Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

An enumerative formula for the spherical cap discrepancy

Holger Heitsch, René Henrion

submitted: August 4, 2020

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: holger.heitsch@wias-berlin.de
rene.henrion@wias-berlin.de

No. 2744
Berlin 2020

2010 *Mathematics Subject Classification.* 11K38, 90C15.

Key words and phrases. Spherical cap discrepancy, uniform distribution on sphere, optimality conditions.

This work is supported by the German Research Foundation (DFG) within the project B04 of CRC/Transregio 154 and by the FMJH Program Gaspard Monge in optimization and operations research including support to this program by EDF.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

An enumerative formula for the spherical cap discrepancy

Holger Heitsch, René Henrion

Abstract

The spherical cap discrepancy is a widely used measure for how uniformly a sample of points on the sphere is distributed. Being hard to compute, this discrepancy measure is typically replaced by some lower or upper estimates when designing optimal sampling schemes for the uniform distribution on the sphere. In this paper, we provide a fully explicit, easy to implement enumerative formula for the spherical cap discrepancy. Not surprisingly, this formula is of combinatorial nature and, thus, its application is limited to spheres of small dimension and moderate sample sizes. Nonetheless, it may serve as a useful calibrating tool for testing the efficiency of sampling schemes and its explicit character might be useful also to establish necessary optimality conditions when minimizing the discrepancy with respect to a sample of given size.

1 Introduction

A discrepancy $\Delta(\mu,\nu)$ quantifies the deviation between two given measures μ and ν. On a local scale, one may compare the two measures with respect to a given set B to obtain the so-called local discrepancy

$$\Delta(B;\mu,\nu) := |\mu(B) - \nu(B)|.$$

In order to arrive at a global deviation measure, one extends the comparison of the two measures to a collection $\mathcal{B}$ of sets and chooses an appropriate L_p norm:

$$\begin{aligned} \Delta_p(\mu,\nu) &:= \left(\int_{\mathcal{B}} \Delta(B;\mu,\nu)^p d\omega(B)\right)^{1/p} \quad (p<\infty), \\ \Delta_\infty(\mu,\nu) &:= \sup_{B\in\mathcal{B}} \Delta(B;\mu,\nu)\,. \end{aligned} \tag{1}$$

For surveys on discrepancies, we refer to, e.g., [2, 4, 10]. Discrepancies play a fundamental role in many mathematical disciplines. For instance, in stochastic programming, the stability of optimal solutions and optimal values with respect to perturbations of the underlying probability measure can be expected only for a problem-adapted choice of a discrepancy [11].

The focus of the present paper will be on the so-called *spherical cap discrepancy*. Our interest in this quantity comes from the algorithmic solution of optimization problems subject to probabilistic constraints. One approach here relies on the so-called *spheric-radial decomposition* of random vectors having elliptically symmetric distribution (e.g., Gaussian). This approach allows for a representation of the decision-dependent probability of some random inequality system as well as of its gradient as integrals with respect to the uniform distribution on a sphere [12]. Hence, for an efficient numerical approximation of these integrals by finite sums, one has to make use of low discrepancy samples for that distribution. It is well known (see, e.g. [1, p. 991]), that the resulting integration error tends to zero (for samples of increasing size) if and only if the spherical cap discrepancy associated with these samples tends to zero. This special discrepancy is obtained from our general setting (1) by defining

DOI 10.20347/WIAS.PREPRINT.2744 Berlin 2020

$p := \infty$, μ as the uniform measure on the sphere, ν as the empirical measure induced by the sample and $\mathcal{B}$ as the collection of all closed half spaces intersected with the sphere (caps). To be more precise, we define the closed halfspace $H(w,t)$ parameterized by (w,t), its empirical and cap measures $\mu^{emp}(w,t)$ and $\mu^{cap}(w,t)$, respectively, and the spherical cap discrepancy Δ associated with the sample $\left\{x^1,\ldots,x^N\right\}$ by

$$\begin{aligned} H(w,t) &:= \{x \in \mathbb{R}^n | \langle w,x\rangle \geq t\} \quad \left(w \in \mathbb{S}^{n-1},\, t \in [-1,1]\right), \\ \mu^{emp}(w,t) &:= N^{-1} \cdot \#\left\{i \in \{1,\ldots,N\} \,|x^i \in H(w,t)\right\}, \\ \mu^{cap}(w,t) &:= \mu\left(\mathbb{S}^{n-1} \cap H(w,t)\right) \ \left(\mu = \text{law of uniform distribution on } \mathbb{S}^{n-1}\right), \\ \Delta(w,t) &:= |\mu^{emp}(w,t) - \mu^{cap}(w,t)|, \\ \Delta &:= \sup_{w\in\mathbb{S}^{n-1},\, t\in[-1,1]} \Delta(w,t). \end{aligned}$$

The following explicit formula for the cap measure - not depending on $w \in \mathbb{S}^{n-1}$ - is well known

$$\mu^{cap}(w,t) = C_n \cdot \begin{cases} \displaystyle\int_0^{\arccos(t)} \sin^{n-2}(\tau)d\tau, & \text{if} \quad 0 \leq t \leq 1, \\ \displaystyle 1 - \int_0^{\arccos(-t)} \sin^{n-2}(\tau)d\tau, & \text{if} \quad -1 \leq t < 0, \end{cases} \tag{2}$$

where

$$C_n := \frac{1}{\int_0^\pi \sin^{n-2}(\tau)d\tau}$$

is some normalizing constant.

To the best of our knowledge, no explicit formula for calculating the spherical cap discrepancy has been known so far. Rather the emphasis in the literature has been laid on suitable estimates with respect to more manageable quantities allowing for asymptotic derivations and constructions of efficient low discrepancy designs (see, e.g., [1, 7]). On the other hand, beyond the asymptotic 'large sample' viewpoint it might be of some interest even for fixed moderate sample sizes to dispose of an easy enumerative formula enabling one to precisely compute the discrepancy and to compare different sampling schemes.

As a rule, L_p discrepancies ($p < \infty$) are easier to compute than L_∞ discrepancies as a consequence of the collection $\mathcal{B}$ of test sets typically having infinite cardinality [3]. As far as explicit formulae for L_∞ discrepancies are available (e.g., for rectangular or general polyhedral sets, see [3, 8, 9]), they are of combinatorial nature which limits their application with respect to the dimension and the size of the sample. More precisely, it has been shown in [6], that computing the star discrepancy is an NP-hard problem. Moreover, the result is improved by [5] who proved that it is indeed W[1]-hard. Therefore, it is not surprising that a similar combinatorial aspect shows up in the enumerative formula for the spheric cap discrepancy we present in Theorem 1 below. In our numerical experiments, we apply the formula to spheres of dimension starting from 2 (2000 samples) up to 5 (100 samples). Even in this rather modest setting, the formula may prove useful for calibration purposes with respect to some given sampling scheme. For instance, in [1, p. 1005] an easy to compute lower bound for the spherical cap discrepancy is used in numerical experiments in order to confirm empirically a certain asymptotic order for a digital net based on a two-dimensional Sobol' point set on $\mathbb{S}^2$. Strictly speaking, the order obtained with respect to the lower bound transfers to the discrepancy only when the ratio between the true value and the lower estimate is approximately constant for increasing sample size. This is what we may confirm indeed in our numerical experiments. We also use the proven formula in order to directly compare discrepancies of a few sampling schemes on $\mathbb{S}^2$ for sample sizes of up to 1000. The results

verify the good quality of a sampling scheme via Lambert's equal-area transform proposed in [1, p. 995]. Finally, we mention that the explicit character of the obtained formula might be of some interest for the derivation of necessary optimality conditions when minimizing the discrepancy as a function of a sample of fixed size.

