Ψ. In particular, Talagrand's convex distance is positively homogeneous of degree 0, whereas the normal distance is positively homogeneous of degree 1: for any $x \in \mathbb{R}^N$, $A \subseteq \mathbb{R}^N$, and $\alpha > 0$,

$$d_{\mathrm{T}}(\alpha x, \alpha A) = d_{\mathrm{T}}(x, A),$$

$$d_{\perp,\Psi}(\alpha x, \alpha A) = \alpha d_{\perp,\Psi}(x, A).$$

Indeed, for a half-space $\mathbb{H}_{p,\nu} \subseteq \mathbb{R}^N$ and weight $w = (w_1, \ldots, w_N)$,

$$d_w(0, \mathbb{H}_{p,\nu}) = \begin{cases} 0, & \text{if } x \in \mathbb{H}_{p,\nu}, \\ \min\{w_n \mid \nu_n \neq 0\}, & \text{if } x \notin \mathbb{H}_{p,\nu}, \end{cases}$$

and, hence, $d_{\mathrm{T}}(x, \mathbb{H}_{p,\nu}) = \mathbf{1}[x \notin \mathbb{H}_{p,\nu}]$: the supremum in Eq. (15) is attained by any weight w that has $w_n = 1$ for some n with $\nu_n \neq 0$, and $w_n = 1$ otherwise.

4.3 Portmanteau Theorem

The geometrical nature of the normal distance, in particular, formula (26), leads to the following equivalence or "portmanteau" theorem for deviation inequalities with respect to $d_{\perp,\Psi}$. In practice, as noted at the beginning of the next section, these inequalities are unlikely to be sharp; their utility lies in the fact that they are geometrically easy to work with.

4.3.1 Theorem 1

Fix $\Psi\colon \mathcal{X}^* \to [0, +\infty]$, homogeneous of degree 1, and let $d_{\perp,\Psi}$ be the corresponding normal distance. For an $\mathcal{X}$-valued random variable X and measurable function $f\colon \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$, consider the inequalities

$$\mathbb{P}[X \in A] \leq \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], A)^2}{2}\right), \tag{27}$$

$$\mathbb{P}[f(X) \leq \theta] \leq \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], f^{-1}([-\infty, \theta]))^2}{2}\right). \tag{28}$$

Then, the following formulations are equivalent:

1. Equation (27) holds for every half-space $A = \mathbb{H}_{p,\nu}$.
2. Equation (27) holds for every convex $A \subseteq \mathcal{X}$.
3. Equation (27) holds for every measurable $A \subseteq \mathcal{X}$.
4. Equation (28) holds for every measurable $f\colon \mathcal{X} \to \mathbb{R}$ and $\theta \in \mathbb{R} \cup \{\pm\infty\}$.
5. Equation (28) holds for quasiconvex $f\colon \mathcal{X} \to \mathbb{R}$ and $\theta \in \mathbb{R} \cup \{\pm\infty\}$.
6. Equation (28) holds for every continuous linear $f\colon \mathcal{X} \to \mathbb{R}$ and $\theta \in \mathbb{R} \cup \{\pm\infty\}$.

Note that if f is quasilinear, then formulation 5 yields concentration inequalities for both the lower and upper tails of $f(X)$.

4.3.2 Proof of Theorem 1

The equivalence will be established by showing that

$$1 \implies 2 \implies 3 \implies 4 \implies 5 \implies 6 \implies 1.$$

Suppose that formulation 1 holds and that $K \subseteq \mathcal{X}$ is convex. Then

$$\begin{aligned}
\mathbb{P}\ \ [X \in K] & \\
&\leq \inf_{\mathbb{H}_{p,\nu} \supseteq K} \mathbb{P}[X \in \mathbb{H}_{p,\nu}] && \text{by monotonicity of } \mathbb{P}, \\
&\leq \inf_{\mathbb{H}_{p,\nu} \supseteq K} \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], \mathbb{H}_{p,\nu})^2}{2}\right) && \text{by formulation 1}, \\
&= \exp\left(-\frac{1}{2} \sup_{\mathbb{H}_{p,\nu} \supseteq K} d_{\perp,\Psi}(\mathbb{E}[X], \mathbb{H}_{p,\nu})^2\right) & \\
&= \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], K)^2}{2}\right) && \text{by Eq. (23)}.
\end{aligned}$$

Hence, formulation 1 implies formulation 2.

Suppose that formulation 2 holds and that $A \subseteq \mathcal{X}$ is measurable. Then

$$\begin{aligned}
\mathbb{P}[X \in A] &\leq \mathbb{P}[X \in \overline{\text{co}}(A)] && \text{since } A \subseteq \overline{\text{co}}(A), \\
&\leq \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], \overline{\text{co}}(A))^2}{2}\right) && \text{by formulation 2}, \\
&= \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], A)^2}{2}\right) && \text{by Eq. (26)},
\end{aligned}$$

and so formulation 2 implies formulation 3. Formulation 4 follows from formulation 3 upon setting $A := \{x \in \mathcal{X} \mid f(x) \leq \theta\}$. Formulation 5 is clearly a special case of formulation 4. Every linear function has convex sublevel sets, and so formulation 5 implies formulation 6. Formulation 1 follows from formulation 6 upon setting $f := \nu$ and $\theta := \langle \nu, p \rangle$.

4.3.3 Remark 1

It is important to note that all the bounds in Theorem 1 may be trivial if the dual space $\mathcal{X}^*$ is not rich enough. For example, given a measure space $(\mathcal{Z}, \mathscr{F}, \mu)$, for $0 < p < 1$, the space

$$\mathcal{L}^p(\mathcal{Z}, \mathscr{F}, \mu; \mathbb{R}) := \left\{ f \colon \mathcal{Z} \to \mathbb{R} \;\middle|\; \|f\|_p := \left(\int_{\mathcal{Z}} |f(z)|^p \, \mathrm{d}\mu(z) \right)^{1/p} < +\infty \right\}$$

is a topological vector space with respect to the quasi-norm topology generated by $\| \cdot \|_p$. This space is not locally convex and has a trivial dual space: the only continuous linear functional on this space is the zero functional, and so the only closed half-space is the whole space. See, e.g., [32, Section 1.47] for further discussion of spaces such as $\mathcal{L}^p([0,1]; \mathbb{R})$ for $0 < p < 1$.