2 Preparatory Results

We have the following elementary (semi-) continuity properties of both considered measures:

Lemma 1. *μ^{cap} is continuous and μ^{emp} is upper semicontinuous on $\mathbb{S}^{n-1} \times [-1,1]$. Moreover, the following relations are satisfied for all $w \in \mathbb{S}^{n-1}$ and $t \in [-1,1]$:*

$$\mu^{emp}(w,t) + \mu^{emp}(-w,-t) \geq 1, \tag{3}$$
$$\mu^{cap}(w,t) + \mu^{cap}(-w,-t) = 1. \tag{4}$$

Proof. The continuity of μ^{cap} and (4) follow immediately from **(2)**. Let $w \in \mathbb{S}^{n-1}$, $t \in [-1,1]$ and $(w_k,t_k) \to (w,t)$ an arbitrary sequence with $w_k \in \mathbb{S}^{n-1}$, $t_k \in [-1,1]$. Define

$$I := \left\{ i \in \{1,\dots,N\} \,|x^i \notin H(w,t) \right\},$$

so that $\langle w, x^i \rangle < t$ for all $i \in I$. Then, by continuity, there is some k_0, such that $\langle w_k, x^i \rangle < t_k$ - i.e., $x^i \notin H(w_k,t_k)$ - for all $k \geq k_0$ and all $i \in I$. It follows that

$$\mu^{emp}(w_k,t_k) \leq \mu^{emp}(w,t) \quad \forall k \geq k_0,$$

whence

$$\limsup_{k\to\infty} \mu^{emp}(w_k,t_k) \leq \mu^{emp}(w,t).$$

This proves the upper semicontinuity of μ^{emp} on $\mathbb{S}^{n-1} \times [-1,1]$. Concerning (3), we employ the relation

$$H(w,t) \cup H(-w,-t) = \mathbb{R}^n,$$

which is evident from the definition of $H(w,t)$ and which entails that

$$\#\left\{ i \in \{1,\dots,N\} \,|x^i \in H(w,t) \right\} + \#\left\{ i \in \{1,\dots,N\} \,|x^i \in H(-w,-t) \right\} \geq N.$$

Thus, (3) follows. □

The next proposition shows that the discrepancy Δ is always realized by a certain closed half space:

Proposition 1. *There are $w^* \in \mathbb{S}^{n-1}$, $t^* \in [-1,1]$, such that*

$$\Delta = \left| \mu^{emp}(w^*,t^*) - \mu^{cap}(w^*,t^*) \right|.$$

Proof. Let $(w_k,t_k) \in \mathbb{S}^{n-1} \times [-1,1]$ be a sequence realizing the supremum in the definition of Δ:

$$\left| \mu^{emp}(w_k,t_k) - \mu^{cap}(w_k,t_k) \right| \to_k \Delta \tag{5}$$

By the compactness of $\mathbb{S}^{n-1} \times [-1,1]$ we may assume that

$$(w_k,t_k) \to (\bar{w},\bar{t}) \in \mathbb{S}^{n-1} \times [-1,1]. \tag{6}$$

According to (5) one may assume one of the following two cases upon passing to a subsequence:

$$\mu^{emp}(w_k,t_k) - \mu^{cap}(w_k,t_k) \;\to\; \Delta, \tag{7}$$
$$\mu^{cap}(w_k,t_k) - \mu^{emp}(w_k,t_k) \;\to\; \Delta. \tag{8}$$

In the case of (7), the continuity of μ^{cap} and the upper semicontinuity of μ^{emp} on $\mathbb{S}^{n-1}\times[-1,1]$ (see Lemma 1) yield along with (6) that:

$$\begin{aligned}\Delta &= \lim_{k\to\infty}\left(\mu^{emp}(w_k,t_k) - \mu^{cap}(w_k,t_k)\right)\\ &= \limsup_{k\to\infty}\left(\mu^{emp}(w_k,t_k) - \mu^{cap}(w_k,t_k)\right)\\ &\le \mu^{emp}(\bar w,\bar t) - \mu^{cap}(\bar w,\bar t) \le \left|\mu^{emp}(\bar w,\bar t) - \mu^{cap}(\bar w,\bar t)\right| \le \Delta.\end{aligned}$$

Hence, $\Delta = \left|\mu^{emp}(\bar w,\bar t) - \mu^{cap}(\bar w,\bar t)\right|$. In the case of (8) one may exploit (3), (4) and once more the upper semicontinuity of μ^{emp} in order to derive that:

$$\begin{aligned}\Delta &= \lim_{k\to\infty}\left(\mu^{cap}(w_k,t_k) - \mu^{emp}(w_k,t_k)\right)\\ &= \lim_{k\to\infty}\left(1-\mu^{emp}(w_k,t_k) - (1-\mu^{cap}(w_k,t_k)\right)\\ &= \limsup_{k\to\infty}\left(1-\mu^{emp}(w_k,t_k) - (1-\mu^{cap}(w_k,t_k)\right)\\ &\le \limsup_{k\to\infty}\left(\mu^{emp}(-w_k,-t_k) - (1-\mu^{cap}(w_k,t_k)\right)\\ &\le \mu^{emp}(-\bar w,-\bar t) - (1-\mu^{cap}(\bar w,\bar t)) = \mu^{emp}(-\bar w,-\bar t) - \mu^{cap}(-\bar w,-\bar t)\\ &\le \left|\mu^{emp}(-\bar w,-\bar t) - \mu^{cap}(-\bar w,-\bar t)\right| \le \Delta.\end{aligned}$$

Hence, $\Delta = \left|\mu^{emp}(-\bar w,-\bar t) - \mu^{cap}(-\bar w,-\bar t)\right|$. Altogether, the assertion follows with $(w^*,t^*) := (\bar w,\bar t)$ in the first case and $(w^*,t^*) := (-\bar w,-\bar t)$ in the second one. □

Proposition 2. *For (w^*,t^*) realizing the discrepancy in Proposition 1 it holds that there is some $i\in\{1,\dots,N\}$ such that $\langle w^*,x^i\rangle = t^*$.*