It is tempting to eliminate these pathologies by working with the algebraic, instead of the topological, dual of $\mathcal{X}$. This can be done, and most results go through mutatis mutandis; in particular, it is necessary to replace all references to the closed convex hull $\overline{\text{co}}(A)$ of $A \subseteq \mathcal{X}$ with the convex hull $\text{co}(A)$; the analog of Eq. (26) (with Ψ now defined on the algebraic dual of $\mathcal{X}$) is

$$d_{\perp,\Psi}(x, A) = d_{\perp,\Psi}(x, \text{co}(A)) \text{ for all } x \in \mathcal{X} \text{ and all } A \subseteq \mathcal{X}.$$

The principal disadvantage of ignoring all topological structure on $\mathcal{X}$ is that there are no longer notions of interior, closure, and frontier—although it still makes sense to discuss the extremal points of convex sets.

5. NORMAL DISTANCE AS A CONCENTRATION RATE

The method of Chernoff bounding (reviewed in Lemma 1) gives bounds on $\mathbb{P}[X \in \mathbb{H}_{p,\nu}]$ in terms of the moment-generating function M_X. If these bounds can be formulated in terms of a suitable normal distance, then Theorem 1 produces equivalent bounds for on $\mathbb{P}[X \in K]$ for convex K, on $\mathbb{P}[X \in A]$ for measurable A, and so on. As noted in [2, Section 2], the best Chernoff bound on $\mathbb{P}[f(X) \geq \theta]$ is never better than the best bound using all the moments of $f(X)$: if f takes only non-negative values, then

$$\inf_{k \in \mathbb{N}} \theta^{-k} \mathbb{E}\big[f(X)^k\big] \leq \inf_{s \geq 0} e^{-s\theta} \mathbb{E}\big[e^{sf(X)}\big]. \tag{29}$$

However, Chernoff bounds have the advantage that they are geometrically very easy to handle.

5.1 Chernoff Bounds

The method of Chernoff bounds [13, 17, Section 7.4.3] is a simple one in which the probability of a subset of $\mathcal{X}$ is bounded by that of a containing half-space, and the probability of that half-space is bounded using the moment-generating function of the probability measure.

5.1.1 Lemma 1: Chernoff Bounds

For any half-space $\mathbb{H}_{p,\nu} \subseteq \mathcal{X}$,

$$\mathbb{P}[X \in \mathbb{H}_{p,\nu}] \leq \inf_{s \geq 0} e^{s\langle \nu, p \rangle} M_X(-s\nu). \tag{30}$$

For any convex set $K \subseteq \mathcal{X}$,

$$\mathbb{P}[X \in K] \leq \inf_{(p,\nu) \in \mathrm{N}^* K} e^{\langle \nu, p \rangle} M_X(-\nu) \tag{31a}$$

$$= \exp\left(- \sup_{p \in K} (\Lambda_X + \chi_{-\mathrm{N}_p^* K})^\star(p) \right). \tag{31b}$$

In particular, for any $x \in \mathcal{X}$,

$$\mathbb{P}[X = x] \leq \exp(-\Lambda_X^\star(x)). \tag{32}$$

5.1.2 Proof

By the definition of the half-space $\mathbb{H}_{p,\nu}$,

$$\begin{aligned}
\mathbb{P}\left[X \in \mathbb{H}_{p,\nu}\right] &= \mathbb{P}\left[\langle \nu, X \rangle \leq \langle \nu, p \rangle\right] \\
&= \mathbb{E}\left[\mathbf{1}_{[\langle \nu, p - X \rangle \geq 0]}\right] \\
&\leq \mathbb{E}\left[e^{s\langle \nu, p - X \rangle}\right] \qquad \text{for any } s \geq 0, \\
&= e^{s\langle \nu, p \rangle} \mathbb{E}\left[e^{\langle -s\nu, X \rangle}\right] \\
&\leq e^{s\langle \nu, p \rangle} M_X(-s\nu).
\end{aligned}$$

Since this inequality holds for any $s \geq 0$, taking the infimum over all such s yields Eq. (30). Recall that the outward normal cone to a convex set at any point is closed under multiplication by non-negative scalars; hence, for any convex set $K \subseteq \mathcal{X}$, taking the infimum of the right-hand side of Eq. (30) over half-spaces $\mathbb{H}_{p,\nu}$ that contain K yields Eq. (31a). Now observe that

$$\begin{aligned}
&\inf\nolimits_{(p,\nu) \in \mathrm{N}^* K} e^{\langle \nu, p \rangle} M_X(-\nu) \\
&= \inf\nolimits_{(p,\nu) \in \mathrm{N}^* K} \exp[\langle \nu, p \rangle + \Lambda_X(-\nu)] \\
&= \exp\Big(\inf\nolimits_{p \in K} \inf\nolimits_{\nu \in \mathrm{N}_p^* K} \left[\langle \nu, p \rangle + \Lambda_X(-\nu)\right]\Big) \\
&= \exp\left(-\sup\nolimits_{p \in K} \sup\nolimits_{\nu \in -\mathrm{N}_p^* K} \left[\langle \nu, p \rangle - \Lambda_X(\nu)\right]\right) \\
&= \exp\left(-\sup\nolimits_{p \in K} (\Lambda_X + \chi_{-\mathrm{N}_p^* K})^\star(p)\right),
\end{aligned}$$

which establishes Eq. (31b); Eq. (32) follows as a special case.

5.2 Families of Gaussian Random Variables

The next result provides the normal distance for an $\mathcal{X}$-valued Gaussian random variable (in fact, for a family of such variables). In the special case of a single Gaussian random vector X on $\mathcal{X} = \mathbb{R}^N$ with covariance operator $C_X = \sigma \mathbb{I}_N$, Proposition 1 yields the classical Chernoff bound for a multivariate normal random variable.