Proof. Assume that $\langle w^*,x^j\rangle \neq t^*$ for all $j\in\{1,\dots,N\}$. Then,

$$\mu^{emp}(w^*,t) = \mu^{emp}(w^*,t^*) \tag{9}$$

for t close to t^*. If $|t^*|<1$, then one may strictly increase ($t>t^*$) or decrease ($t<t^*$) $\mu^{cap}(w^*,t)$, so that by virtue of (9) the local discrepancy $\Delta(w^*,t)$ can be strictly increased in comparison with the maximal one $\Delta(w^*,t^*)=\Delta$. This is a contradiction. If $t^*=1$, then

$$\begin{aligned}-1 < \langle w^*,x^j\rangle &< 1 \quad \forall j\in\{1,\dots,N\},\\ \mu^{cap}(w^*,t^*) &= \mu^{emp}(w^*,t^*) = 0.\end{aligned}$$

Since $\mu^{cap}(w^*,t)$ is strictly increased for $t<t^*=1$ while $\mu^{emp}(w^*,t)=0$ for t close to t^* (see (9)), one may strictly increase the local discrepancy again, so that the same contradiction results. The case $t^*=-1$ follows analogously. □

Lemma 2. *Let $\{x^1,\dots,x^k\}\subseteq\mathbb{S}^{n-1}$ for some $k\in\mathbb{N}$ and let*

$$S := \left\{(w,t)\;\middle|\;\langle w,x^i\rangle = t\;\;(i=1,\dots,k),\;\langle w,w\rangle = 1\right\}\subseteq\mathbb{S}^{n-1}\times[-1,1]. \tag{10}$$

Let

$$p := \operatorname{rank} \left\{ \begin{pmatrix} x^i \\ -1 \end{pmatrix} \right\}_{i=1,\dots,k}.$$

Then, assuming without loss of generality that

$$\operatorname{rank} \left\{ \begin{pmatrix} x^i \\ -1 \end{pmatrix} \right\}_{i=1,\dots,p} = p,$$

the set S defined in (10) has a reduced representation

$$S = \left\{ (w,t) \;\middle|\; \left\langle w, x^i \right\rangle = t \;\; (i = 1,\dots,p) \,,\; \langle w, w \rangle = 1 \right\}. \tag{11}$$

Proof. By $p \leq k$ it is sufficient to show that the right-hand side of (11) is contained in S as defined in (10). It is therefore enough to show the implication

$$\left\langle w, x^j \right\rangle = t \;\; (j = 1,\dots,p) \Longrightarrow \left\langle w, x^i \right\rangle = t \;\; (i = p+1,\dots,k). \tag{12}$$

By definition of p, the vectors $\begin{pmatrix} x^i \\ -1 \end{pmatrix}$ $(i = p+1,\dots,k)$ are linear combinations of the vectors $\begin{pmatrix} x^j \\ -1 \end{pmatrix}$ $(j = 1,\dots,p)$. Hence, for an arbitrarily fixed $i \in \{p+1,\dots,k\}$ there exists some $\lambda \in \mathbb{R}^p$ such that

$$\begin{pmatrix} x^i \\ -1 \end{pmatrix} = \sum_{j=1}^{p} \lambda_j \begin{pmatrix} x^j \\ -1 \end{pmatrix}.$$

Along with the assumption in (12), both components of this last identity yield that

$$\left\langle w, x^i \right\rangle = \sum_{j=1}^{p} \lambda_j \left\langle w, x^j \right\rangle = t \sum_{j=1}^{p} \lambda_j = t$$

which is the conclusion of (12). □

Lemma 3. *Let $\left\{x^1,\dots,x^k\right\} \subseteq \mathbb{S}^{n-1}$ be such that*

$$\operatorname{rank} \left\{ \begin{pmatrix} x^i \\ -1 \end{pmatrix} \right\}_{i=1,\dots,k} = k.$$

Denote by X_* the matrix whose columns are generated by x^i for $i = 1,\dots,k$ and define

$$\tilde{X}_* := \begin{pmatrix} X_* \\ -\mathbf{1}^T \end{pmatrix}; \quad \gamma := \mathbf{1}^T \left(\tilde{X}_*^T \tilde{X}_* \right)^{-1} \mathbf{1}; \quad \mathbf{1} := (1,\dots,1)^T.$$

Let (w^*, t^*) be a local solution of one of the two optimization problems

$$\max_{w,t} / \min_{w,t} \left\{ t \;\middle|\; \left\langle w, x^i \right\rangle = t \;\; (i = 1,\dots,k) \,,\; \langle w, w \rangle = 1 \right\}. \tag{13}$$

Then, it holds that $0 < \gamma \leq 1$. If $\gamma < 1$, then

$$(w^*, t^*) \in \left\{ \left(\hat{w}, \hat{t}\right), \left(-\hat{w}, -\hat{t}\right) \right\},$$

where

$$\hat{t} := \left(\frac{1-\gamma}{\gamma} \right)^{1/2}, \quad \hat{w} := \frac{1+\hat{t}^2}{\hat{t}} X_* \left(\tilde{X}_*^T \tilde{X}_* \right)^{-1} \mathbf{1}.$$

Moreover, $\gamma = 1$ is equivalent to $t^* = 0$ and we then have $\operatorname{rank} X_* = k-1$.

Proof. In order to identify (w^*, t^*) via necessary optimality conditions we have first to check if the gradients

$$\left\{ \begin{pmatrix} x^1 \\ -1 \end{pmatrix}, \ldots, \begin{pmatrix} x^k \\ -1 \end{pmatrix}, \begin{pmatrix} 2w \\ 0 \end{pmatrix} \right\}$$

with respect to (w, t) of the equality constraints in (13) are linearly independent. We assume a linear combination

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \sum_{i=1}^{k} \lambda_i \begin{pmatrix} x^i \\ -1 \end{pmatrix} + \mu \begin{pmatrix} 2w \\ 0 \end{pmatrix}.$$

Multiplication of the first component with w yields - taking into account the equality constraints in (13) and comparing the second component - that

$$0 = \sum_{i=1}^{k} \lambda_i \left\langle w, x^i \right\rangle + 2\mu \left\langle w, w \right\rangle = t \sum_{i=1}^{k} \lambda_i + 2\mu = 2\mu.$$

Hence, the first component may be rewritten as

$$\sum_{i=1}^{k} \lambda_i \begin{pmatrix} x^i \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

By assumption of the Lemma, the vectors $\left\{ \binom{x^i}{-1} \right\}_{i=1,\ldots,k}$ are linearly independent, whence $\lambda_i = 0$ for $i = 1, \ldots, k$. Furthermore, $\mu = 0$, which altogether proves the linear independence of the gradients of equality constraints in (13).

This allows us to derive the following simultaneous necessary optimality conditions for a local solution (w^*, t^*) of any of the two problems (13). Here the gradient of the objective function t appears on the left-hand side:

$$\exists\, \lambda_1, \ldots, \lambda_k, \mu \in \mathbb{R} : \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sum_{i=1}^{k} \lambda_i \begin{pmatrix} x^i \\ -1 \end{pmatrix} + \mu \begin{pmatrix} 2w^* \\ 0 \end{pmatrix}. \tag{14}$$

The second component implies that $\sum_{i=1}^{k} \lambda_i = -1$. Multiplication of the first component by w^* and exploiting the equality constraints in (13) yields that

$$0 = \sum_{i=1}^{k} \lambda_i \left\langle w^*, x^i \right\rangle + 2\mu \left\langle w^*, w^* \right\rangle = t^* \sum_{i=1}^{k} \lambda_i + 2\mu,$$

in particular, $t^* = 2\mu$. Putting $\lambda := (\, \lambda_1, \ldots, \lambda_k)^T$ the equation in (14) reads as

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \tilde{X}_* \lambda + t^* \begin{pmatrix} w^* \\ 0 \end{pmatrix}. \tag{15}$$

Multiplication of both sides from the left by $\tilde{X}_*^T$ and using the first feasibility constraint $X_*^T w^* = t^* \mathbf{1}$ in (13) results in

$$-\mathbf{1} = \tilde{X}_*^T \tilde{X}_* \lambda + t^{*2} \mathbf{1}.$$

By the assumption of this Lemma, the matrix $\tilde{X}_*^T \tilde{X}_*$ is regular and we can solve the last equation for λ:

$$\lambda = -(1 + t^{*2}) \left(\tilde{X}_*^T \tilde{X}_* \right)^{-1} \mathbf{1}. \tag{16}$$

Recalling that $\mathbf{1}^T\lambda = -1$ we arrive at

$$(1+t^{*2})\mathbf{1}^T\left(\tilde{X}_*^T\tilde{X}_*\right)^{-1}\mathbf{1} = 1\,. \tag{17}$$

By definition of γ, the latter equation implies that we necessarily have $0 < \gamma \leq 1$, and $\gamma = 1$ if and only if $t^* = 0$. In the case $\gamma = 1$ we have that $w^* \in \operatorname{Ker} X_*^T \cap \mathbb{S}^{n-1}$ by feasibility of w^* in (13). Moreover, from (15) we see that $X_*\lambda = 0$ for some $\lambda \neq 0$ which then implies that $\operatorname{rank} X_* = k-1$ due to

$$k-1 = \operatorname{rank}\tilde{X}_* - 1 \leq \operatorname{rank} X_* = k - \dim\operatorname{Ker} X_* \leq k-1.$$

If, in contrast, $0 < \gamma < 1$, then - with $\hat{t}$ defined in the Lemma -

$$t^* = \pm\left(\frac{1-\gamma}{\gamma}\right)^{1/2} = \pm\hat{t}. \tag{18}$$

The first component of (15) reads

$$0 = X_*\lambda + t^*w^*.$$

Using the representation (16) for λ we obtain that

$$w^* = \frac{1+t^{*2}}{t^*}X_*\left(\tilde{X}_*^T\tilde{X}_*\right)^{-1}\mathbf{1}\,.$$

Hence, with $\hat{w}$ defined in the Lemma, we have $(w^*,t^*) = \left(\hat{w},\hat{t}\right)$ in case that $t^* = \hat{t}$ or $(w^*,t^*) = \left(-\hat{w},-\hat{t}\right)$ in case that $t^* = -\hat{t}$. This completes the proof. □

Lemma 4. *Let $\left\{x^1,\dots,x^N\right\} \subseteq \mathbb{S}^{n-1}$. For any $I \subseteq \{1,\dots,N\}$ let X_I be the matrix whose columns are x^i, $i \in I$. Define $\tilde{X}_I := \left(\begin{smallmatrix} X_I \\ -\mathbf{1}^T \end{smallmatrix}\right)$ and let be $\tilde{X} := \tilde{X}_{\{1,\dots,N\}}$. Let $w_0 \in \mathbb{S}^{n-1}$ be given such that*

$$I_0 := \left\{i \in \{1,\dots,N\} \,|\, \langle w_0, x^i\rangle = 0\right\} \neq \emptyset$$

and such that it holds

$$w_0 \in \operatorname*{arg\,max}_{w\in\operatorname{Ker} X_{I_0}^T\cap\mathbb{S}^{n-1}} \mu^{emp}(w,0)\,, \tag{19}$$

$$\operatorname{rank}\tilde{X}_{I_0} < \min\left\{n, \operatorname{rank}\tilde{X}\right\}, \quad \operatorname{rank} X_{I_0} = \operatorname{rank}\tilde{X}_{I_0} - 1\,. \tag{20}$$

Then there exists $w_1 \in \mathbb{S}^{n-1}$ such that for $I_1 := \left\{i \in \{1,\dots,N\} \,|\, \langle w_1, x^i\rangle = 0\right\}$ it holds $I_0 \subseteq I_1$, and, moreover, it holds

$$w_1 \in \operatorname{Ker} X_{I_1}^T \cap \mathbb{S}^{n-1}, \quad \mu^{emp}(w_1,0) = \mu^{emp}(w_0,0), \tag{21}$$

$$\operatorname{rank} X_{I_1} = \operatorname{rank} X_{I_0} + z, \quad \operatorname{rank}\tilde{X}_{I_1} = \operatorname{rank}\tilde{X}_{I_0} + z \tag{22}$$

for some natural number $z \geq 1$.

Proof. We claim that assumptions (19) and (20) imply that the index set

$$J_0 := \left\{j \in \{1,\dots,N\} \,|\, \langle w_0, x^j\rangle > 0\right\}$$

is nonempty. Indeed, in case that $J_0 = \emptyset$, we would have that $\mu^{emp}(w_0,0) = N^{-1}\#I_0$ and that $\langle w_0, x^i\rangle \leq 0$ for all $i \in \{1,\dots,N\}$, which amounts to $\mu^{emp}(-w_0,0) = 1$. Then, since $-w_0 \in \operatorname{Ker} X_{I_0}^T \cap \mathbb{S}^{n-1}$, it would follow that

$$1 = \mu^{emp}(-w_0,0) \leq \mu^{emp}(w_0,0) = N^{-1}\#I_0 \leq 1.$$

Consequently, $N = \#I_0$, hence $\tilde{X} = \tilde{X}_{I_0}$ and we arrive at the contradiction

$$\operatorname{rank} \tilde{X} = \operatorname{rank} \tilde{X}_{I_0} < \min\left\{n, \operatorname{rank} \tilde{X}\right\} \leq \operatorname{rank} \tilde{X}.$$

Therefore, $J_0 \neq \emptyset$.