5.2.1 Proposition 1

Let Γ be a family of Gaussian random vectors in $\mathcal{X}$. For each $X \in \Gamma$, let $C_X \colon \mathcal{X}^* \to \mathcal{X}^{**}$ be its covariance operator defined by

$$\langle C_X \ell, \nu \rangle := \mathbb{E}\left[\langle \ell, X \rangle \langle \nu, X \rangle\right]. \tag{33}$$

Let $E := \{\mathbb{E}[X] \mid X \in \Gamma\}$, let

$$\Psi(\nu) := \sup_{X \in \Gamma} \sqrt{\langle C_X \nu, \nu \rangle}, \tag{34}$$

and let $d_{\perp,\Psi}$ be the corresponding normal distance. Then, for any $A \subseteq \mathcal{X}$,

$$\sup_{X \in \Gamma} \mathbb{P}[X \in A] \leq \exp\left(-\frac{d_{\perp,\Psi}(E, A)^2}{2}\right). \tag{35}$$

5.2.2 Proof

For each $X \in \Gamma$, the moment-generating function for X is given by

$$M_X(\ell) := \mathbb{E}\left[e^{\langle \ell, X \rangle}\right] = \exp\left(\langle \ell, \mathbb{E}[X] \rangle + \frac{\langle C_X \ell, \ell \rangle}{2}\right). \tag{36}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\ [X \in \mathbb{H}_{p,\nu}] \\
&\quad \leq \inf_{s \geq 0} \exp\left(s\langle \nu, p - \mathbb{E}[X] \rangle + s^2 \frac{\langle C_X \nu, \nu \rangle}{2}\right) \quad \text{by Eq. (36) and Lemma 1,} \\
&\quad = \exp\left(-\frac{\langle \nu, \mathbb{E}[X] - p \rangle_+^2}{2\langle C_X \nu, \nu \rangle^2}\right) \\
&\quad \leq \exp\left(-\frac{\langle \nu, \mathbb{E}[X] - p \rangle_+^2}{2\Psi(\nu)^2}\right) \qquad \text{by Eq. (34),} \\
&\quad = \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], \mathbb{H}_{p,\nu})^2}{2}\right) \qquad \text{by Eq. (23).}
\end{aligned}$$

Hence, by Theorem 1,

$$\mathbb{P}[X \in A] \leq \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], A)^2}{2}\right),$$

and so

$$\sup_{X \in \Gamma} \mathbb{P}[X \in A] \leq \sup_{X \in \Gamma} \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], A)^2}{2}\right) = \exp\left(-\inf_{X \in \Gamma} \frac{d_{\perp,\Psi}(\mathbb{E}[X], A)^2}{2}\right) = \exp\left(-\frac{d_{\perp,\Psi}(E, A)^2}{2}\right).$$

5.3 Families of Bounded Random Variables

Lemma 1 also has the following consequences for random vectors supported in a cuboid in $\mathbb{R}^N$; this encompasses two standard situations in which concentration is often studied, namely concentration for functions on the Euclidean unit cube and on the Hamming cube.

5.3.1 Proposition 2

Let X be a random vector in $\mathbb{R}^N$ with independent components such that each component X_n almost surely takes values in a fixed interval of length L_n. Let

$$\Psi(\nu) := \frac{1}{2}\sqrt{\sum_{n=1}^{N} L_n^2 \nu_n^2} \tag{37}$$

and let $d_{\perp,\Psi}$ be the corresponding normal distance. Then, for any $A \subseteq \mathbb{R}^N$,

$$\mathbb{P}[X \in A] \leq \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[X], A)^2}{2}\right). \tag{38}$$

A fortiori, if X takes values in (a translate of) the unit cube $[0,1]^N$, then

$$\mathbb{P}[X \in A] \leq \exp\left(-2d_{\perp}(\mathbb{E}[X], A)^2\right), \tag{39}$$

and if X takes values in (a translate of) the Hamming cube $\{-1,+1\}^N$, then

$$\mathbb{P}[X \in A] \leq \exp\left(-\frac{d_{\perp}(\mathbb{E}[X], A)^2}{2}\right). \tag{40}$$

5.3.2 Proof

The proof is similar to the Gaussian case: it is an application of Lemma 1 and Hoeffding's lemma [20, Lemma 1 and Eq. (4.16)], which bounds the moment-generating function of X_n as follows:

$$M_{X_n}(\ell_n) := \mathbb{E}\left[\exp(\ell_n X_n)\right] \leq \exp\left(\ell_n \mathbb{E}[X_n] + \frac{\ell_n^2 L_n^2}{8}\right).$$

Note that the claim also can be proved by applying McDiarmid's inequality to the function $\langle \nu, \cdot \rangle$, which has mean $\mathbb{E}[\langle \nu, X \rangle] = \langle \nu, \mathbb{E}[X] \rangle$ and McDiarmid diameter $\sqrt{L_1^2 + \cdots + L_N^2}$.

5.3.3 Remark 2

Note the similarity between the normal distances of Propositions 1 and 2. In the Gaussian case, the norm on $\mathcal{X}^*$ is the one induced by the "largest" covariance operator in the family of random variables Γ. In the bounded-range case, the norm on $\mathcal{X}^*$ is the one induced by the largest covariance operator for random variables satisfying the range constraint: if X is a real-valued random variable taking values in an interval $[a,b]$, then $\Psi(\nu)^2 = (1/4)(b-a)^2\nu^2$ and $\mathrm{Var}[X] \leq (1/4)(b-a)^2$; this upper bound on the variance is attained by a Bernoulli random variable with law $(1/2)\delta_a + (1/2)\delta_b$.

5.4 Functions of Empirical Means

The next result identifies the normal distance that corresponds to the concentration of a vector, the entries of which are the empirical (sampled) means of functions of independent random variables.