From the assumption $w_0 \in \operatorname{Ker} X_{I_0}^T \cap \mathbb{S}^{n-1}$ and from the definitions of I_0, J_0 we observe that

$$\mu^{emp}(w_0, 0) = N^{-1}\,(\#I_0 + \#J_0). \tag{23}$$

In order to show the existence of some suitable w_1 let us consider the following optimization problem:

$$\min_w \left\{\varphi(w) \;\middle|\; w \in \operatorname{Ker} X_{I_0}^T \cap \mathbb{S}^{n-1},\, \varphi(w) \geq 0\right\}, \quad \varphi(w) := \min_{j \in J_0} \langle w, x^j \rangle. \tag{24}$$

Observe that the feasible set of this problem is nonempty (it contains w_0) and compact by continuity of φ. Hence, once more by continuity of φ, the problem admits a solution w_1.

Next, we prove that $\varphi(w_1) = 0$. Assume to the contrary that $\varphi(w_1) > 0$. Select $j_1 \in J_0$ satisfying $\langle w_1, x^{j_1} \rangle = \varphi(w_1)$ and put $K := I_0 \cup \{j_1\}$. Because $x^{j_1} \notin \operatorname{span}\left\{x^i\right\}_{i \in I_0}$ by definition of the index sets I_0 and J_0, we observe that

$$\operatorname{rank} X_K = \operatorname{rank} X_{I_0} + 1. \tag{25}$$

Assumption (20) and property (25) imply that

$$\dim \operatorname{Ker} X_K^T = n - \operatorname{rank} X_K = n - \operatorname{rank} X_{I_0} - 1 > 0,$$

whence $\operatorname{Ker} X_K^T \cap \mathbb{S}^{n-1} \neq \emptyset$. Select some $\bar{w} \in \operatorname{Ker} X_K^T \cap \mathbb{S}^{n-1}$ moreover satisfying $\langle w_1, \bar{w} \rangle \geq 0$ and define

$$\bar{w}_t := t\bar{w} + (1-t)\, w_1 \quad \forall t \in [0,1]\,.$$

Then, with $\|\cdot\|$ referring to the Euclidean norm, we derive that

$$\|\bar{w}_t\| > 1 - t > 0 \quad \forall t \in (0,1). \tag{26}$$

In particular, recalling that $w_1 \in \operatorname{Ker} X_{I_0}^T \cap \mathbb{S}^{n-1}$ is a solution of (24) and that

$$\bar{w} \in \operatorname{Ker} X_K^T \cap \mathbb{S}^{n-1} \subseteq \operatorname{Ker} X_{I_0}^T \cap \mathbb{S}^{n-1},$$

we may define

$$\tilde{w}_t := \bar{w}_t / \|\bar{w}_t\| \in \operatorname{Ker} X_{I_0}^T \cap \mathbb{S}^{n-1} \quad \forall t \in (0,1).$$

Now, since $\lim_{t \downarrow 0} \|\bar{w}_t\| = 1$, we infer that for all $j \in J_0$

$$\lim_{t \downarrow 0} \left\langle \tilde{w}_t, x^j \right\rangle = \lim_{t \downarrow 0} \left(t \left\langle \bar{w}, x^j \right\rangle + (1-t) \left\langle w_1, x^j \right\rangle\right) / \|\bar{w}_t\| = \left\langle w_1, x^j \right\rangle \geq \varphi(w_1) > 0.$$

Consequently, $\varphi(\tilde{w}_t) \geq 0$ for small enough $t > 0$ which entails that $\tilde{w}_t$ is feasible in problem (24) for small enough $t > 0$. On the other hand, since $\bar{w} \in \operatorname{Ker} X_K^T$, we may exploit the relation $\langle \bar{w}, x^{j_1} \rangle = 0$, in order to derive from (26) that

$$\begin{aligned} \varphi(\tilde{w}_t) &\leq \left\langle \tilde{w}_t, x^{j_1} \right\rangle = \left(t \left\langle \bar{w}, x^{j_1} \right\rangle + (1-t) \left\langle w_1, x^{j_1} \right\rangle\right) / \|\bar{w}_t\| \\ &= (1-t) \left\langle w_1, x^{j_1} \right\rangle / \|\bar{w}_t\| < \left\langle w_1, x^{j_1} \right\rangle = \varphi(w_1) \end{aligned}$$

for all $t \in (0,1)$, whence the contradiction that for small enough $t > 0$ $\tilde{w}_t$ is feasible in problem (24) and realizes a strictly smaller objective value than the solution w_1. Hence, we have shown that $\varphi(w_1) = 0$.

Therefore, we have that $\langle w_1, x^{j_1}\rangle = \varphi(w_1) = 0$ (due to the choice of j_1), and hence, $w_1 \in \operatorname{Ker} X_K^T \cap \mathbb{S}^{n-1}$. Put $I_1 := \{i \in \{1,\dots,N\} \mid \langle w_1, x^i\rangle = 0\}$ and obtain that $I_0 \subset K \subseteq I_1$. Moreover, the relation

$$\langle w_1, x^j\rangle \geq \varphi(w_1) = 0 \quad \forall j \in J_0$$

implies together with equation (23) and assumption (19) that

$$\mu^{emp}(w_1, 0) \geq N^{-1}\,(\#I_0 + \#J_0) = \mu^{emp}(w_0, 0) \geq \mu^{emp}(w_1, 0).$$

Hence, $\mu^{emp}(w_1, 0) = \mu^{emp}(w_0, 0)$. This, along with the definition of I_1 shows the two relations claimed in (21).

In order to verify (22), let finally $I_1 \setminus I_0 = \{k_1, \dots, k_s\}$ and put $K_0 := I_0$, $K_\ell := I_0 \cup \{k_1, \dots, k_\ell\}$ for $\ell = 1, \dots, s$. Obviously, we have that $\begin{pmatrix} x^\ell \\ -1 \end{pmatrix} \in \operatorname{range} \tilde{X}_{K_{\ell-1}}$ implies that $x^\ell \in \operatorname{range} X_{K_{\ell-1}}$ for any $\ell = 1, \dots, s$. Thus, we obtain

$$\operatorname{rank} \tilde{X}_{K_\ell} - \operatorname{rank} \tilde{X}_{K_{\ell-1}} \geq \operatorname{rank} X_{K_\ell} - \operatorname{rank} X_{K_{\ell-1}} \tag{27}$$

for all $\ell = 1, \dots, s$. Applying (27) recursively for K_ℓ ($\ell = 1, \dots, s$) shows the following estimation (note that $I_1 = K_s$):

$$\begin{aligned} \operatorname{rank} \tilde{X}_{I_1} - \operatorname{rank} \tilde{X}_{I_0} &= \sum_{\ell=1}^{s} \left(\operatorname{rank} \tilde{X}_{K_\ell} - \operatorname{rank} \tilde{X}_{K_{\ell-1}}\right) \\ &\geq \sum_{\ell=1}^{s} \left(\operatorname{rank} X_{K_\ell} - \operatorname{rank} X_{K_{\ell-1}}\right) \\ &= \operatorname{rank} X_{I_1} - \operatorname{rank} X_{I_0}\,. \end{aligned} \tag{28}$$

Define $z := \operatorname{rank} X_{I_1} - \operatorname{rank} X_{I_0}$ and note that all summands above are non-negative. Moreover, because $j_1 \in I_1 \setminus I_0$ we may assume w.l.o.g. $k_1 = j_1$. Hence, the first summand of the second sum (applying $K_1 = K$, $K_0 = I_0$) reads $\operatorname{rank} X_K - \operatorname{rank} X_{I_0}$ which equals 1, due to (25). Therefore, we obtain $z \geq 1$. On the other hand, by the definition of z and by assumption (20) it holds that

$$\begin{aligned} \operatorname{rank} \tilde{X}_{I_1} - \operatorname{rank} \tilde{X}_{I_0} &\leq \operatorname{rank} X_{I_1} + 1 - \operatorname{rank} \tilde{X}_{I_0} \\ &= z + \operatorname{rank} X_{I_0} + 1 - (\operatorname{rank} X_{I_0} + 1) = z\,. \end{aligned} \tag{29}$$

Estimations (28) and (29) show the relations claimed in (22) and we are done. □

We finish this section by a simple implication which will be needed several times in the proof of the main result below and which uses the notation introduced in Lemma 4:

$$\begin{aligned} \operatorname{Ker} X_I^T \cap \mathbb{S}^{n-1} \neq \emptyset &\implies \dim \operatorname{Ker} X_I^T \geq 1 \implies \operatorname{rank} \tilde{X}_I \leq \operatorname{rank} X_I + 1 \leq n \\ &\implies \operatorname{rank} \tilde{X}_I \leq \min\{n, \operatorname{rank} \tilde{X}\}. \end{aligned} \tag{30}$$

3 Main Result

After the preparations of the previous section, we are in a position to derive a formula allowing for the computation of the cap discrepancy of any sample on the sphere by means of explicit finite enumeration. The Theorem divides into a simpler part for the case that the half space realizing the discrepancy does not contain the origin on its boundary (i.e., $t^* \neq 0$ for the couple (w^*, t^*) in Proposition 1) and a technically more delicate part in case that the origin does belong to that boundary (i.e., $t^* = 0$).

Theorem 1. *Let $\left\{x^1, \ldots, x^N\right\} \subseteq \mathbb{S}^{n-1}$. For any $I \subseteq \{1, \ldots, N\}$ with $I \neq \emptyset$, let X_I be the matrix whose columns are x^i ($i \in I$) and define $\tilde{X}_I := \left(\begin{smallmatrix} X_I \\ -\mathbf{1}^T \end{smallmatrix}\right)$ as well as $\tilde{X} := \tilde{X}_{\{1,\ldots,N\}}$. Consider the following finite families of index sets:*

$$\begin{aligned} \Phi_1 &:= \left\{ I \subseteq \{1, \ldots, N\} \;\middle|\; 1 \leq \operatorname{rank} \tilde{X}_I = \#I \leq \min\left\{n, \operatorname{rank} \tilde{X}\right\};\ \gamma_I < 1 \right\}, \\ \Phi_0 &:= \left\{ I \subseteq \{1, \ldots, N\} \;\middle|\; 1 \leq \operatorname{rank} \tilde{X}_I = \#I = \min\left\{n, \operatorname{rank} \tilde{X}\right\};\ \gamma_I = 1 \right\}, \end{aligned}$$

where $\gamma_I := \mathbf{1}^T \left(\tilde{X}_I^T \tilde{X}_I\right)^{-1} \mathbf{1}$. For $I \in \Phi_1 \cup \Phi_0$ put

$$t_I := \begin{cases} \left(\frac{1-\gamma_I}{\gamma_I}\right)^{1/2} & I \in \Phi_1 \\ 0 & I \in \Phi_0 \end{cases}, \quad w_I := \begin{cases} \frac{1+{t_I}^2}{t_I} X_I \left(\tilde{X}_I^T \tilde{X}_I\right)^{-1} \mathbf{1} & I \in \Phi_1 \\ \in \operatorname{Ker} X_I^T \cap \mathbb{S}^{n-1} & I \in \Phi_0 \end{cases},$$

where the selection of w_I in case of $I \in \Phi_0$ is arbitrary. Then, for the cap discrepancy it holds that $\Delta = \max\{\Delta_1, \Delta_0\}$, where

$$\Delta_1 := \begin{cases} \max\limits_{I \in \Phi_1} \max\left\{\Delta(w_I, t_I), \Delta(-w_I, -t_I)\right\} & \text{if } \Phi_1 \neq \emptyset \\ 0 & \text{else} \end{cases},$$

$$\Delta_0 := \begin{cases} \max\limits_{I \in \Phi_0} \max\left\{\Delta(w_I, 0), \Delta(-w_I, 0)\right\} & \text{if } \Phi_0 \neq \emptyset \\ 0 & \text{else} \end{cases}.$$

Proof. Let $(w^*, t^*) \in \mathbb{S}^{n-1} \times [-1, 1]$ be such that (see Prop. 1)

$$\Delta = \Delta(w^*, t^*) = \left|\mu^{emp}(w^*, t^*) - \mu^{cap}(w^*, t^*)\right|. \tag{31}$$

We define the (disjoint) index sets

$$I^* := \{i \in \{1, \ldots, N\} \mid \left\langle w^*, x^i \right\rangle = t^*\}, \ J^* := \{i \in \{1, \ldots, N\} \mid \left\langle w^*, x^i \right\rangle > t^*\}.$$

From Proposition 2, we infer that $I^* \neq \emptyset$. Let

$$S := \left\{ (w, t) \in \mathbb{S}^{n-1} \times [-1, 1] \;\middle|\; \begin{array}{lll} \langle w, x^i \rangle = t & & i \in I^* \\ \langle w, x^i \rangle > t & & i \in J^* \\ \langle w, x^i \rangle < t & i \in \{1, \ldots, N\} \setminus (I^* \cup J^*) & \end{array} \right\}.$$

The choice of (w^*, t^*) implies that

$$(w^*, t^*) \in \operatorname*{arg\,max}_{(w,t) \in \mathbb{S}^{n-1} \times [-1,1]} |\mu^{emp}(w, t) - \mu^{cap}(w, t)|. \tag{32}$$

Since $(w^*, t^*) \in S$ it follows that even

$$(w^*, t^*) \in \operatorname*{arg\,max}_{(w,t)\in S} |\mu^{emp}(w,t) - \mu^{cap}(w,t)|.$$