5.4.1 Proposition 3

For $n = 1, \ldots, N$, let $Z_n := f_n(Y_{n,1}, \ldots, Y_{n,K(n)})$ be a real-valued function of independent random variables $Y_{n,k}$, and suppose that f_n has finite McDiarmid diameter $\mathcal{D}[f_n]$. Let $Z = (Z_1, \ldots, Z_N)$. Suppose that the random inputs of each f_n are sampled independently $M(n)$ times according to the distribution $\mathbb{P}$ and that the empirical average

$$\widehat{\mathbb{E}}[Z] = \left(\frac{1}{M(n)} \sum_{m=1}^{M(n)} f_n \left(Y_{n,1}^{(m)}, \ldots, Y_{n,K(n)}^{(m)} \right) \right)_{n=1}^{N} \in \mathbb{R}^N \tag{41}$$

is formed. Then, for any $A \subseteq \mathbb{R}^N$,

$$\mathbb{P}\left[\widehat{\mathbb{E}}[Z] \in A \right] \leq \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[Z], A)^2}{2} \right), \tag{42}$$

where $\Psi \colon (\mathbb{R}^N)^* \to [0, +\infty)$ is given in terms of the McDiarmid diameters of the functions $f_1, \ldots, f_N$ and the sample sizes $M(1), \ldots, M(N)$:

$$\Psi(\nu) := \frac{1}{2} \left(\sum_{n=1}^{N} \frac{\nu_n^2 \mathcal{D}[f_n]^2}{M(n)} \right)^{1/2}. \tag{43}$$

5.4.2 Proof

Let $\mathbb{H}_{p,\nu} \subsetneq \mathbb{R}^N$ be a half-space. Consider the real-valued random variable $\left\langle \nu, \widehat{\mathbb{E}}[Z] \right\rangle$ as a function of the sampled input random variables $Y_{n,k}^{(m)}$. Suppose that the McDiarmid subdiameter of f_n with respect to $Y_{n,k}$ is $D_{n,k}$. Then the McDiarmid subdiameter of $\left\langle \nu, \widehat{\mathbb{E}}[Z] \right\rangle$ with respect to the mth sample of $Y_{n,k}$ is $\nu_n D_{n,k}/M(n)$. Hence, the McDiarmid diameter of $\left\langle \nu, \widehat{\mathbb{E}}[Z] \right\rangle$ is

$$\sqrt{\sum_{k,n,m} \frac{\nu_n^2 D_{n,k}^2}{M(n)^2}} = \sqrt{\sum_{n,m} \frac{\nu_n^2 \mathcal{D}[f_n]^2}{M(n)^2}} = \sqrt{\sum_{n} \frac{\nu_n^2 \mathcal{D}[f_n]^2}{M(n)}}$$

Therefore, since $\widehat{\mathbb{E}}[Z]$ is an unbiased estimator for $\mathbb{E}[Z]$ (i.e., $\mathbb{E}\left[\widehat{\mathbb{E}}[Z]\right] = \widehat{\mathbb{E}}[Z]$), McDiarmid's inequality (21a) implies that

$$\begin{aligned} \mathbb{P}\left[\widehat{\mathbb{E}}[Z] \in \mathbb{H}_{p,\nu} \right] &= \mathbb{P}\left[\left\langle \nu, \widehat{\mathbb{E}}[Z] \right\rangle \leq \langle \nu, p \rangle \right] \leq \exp\left(-\frac{2\left(\langle \nu, \mathbb{E}[Z] \rangle - \langle \nu, p \rangle \right)_+^2}{\sum_{n=1}^{N} (\nu_n^2 \mathcal{D}[f_n]^2)/M(n)} \right) \\ &= \exp\left(-\frac{\langle \nu, \mathbb{E}[Z] - p \rangle_+^2}{2 \cdot (1/4) \cdot \sum_{n=1}^{N} (\nu_n^2 \mathcal{D}[f_n]^2)/M(n)} \right) = \exp\left(-\frac{d_{\perp,\Psi}(\mathbb{E}[Z], \mathbb{H}_{p,\nu})^2}{2} \right). \end{aligned}$$

The claim now follows from Theorem 1.

An example of the application of Proposition 3 is the following.

5.4.3 Example 1: Functions of Empirical Means

The Chernoff bounding method can be used to provide much-improved confidence levels for quantities derived from many empirical—as opposed to exact—means. For example, consider the problem of [33, Section 5]: an input parameter space $\mathcal{X}$ is partitioned into N sub-rectangles, and the probability that a function of interest $\phi \colon \mathcal{X} \to \mathbb{R}$ takes values below $\theta \in \mathbb{R}$ is bounded from above using the following variant of McDiarmid's inequality:

$$\mathbb{P}[\phi \leq \theta] \leq \sum_{n=1}^{N} \mathbb{P}(A_n) \exp\left(-\frac{2(\mathbb{E}[\phi|_{A_n}] - \theta)_+^2}{\mathcal{D}[\phi|_{A_n}]} \right). \tag{44}$$

Suppose, however, that the local (conditioned) means $\mathbb{E}[\phi|_{A_n}]$ are not known exactly; instead, through a finite number of independent samples, the empirical means $\widehat{\mathbb{E}}[\phi|_{A_n}]$ are known. Given $\alpha_1, \ldots, \alpha_N > 0$, with what probability is it true that

$$\mathbb{P}[\phi \leq \theta] \leq \sum_{n=1}^{N} \mathbb{P}(A_n) \exp\left(-\frac{2\left(\widehat{\mathbb{E}}[\phi|_{A_n}] + \alpha_n - \theta\right)_+^2}{\mathcal{D}[\phi|_{A_n}]} \right) ? \tag{45}$$

Furthermore, consider the following problem of optimal allocation of sampling resources: suppose that all the terms but the empirical means $\widehat{\mathbb{E}}[\phi|_{A_n}]$ are known, and that a prescribed total number of samples—M, say—are available for sampling these N means; how should those M samples be assigned to those N "bins" (i.e., to the various subsets A_n) so as to maximize the probability that (45) holds true?

More generally, suppose that $H_0 \colon \mathbb{R}^N \to \mathbb{R}$ is some function of interest: in particular, the quantity of interest is $H_0\left(\mathbb{E}[Z_1], \ldots, \mathbb{E}[Z_N]\right)$ for some absolutely integrable real-valued random variables $Z_1, \ldots, Z_N$. Bear in mind that different Z_n may be physically incomparable: for example, Z_1 may have units of area, Z_2 may have units of temperature, and so on. Therefore, it is not immediately obvious how to combine such apparently incommensurable uncertainties.

If the exact means $\mathbb{E}[Z_n]$ are unknown, then empirical means $\widehat{\mathbb{E}}[Z_n]$ may be used in their place if appropriate confidence corrections are made. Suppose that "error" corresponds to concluding, based on the empirical means, that $H_0(\mathbb{E}[Z])$ is smaller than it actually is. Given $\alpha \in \mathbb{R}^N$, set

$$H_\alpha(z_1, \ldots, z_N) := H_0(z_1 + \alpha_1, \ldots, z_N + \alpha_N). \tag{46}$$

Therefore, given any $\varepsilon > 0$, we seek an appropriate "margin hit" $\alpha = \alpha(\varepsilon) \in \mathbb{R}^N$ (typically, $\alpha_n \geq 0$ for each $n \in \{1, \ldots, N\}$) such that

$$\mathbb{P}\left[H_\alpha\left(\widehat{\mathbb{E}}[Z_1], \ldots, \widehat{\mathbb{E}}[Z_N] \right) \geq H_0\left(\mathbb{E}[Z_1], \ldots, \mathbb{E}[Z_N] \right) \right] \geq 1 - \varepsilon.$$

Dually, given $\alpha \in \mathbb{R}^N$, we seek a sharp upper bound on the probability of error; i.e., on

$$\mathbb{P}\left[H_\alpha\left(\widehat{\mathbb{E}}[Z_1], \ldots, \widehat{\mathbb{E}}[Z_N] \right) \leq H_0\left(\mathbb{E}[Z_1], \ldots, \mathbb{E}[Z_N] \right) \right].$$

If H_0 (and, hence, H_α) is monotonic in each of its N arguments and $Z_1, \ldots, Z_N$ are independent, then the probability of non-error can be bounded from below as follows:

$$\begin{aligned} \mathbb{P}\left[H_\alpha\left(\widehat{\mathbb{E}}[Z] \right) \leq H_0(\mathbb{E}[Z]) \right] &= \mathbb{P}\left[H_\alpha\left(\widehat{\mathbb{E}}[Z] \right) \leq H_\alpha(\mathbb{E}[Z] - \alpha) \right] \leq \textstyle\prod_{n=1}^{N} \mathbb{P}\left[\widehat{\mathbb{E}}[Z_n] \leq \mathbb{E}[Z_n] - \alpha_n \right] \\ &\leq 1 - \prod_{n=1}^{N} \left[1 - \exp\left(-\frac{-2M(n)(\alpha_n)_+^2}{\mathcal{D}[f_n]^2} \right) \right]. \end{aligned}$$

Unfortunately, when N is large, the last line of this inequality, typically, is close to zero unless the sample sizes are very large, and so this bound is of very limited utility. Geometrically, this is analogous to the fact that a high-dimensional orthant (product of half-lines) appears to be very narrow from the perspective of an observer at its vertex. In contrast, half-spaces always fill a half of the observer's field of view. To bound the probability of sublevel or superlevel sets using half-spaces requires H_α to have some convexity—not monotonicity—properties.

If H_α is quasi-convex, then the bounds using normal distances can be applied to good effect, and yield estimates that actually perform better the larger N is. In particular, if H_α is both quasi-convex and differentiable, then the outward normal to its t-level set at some point p is just any positive multiple of the derivative of H_α at p, and this yields the bound

$$\mathbb{P}\left[H_\alpha\left(\widehat{\mathbb{E}}[Z] \right) \leq \theta \right] \leq \inf_{p : H_\alpha(p) \leq \theta} \exp\left(-\frac{2\left(\sum_{n=1}^{N} \partial_n H_\alpha(p) (\mathbb{E}[Z_n] - p_n) \right)_+^2}{\sum_{n=1}^{N} [\partial_n H_\alpha(p)]^2 \mathcal{D}[f_n]^2 / M(n)} \right). \tag{47}$$

In particular, taking $\theta = H_0(\mathbb{E}[Z]) = H_\alpha(\mathbb{E}[Z] - \alpha)$ and evaluating the exponential in Eq. (47) at $p = \mathbb{E}[Z] - \alpha \in \mathbb{R}^N$ yields that

$$\mathbb{P}\left[H_\alpha\left(\widehat{\mathbb{E}}[Z]\right) \le H_0(\mathbb{E}[Z])\right] \le \exp\left(-\frac{2\left(\sum_{n=1}^N \partial_n H_\alpha(p)\alpha_n\right)_+^2}{\sum_{n=1}^N \{[\partial_n H_\alpha(p)]^2 \mathcal{D}[f_n]^2\}/M(n)}\right). \tag{48}$$

5.4.4 Remark 3

Formula (48) is particularly useful since it links the margin hits α_n, the sample sizes $M(n)$, and the maximum probability of error. For example, given a desired level of confidence, margin hits α_n, and a total number of samples $M \in \mathbb{N}$, one can choose sample sizes $M(1), \ldots, M(N)$ that sum to M and minimize the right-hand side of Eq. (48); this yields an optimal distribution of sampling resources so as to ensure that $H_\alpha\left(\widehat{\mathbb{E}}[Z]\right) \ge H_0(\mathbb{E}[Z])$ with the desired level of confidence. That is, from the point of view of minimizing error probabilities, an optimal assignment of sampling resources is given by the minimizer of the right-hand side of Eq. (48) among all $(M_1, \ldots, M_N) \in \mathbb{N}_0^N$ such that $\sum_{n=1}^N M_n = M$.

6. HIGH-DIMENSIONAL ASYMPTOTICS

The topic of this section is the asymptotic sharpness of the bounds introduced above as the dimension of the space $\mathcal{X}$ becomes large. We begin with a comparison of the McDiarmid and half-space bounds for a simple function: a quadratic form on $\mathbb{R}^N$.