We observe that $\mu^{emp}(w,t) = \#I^* + \#J^* =: e^*$ for all $(w,t) \in S$. Therefore, depending on the sign of $e^* - \mu^{cap}(w^*, t^*)$, one has that

$$(w^*, t^*) \in \operatorname*{arg\,min}_{(w,t)\in S} \mu^{cap}(w,t) \quad \text{or} \quad (w^*, t^*) \in \operatorname*{arg\,max}_{(w,t)\in S} \mu^{cap}(w,t)\,,$$

respectively. Because, $\mu^{cap}(w,t)$ depends on t only and is monotonically decreasing with t (see(2)), (w^*, t^*) is a solution of one of the two optimization problems

$$\max_{w,t} / \min_{w,t} \left\{ t \;\middle|\; \begin{array}{ll} \langle w, x^i\rangle = t & i \in I^* \\ \langle w, x^i\rangle > t & i \in J^* \\ \langle w, x^i\rangle < t & i \in \{1,\dots,N\} \setminus (I^* \cup J^*) \\ \langle w, w\rangle = 1 & \end{array} \right\}.$$

Next, choose a subset $\bar{I}^* \subseteq I^*$ such that

$$\#\bar{I}^* = \operatorname{rank} \tilde{X}_{\bar{I}^*} = \operatorname{rank} \tilde{X}_{I^*} \tag{33}$$

(using the notation introduced in the statement of the Theorem). By Lemma 2, the two optimization problems above can be reformulated as

$$\max_{w,t} / \min_{w,t} \left\{ t \;\middle|\; \begin{array}{ll} \langle w, x^i\rangle = t & i \in \bar{I}^* \\ \langle w, x^i\rangle > t & i \in J^* \\ \langle w, x^i\rangle < t & i \in \{1,\dots,N\} \setminus (I^* \cup J^*) \\ \langle w, w\rangle = 1 & \end{array} \right\}. \tag{34}$$

Since $\langle w^*, x^i\rangle > t^*$ for $i \in J^*$ and $\langle w^*, x^i\rangle < t^*$ for $i \in \{1,\dots,N\} \setminus (I^* \cup J^*)$ and (w^*, t^*) is a solution of one of the two optimization problems (34), it follows that (w^*, t^*) must be a local solution of one of the two optimization problems

$$\max_{w,t} / \min_{w,t} \left\{ t \;\middle|\; \langle w, x^i\rangle = t \quad \left(i \in \bar{I}^*\right); \langle w, w\rangle = 1 \right\}. \tag{35}$$

By (33), these problems satisfy the assumption of Lemma 3 with $X_* := X_{\bar{I}^*}$ and $k := \#\bar{I}^*$. According to that Lemma we have that $0 < \gamma_{\bar{I}^*} \leq 1$ with γ_I as introduced in the statement of this Theorem.

In the case of $\gamma_{\bar{I}^*} < 1$ it follows from Lemma 3 (last statement), that $t^* \neq 0$. Then, by feasibility of (w^*, t^*) in (35), we have that

$$(t^*)^{-1} \tilde{X}_{\bar{I}^*}^T w^* = \mathbf{1} \quad (= (1,\dots,1) \in \mathbb{R}^{\#\bar{I}^*}).$$

Consequently, $-\mathbf{1} \in \operatorname{range} X_{\bar{I}^*}^T$, and thus,

$$\operatorname{rank} \tilde{X}_{\bar{I}^*} = \operatorname{rank} \begin{pmatrix} X_{\bar{I}^*} \\ -\mathbf{1}^T \end{pmatrix} = \operatorname{rank} (X_{\bar{I}^*}^T \mid -\mathbf{1}) = \operatorname{rank} X_{\bar{I}^*}^T \leq n.$$

Since also $\operatorname{rank} \tilde{X}_{\bar{I}^*} \leq \operatorname{rank} \tilde{X}$, we have shown that $\bar{I}^* \in \Phi_1$. Therefore, with the definitions of t_I, w_I in the statement of this Theorem, we infer from Lemma 3 that

$$(w^*, t^*) \in \{(w_{\bar{I}^*}, t_{\bar{I}^*}), (-w_{\bar{I}^*}, -t_{\bar{I}^*})\}.$$

Thus,

$$\begin{aligned}\Delta = \Delta(w^*, t^*) &\leq \max\left\{\Delta(w_{\bar{I}^*}, t_{\bar{I}^*}), \Delta(-w_{\bar{I}^*}, -t_{\bar{I}^*})\right\} \\ &\leq \max_{I\in\Phi_1} \max\left\{\Delta(w_I, t_I), \Delta(-w_I, -t_I)\right\} = \Delta_1 \end{aligned} \tag{36}$$

with Δ_1 as introduced in the statement of this Theorem.

If, in contrast, $\gamma_{\bar{I}^*} = 1$, then by Lemma 3 we observe that $t^* = 0$, and, $\operatorname{rank} X_{\bar{I}^*} = \#\bar{I}^* - 1$. The second equality in (33) along with $\bar{I}^* \subseteq I^*$ yields that $\operatorname{rank} X_{I^*} = \operatorname{rank} X_{\bar{I}^*}$. Hence, the first equality in (33) provides the relation

$$\operatorname{rank} X_{I^*} = \operatorname{rank} \tilde{X}_{I^*} - 1. \tag{37}$$

Moreover, by definition of I^*, one has that $w^* \in \operatorname{Ker} X_{I^*}^T \cap \mathbb{S}^{n-1}$, so that

$$w^* \in \operatorname*{arg\,max}_{w\in\operatorname{Ker} X_{I^*}^T\cap\mathbb{S}^{n-1}} |\mu^{emp}(w,0) - \mu^{cap}(w,0)| = \operatorname*{arg\,max}_{w\in\operatorname{Ker} X_{I^*}^T\cap\mathbb{S}^{n-1}} |\mu^{emp}(w,0) - \tfrac{1}{2}| \tag{38}$$

as a consequence of (32). We claim that

$$w^* \in \pm A, \quad \text{where } A := \operatorname*{arg\,max}_{w\in\operatorname{Ker} X_{I^*}^T\cap\mathbb{S}^{n-1}} \mu^{emp}(w,0). \tag{39}$$

Observe first that

$$\mu^{emp}(w,0) \geq \tfrac{1}{2} \quad \forall w \in A \tag{40}$$

due to the following relation resulting from (3):

$$2\mu^{emp}(w,0) \geq \mu^{emp}(w,0) + \mu^{emp}(-w,0) \geq 1 \quad \forall w \in A.$$