6.1 Example 2: Comparison with McDiarmid's Inequality

The following example serves to illustrate how the half-space method can produce upper bounds on the measure of suitable sublevel sets that are superior to those offered by McDiarmid's inequality; it also shows how this effect is more pronounced in higher-dimensional spaces. Consider the following quadratic form Q_N on $\mathbb{R}^N$:

$$Q_N(x) := \frac{1}{2}\left\|x - \left(\frac{1}{2}, \ldots, \frac{1}{2}\right)\right\|_2^2. \tag{49}$$

For any $\theta > 0$, the sublevel set $Q_N^{-1}([-\infty, \theta])$ is simply a ball of radius $\sqrt{2\theta}$ about the point $(1/2, \ldots, 1/2)$. Suppose that a random vector X takes values in $[-(1/2), +(1/2)]^N$ with independent components. McDiarmid's inequality [Eq. (21a)] implies that

$$\mathbb{P}[Q_N(X) \le \theta] \le \exp\left[-8\left(\frac{\sqrt{N}}{6} - \frac{\theta}{\sqrt{N}}\right)_+^2\right],$$

If also $\mathbb{E}[X] = 0$, then Proposition 2 implies that

$$\mathbb{P}[Q_N(X) \le \theta] \le \exp\left(-\frac{(\sqrt{N} - \sqrt{8\theta})_+^2}{2}\right).$$

For small N and large θ, McDiarmid's bound is the sharper of the two. However, for small θ (and, notably, as $N \to \infty$ for any fixed θ), the half-space bound is the sharper bound (see Fig. 3 for an illustration).

The previous example suggests that bounds constructed using the half-space method may perform very well in high dimension but also that the sharpness of the bound may depend on "how round" the set whose measure we wish to bound is. To fix ideas, suppose that $X = (X_1, \ldots, X_N) \colon \Omega \to \mathbb{R}^N$ is a random vector with independent components, where X_n is supported on an interval of length L_n. For $A \subseteq \mathbb{R}^N$, how sharp is the bound

$$\mathbb{P}[X \in A] \le \exp\left(-\frac{d_\perp(\mathbb{E}[X], A)^2}{2}\right)? \tag{50}$$

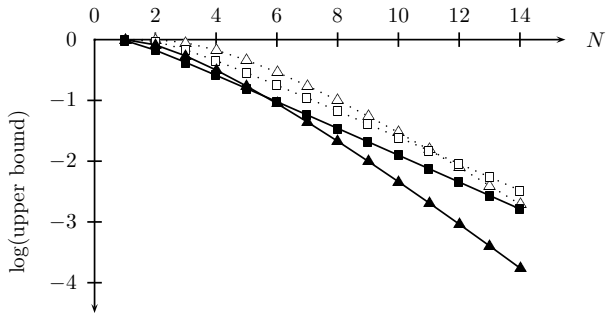

FIG. 3: For the quadratic form Q_N on $\mathbb{R}^N$ given in Eq. (49), a comparison of the McDiarmid upper bound (squares) and half-space upper bound (triangles) on $\mathbb{P}[Q_N(X) \leq \mathrm{m}]$ in the cases $\mathrm{m} = 1/4$ (dotted line and hollow polygons) and $\mathrm{m} = 1/8$ (solid line and filled polygons).

First, note that since $d_\perp(\mathbb{E}[X], A) = d_\perp(\mathbb{E}[X], \overline{\mathrm{co}}(A))$, the bound cannot be expected to be sharp if A differs greatly from its closed convex hull, and so it makes sense to restrict investigation to the case that $A = K$, a closed and convex subset of $\mathbb{R}^N$. Second, it is not reasonable to expect the bound [Eq. (50)] on $\mathbb{P}[X \in K]$ to be sharp if K is sharply pointed; e.g., if K is the narrow wedge K_ε of angle $\varepsilon \ll 1$ based at $e_1 := (1, 0, \ldots, 0)$ in $\mathbb{R}^N$:

$$K_\varepsilon := \left\{ x \in \mathbb{R}^N \,\middle|\, \frac{(x - e_1) \cdot e_1}{\|x - e_1\|_2} \leq \varepsilon \right\}. \tag{51}$$

Therefore, we wish to consider the opposite situation in which K has no sharp points, which will be made precise by requiring that K satisfy an interior ball condition.

Suppose that $(p, \nu) \in \mathrm{N}^* K$ is such that $d_\perp(x, \mathbb{H}_{p,\nu}) = d_\perp(x, K)$. Suppose also that $\mathbb{B}_r(p - r\omega) \subseteq K$, with $r > 0$ and $\omega \in \mathbb{R}^N$ a unit vector, is an interior ball for K at $p \in \partial K$ (cf. Fig. 4). If the law of X on $\mathbb{R}^N$ is highly singular, then it cannot be expected that the bound [Eq. (50)] is sharp, so suppose that the law of X has a density with respect to Lebesgue measure that is bounded above by some constant $C > 0$. Then, the bound [Eq. (50)] is

$$\mathbb{P}[X \in K] \leq \exp\left(- \frac{2 \langle \nu, \mathbb{E}[X] - p \rangle_+^2}{\sum_{n=1}^N \nu_n^2 L_n^2} \right).$$

In the extreme case, K is precisely the closed ball $\overline{\mathbb{B}}_r(p - r\omega)$, the $\mathbb{P}$ measure of which is at most $C r^N \pi^{N/2} / \Gamma(1 + N/2)$.

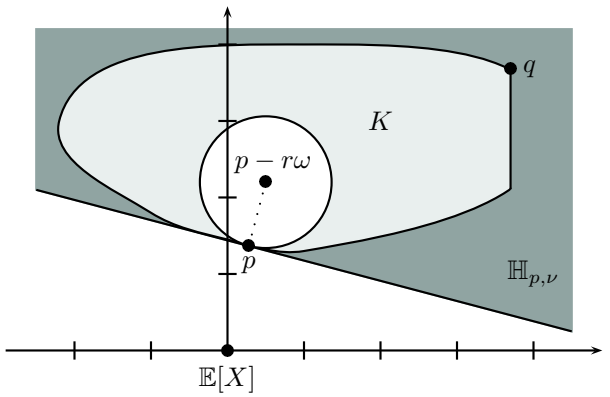

FIG. 4: An interior ball of radius r for the closed convex set K at the frontier point p. Necessarily, p is a point at which ∂K is smooth; K admits no interior ball of positive radius at the vertex q. For convenience, the unit vector $\omega \in \mathbb{R}^N$ has been identified with $\nu \in \mathrm{N}_p^* K \subseteq (\mathbb{R}^N)^*$.

In large deviations theory, the standard notion of asymptotic sharpness is logarithmic equivalence [15, Section I.1]; see, also, [14, 16] for surveys of the large deviations literature. Two sequences $(\alpha_n)_{n\in\mathbb{N}}$ and $(\beta_n)_{n\in\mathbb{N}}$ are said to be logarithmically equivalent, denoted $\alpha_n \simeq \beta_n$, if

$$\frac{1}{n}\log\alpha_n - \frac{1}{n}\log\beta_n \equiv \log\left(\frac{\alpha_n}{\beta_n}\right)^{1/n} \to 0 \text{ as } n\to\infty. \tag{52}$$

Are the half-space bound [Eq. (50)] and the measure of $\overline{\mathbb{B}}_r(p-r\omega)$ logarithmically equivalent? That is, does the conditional probability $\mathbb{P}\left[X\in\overline{\mathbb{B}}_r(p-r\omega)\,\middle|\,X\in\mathbb{H}_{p,\nu}\right]$, when raised to the power $1/N$, converge to 1 as $N\to\infty$? To simplify the asymptotic expansions below, in all lines after the first two, we shall take $\mathbb{E}[X]=0$ and $L_1=\cdots=L_N=1$. Then,

$$\begin{aligned}
&\frac{1}{N}\log\mathbb{P}\left[X\in\overline{\mathbb{B}}_r(p-r\omega)\right] - \frac{1}{N}\log\left(\text{right-hand side of Eq. (50)}\right)\\
&\le \frac{1}{N}\left(\log\frac{Cr^N\pi^{N/2}}{\Gamma(1+N/2)} + \frac{2\langle\nu,\mathbb{E}[X]-p\rangle_+^2}{\sum_{n=1}^N\nu_n^2L_n^2}\right)\\
&= \frac{2\langle\nu,p\rangle_-^2}{N\|\nu\|_2^2} + \frac{\log(Cr^N\pi^{N/2})}{N} - \frac{\log\Gamma(1+N/2)}{N}
\end{aligned}$$

which, by Stirling's approximation for the Gamma function [34, p. 256, Eq. (6.1.37)], is approximately

$$\begin{aligned}
&\approx \frac{2\langle\nu,p\rangle_-^2}{N\|\nu\|_2^2} + \frac{\log(Cr^N\pi^{N/2})}{N} - \frac{1}{N}\log\left(\sqrt{\frac{2\pi}{1+N/2}}\left(\frac{1+N/2}{e}\right)^{1+N/2}\right)\\
&\sim \frac{2\langle\nu,p\rangle_-^2}{N\|\nu\|_2^2} + \frac{\log C}{N} - \frac{1}{2N}\log\frac{4\pi}{N} - \frac{1+N/2}{N}\log\frac{N}{2e}\\
&\sim \frac{2\langle\nu,p\rangle_-^2}{N\|\nu\|_2^2} + \log r - \log\sqrt{N}
\end{aligned}$$

Note that $\langle\nu,p\rangle_-/\|\nu\|_2 \le \sqrt{N}d_{\mathbf{1}}(0,p)$, where $d_{\mathbf{1}}$ denotes the weighted Hamming distance with weight $w=(1,\ldots,1)$. Therefore, a necessary (but not sufficient) condition for the half-space bound to be asymptotically sharp in the sense of logarithmic equivalence is that r is of the same order as $\sqrt{N}$. That is, it is necessary that K is sufficiently round that it has an interior ball of radius comparable to $\sqrt{N}$ at those frontier points where the normal distance $d_\perp(\mathbb{E}[X],K)$ is attained.

Now suppose that $K=f^{-1}([-\infty,\theta])$ is a convex sublevel set for twice-differentiable function f. Let $\eta_1,\ldots,\eta_{N-1},\nu$ be a basis of $\mathbb{R}^N$ such that

$$\|\eta_1\|_2=\cdots=\|\eta_{N-1}\|_2=\|\nu\|_2=1$$

and, for each $n\in\{1,\ldots,N-1\}$, η_n is perpendicular to ν. Suppose that, in this system of normal coordinates, near p, the frontier of K can be approximated by a parabola:

$$\partial K = \left\{ y_1\eta_1+\ldots y_{N-1}\eta_{N-1}-y_N\nu \;\middle|\; y_N=\sum_{n=1}^{N-1}\lambda_n y_n^2\right\}$$

with $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_{N-1}\ge 0$. Then, the condition that K admits an interior ball of radius r at p is the inequality

$$r-\sqrt{r^2-\sum_{n=1}^{N-1}y_n^2} \ge \sum_{n=1}^{N-1}\lambda_n y_n^2 \quad\text{whenever}\quad \sum_{n=1}^{N-1}y_n^2\le r^2.$$

This, in turn, leads to the following condition on λ_1: it must hold that $\lambda_1\le(1/2r)$. Put another way, the half-space method cannot be expected to provide asymptotically sharp bounds for $\mathbb{P}[f(X)\le\theta]$ if, when f is approximated in normal coordinates near the closest point of $f^{-1}([-\infty,\theta])$ to $\mathbb{E}[X]$ by a non-negative quadratic form, that quadratic form has an eigenvalue greater than $(4N)^{-1/2}$.

7. CONCLUSIONS

In this paper we have reviewed some well-established notions of distance and diameter in the concentration-of-measure literature, and paid particular attention to the distance associated with the method of Chernoff bounds. In so doing, we observe that associated with a deviation inequality that depends on a family of (exact) expected values, there is a way to assign sampling resources to the estimation of those expected values that is both natural with respect to the distance (concentration rate) and optimal with respect to error probabilities. We note, however, that this optimality is with respect to the deviation inequality that is being estimated: if the deviation inequality itself is not sharp, then the "optimal" sampling assignment will have a similarly non-sharp character. Hence, we expect that the full power of this approach is contingent upon applying it to optimal concentration-of-measure inequalities, as in [35, 36]

ACKNOWLEDGMENTS

The authors acknowledge portions of this work have been supported by the United States Department of Energy National Nuclear Security Administration under Award Number DE-FC52-08NA28613 through the California Institute of Technology's ASC/PSAAP Center for the Predictive Modeling and Simulation of High Energy Density Dynamic Response of Materials.

REFERENCES

1. Ledoux, M., *The Concentration of Measure Phenomenon*, vol. 89, American Mathematical Society, Providence, RI, 2001.
2. Lugosi, G., Concentration-of-measure inequalities, *Lecture Notes*, Pompeu Fabra University, Barcelona, Spain, 2009
3. McDiarmid, C., Concentration, in *Probabilistic Methods for Algorithmic Discrete Mathematics*, vol. 16, pp. 195–248, Springer, Berlin, 1998.
4. Milman, V. D. and Schechtman, G., *Asymptotic Theory of Finite-Dimensional Normed Spaces*, in Lecture Notes in Mathematics, vol. 1200, Springer, Berlin, 1986.
5. McDiarmid, C., On the method of bounded differences, in *Surveys in Combinatorics*, London Mathematical Society Lecture Note Series, vol. 141, pp. 148–188, Cambridge University Press, Cambridge, 1989.
6. McDiarmid, C., Centering sequences with bounded differences, *Combin. Probab. Comput.*, 6(1):79–86, 1997.
7. Vu, V. H., Concentration of non-Lipschitz functions and applications, *Random Struct. Algorithms*, 20(3):262–316, 2002.
8. Lucas, L. J., Owhadi, H., and Ortiz, M., Rigorous verification, validation, uncertainty quantification and certification through concentration-of-measure inequalities, *Comput. Methods Appl. Mech. Engrg.*, 197(51-52):4591–4609, 2008.
9. Steinwart, I. and Christmann, A., *Support Vector Machines*, *Information Science and Statistics*, Springer, New York, 2008.
10. Dubhashi, D. P., Talagrand's inequality and locality in distributed computing, *Lect. Notes Comput. Sci.*, 1518:60–70, 1998.
11. Lévy, P., *Problèmes Concrets d'Analyse Fonctionnelle. Avec un complément sur les fonctionnelles analytiques par F. Pellegrino*, 2nd ed., Gauthier-Villars, Paris, 1951.
12. Talagrand, M., Concentration of measure and isoperimetric inequalities in product spaces, *Publ. Math., Inst. Hautes Etud. Sci.*, 81:73–205, 1995.
13. Chernoff, H., A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Math. Stat.*, 23:493–507, 1952.
14. Dembo, A. and Zeitouni, O., *Large Deviations Techniques and Applications*, 2nd ed., vol. 38, Springer, New York, 1998.
15. den Hollander, F., *Large Deviations*, in *Fields Institute Monographs*, vol. 14, American Mathematical Society, Providence, RI, 2000.
16. Varadhan, S. R. S., Large deviations, *Ann. Probab.*, 36(2):397–419, 2008.
17. Boyd, S. and Vandenberghe, L., *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
18. Talagrand, M., An isoperimetric theorem on the cube and the Kintchine–Kahane inequalities, *Proc. Am. Math. Soc.*, 104(3):905–909, 1988.

19. Johnson, W. B. and Schechtman, G., Remarks on Talagrand's deviation inequality for Rademacher functions, *Lect. Notes Math.*, 1470:72–77, 1991.
20. Hoeffding, W., Probability inequalities for sums of bounded random variables, *J. Am. Statist. Assoc.*, 58(301):13–30, 1963.
21. Adams, M., Lashgari, A., Li, B., McKerns, M., Mihaly, J. M., Ortiz, M., Owhadi, H., Rosakis, A. J., Stalzer, M., and Sullivan, T. J., Rigorous model-based uncertainty quantification with application to terminal ballistics. Part II: Systems with uncontrollable inputs and large scatter, *J. Mech. Phys. Solids*, in press, 2011.
22. Kidane, A. A., Lashgari, A., Li, B., McKerns, M., Ortiz, M., Owhadi, H., Ravichandran, G., Stalzer, M., and Sullivan, T. J., Rigorous model-based uncertainty quantification with application to terminal ballistics. Part I: Systems with controllable inputs and small scatter, *J. Mech. Phys. Solids*, in press, 2011.
23. Bakry, D. and Émery, M., Diffusions hypercontractives, *Lect. Notes Math.*, 1123:177–206, 1985.
24. Gross, L., Logarithmic Sobolev inequalities, *Am. J. Math.*, 97(4):1061–1083, 1975.
25. Holley, R. and Stroock, D., Logarithmic Sobolev inequalities and stochastic Ising models, *J. Stat. Phys.*, 46(5-6):1159–1194, 1987.
26. Bobkov, S. G. and Ledoux, M., On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures, *J. Funct. Anal.*, 156(2):347–365, 1998.
27. Boucheron, S., Lugosi, G., and Massart, P., Concentration inequalities using the entropy method, *Ann. Probab.*, 31(3):1583–1614, 2003.
28. Ledoux, M., On Talagrand's deviation inequalities for product measures, *ESAIM: Probab. Stat.*, 1:63–87, 1996.
29. Dembo, A., Information inequalities and concentration of measure, *Ann. Probab.*, 25(2):927–939, 1997.
30. Marton, K., Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration, *Ann. Probab.*, 24(2):857–866, 1996.
31. Ledoux, M. and Talagrand, M., Probability in Banach spaces: Isoperimetry and processes, in *Results in Mathematics and Related Areas. 3*, vol. 23, Springer, Berlin, 1991.
32. Rudin, W., Functional analysis, in *International Series in Pure and Applied Mathematics*, 2nd ed., McGraw-Hill, New York,, 1991.
33. Sullivan, T. J., Topcu, U., McKerns, M., and Owhadi, H., Uncertainty quantification via codimension-one partitioning, *Int. J. Numer. Meth. Eng.*, 85(12):1499–1521, 2011.
34. Abramowitz, M. and Stegun, I. A. (Eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover New York, 1992.
35. Bertsimas, D. and Popescu, I., Optimal inequalities in probability theory: A convex optimization approach, *SIAM J. Optim.*, 15(3):780–804, 2005.
36. Owhadi, H., Scovel, C., Sullivan, T. J., McKerns, M., and Ortiz, M., Optimal uncertainty quantification, *SIAM Rev.*, under review, 2011.