Moreover, we have that $A \neq \emptyset$. Indeed, $\operatorname{Ker} X_{I^*}^T \cap \mathbb{S}^{n-1}$ is nonempty (it contains w^*), and it is compact, so that the upper semicontionuous function $\mu^{emp}(w,0)$ (see Lemma 1) attains its maximum over this set. Hence, we may choose some $w' \in A$. From (40) we know that $\mu^{emp}(w',0) \geq \frac{1}{2}$. Assume that $w^* \notin A$. Then by (38),

$$\left|\mu^{emp}(w^*,0) - \tfrac{1}{2}\right| \geq \left|\mu^{emp}(w',0) - \tfrac{1}{2}\right| = \mu^{emp}(w',0) - \tfrac{1}{2} > \mu^{emp}(w^*,0) - \tfrac{1}{2}, \tag{41}$$

whence $\mu^{emp}(w^*,0) < \frac{1}{2}$. Once more, (3) yields that

$$\mu^{emp}(w^*,0) + \mu^{emp}(-w^*,0) \geq 1.$$

Taking into account that $-w^* \in \operatorname{Ker} X_{I^*}^T \cap \mathbb{S}^{n-1}$, this leads to

$$\begin{aligned}|\mu^{emp}(w^*,0) - \tfrac{1}{2}| &= \tfrac{1}{2} - \mu^{emp}(w^*,0) \leq \mu^{emp}(-w^*,0) - \tfrac{1}{2} \\ &\leq \mu^{emp}(w',0) - \tfrac{1}{2} \leq |\mu^{emp}(w^*,0) - \tfrac{1}{2}|,\end{aligned} \tag{42}$$

so that the whole chain becomes an equality. Therefore, $\mu^{emp}(-w^*,0) = \mu^{emp}(w',0)$ or $-w^* \in A$. This proves (39).

At the same time, (42) shows that if $w^* \notin A$, then not only w^* but also $-w^* \in A$ realizes the discrepancy (see (31)):

$$\Delta = |\mu^{emp}(w^*,0) - \tfrac{1}{2}| \leq |\mu^{emp}(-w^*,0) - \tfrac{1}{2}| \leq \Delta.$$

Therefore, in case of $w^* \notin A$ we may assume that we have chosen in (31) $-w^* \in A$ rather than w^* from the very beginning. Consequently, we may assume w.l.o.g. that $w^* \in A$:

$$w^* \in \operatorname*{arg\,max}_{w \in \operatorname{Ker} X_{I^*}^T \cap \mathbb{S}^{n-1}} \mu^{emp}(w,0) \tag{43}$$

in (39). By (30), $w^* \in \operatorname{Ker} X_{I^*}^T \cap \mathbb{S}^{n-1}$ implies that $\operatorname{rank} \tilde{X}_{I^*} \leq \min\left\{n, \operatorname{rank} \tilde{X}\right\}$.

We claim the existence of some index set $\hat{I}$ and of some vector $\hat{w}$ such that

$$\begin{aligned} &I^* \subseteq \hat{I} \subseteq \{1,\ldots,N\},\ \operatorname{rank} \tilde{X}_{\hat{I}} = \min\left\{n, \operatorname{rank} \tilde{X}\right\}, \\ &\hat{w} \in \operatorname{Ker} X_{\hat{I}}^T \cap \mathbb{S}^{n-1},\ \mu^{emp}(\hat{w},0) = \mu^{emp}(w^*,0). \end{aligned} \tag{44}$$

If $\operatorname{rank} \tilde{X}_{I^*} = \min\left\{n, \operatorname{rank} \tilde{X}\right\}$, then we may choose $\hat{I} := I^*$ and $\hat{w} := w^*$ in (44). Otherwise, $\operatorname{rank} \tilde{X}_{I^*} < \min\left\{n, \operatorname{rank} \tilde{X}\right\}$ and we make use of Lemma 4 starting with the data $I_0 := I^*$ and $w_0 := w^*$. Observe that by virtue of (37) and (43), I_0 and w_0 satisfy the assumptions (19) and (20) of that Lemma. Accordingly, we derive the existence of some index set $I_1 \supseteq I_0$ and w_1 satisfying the relations (21) and (22). In particular, the first relation in (21) yields that $w_1 \in \operatorname{Ker} X_{I_0}^T \cap \mathbb{S}^{n-1}$ due to $I_0 \subseteq I_1$, whence the second relation in (21) entails that

$$w_1 \in \operatorname*{arg\,max}_{w \in \operatorname{Ker} X_{I_1}^T \cap \mathbb{S}^{n-1}} \mu^{emp}(w,0).$$

Moreover, we infer from (20) and (22) that $\operatorname{rank} X_{I_1} = \operatorname{rank} \tilde{X}_{I_1} - 1$ and from $w_1 \in \operatorname{Ker} X_{I_1}^T$ and (30) that $\operatorname{rank} \tilde{X}_{I_1} \leq \min\left\{n, \operatorname{rank} \tilde{X}\right\}$.

Now, if $\operatorname{rank} \tilde{X}_{I_1} = \min\left\{n, \operatorname{rank} \tilde{X}\right\}$, then we may choose $\hat{I} := I_1$ and $\hat{w} := w_1$ in (44) due to (21) and $w_0 = w^*$. Otherwise, $\operatorname{rank} \tilde{X}_{I_1} < \min\left\{n, \operatorname{rank} \tilde{X}\right\}$ and so the assumptions (19) and (20) of Lemma 4 are also satisfied for I_1 and w_1 instead of I_0 and w_0. This allows us to apply Lemma 4 again. In this way, a sequence of index sets I_k and of points w_k ($k = 1, 2, \ldots$) is obtained for which $I^* = I_0 \subseteq I_k$ and by (21) and (22)

$$w_k \in \operatorname{Ker} X_{I_k}^T \cap \mathbb{S}^{n-1},\ \mu^{emp}(w_k,0) = \mu^{emp}(w_0,0),\ \operatorname{rank} \tilde{X}_{I_k} = \operatorname{rank} \tilde{X}_{I_{k-1}} + z_k,$$

where $z_k \in \mathbb{N}$ and $z_k \geq 1$. Since $\operatorname{rank} \tilde{X}_{I_k} \leq \min\left\{n, \operatorname{rank} \tilde{X}\right\}$ by (30), the last relation implies that, after finitely many steps, we arrive at the situation $\operatorname{rank} \tilde{X}_{I_k} = \min\left\{n, \operatorname{rank} \tilde{X}\right\}$, so that we may define $\hat{I} := I_k$ and $\hat{w} := w_k$ in (44).

Next, recall that $\mu^{emp}(w^*,0) \geq \frac{1}{2}$ by (43) and (40). Hence, by (31), (38) and (44),

$$\Delta = \left|\mu^{emp}(w^*,0) - \tfrac{1}{2}\right| = \mu^{emp}(w^*,0) - \tfrac{1}{2} = \mu^{emp}(\hat{w},0) - \tfrac{1}{2} = \left|\mu^{emp}(\hat{w},0) - \tfrac{1}{2}\right|. \tag{45}$$

This relation shows that $(\hat{w}, 0)$ realizes the discrepancy Δ as much as $(w^*, t^* = 0)$. Therefore, we may repeat the beginning of this proof until (35) just replacing (w^*, t^*) by $(\hat{w}, 0)$. In particular, analogously to the index set I^* introduced there, we define

$$I_* := \{i \in \{1,\ldots,N\} \mid \langle \hat{w}, x^i \rangle = 0\}$$

Following the previous arguments from (33) to (35), we may find an index set $\bar{I}_* \subseteq I_*$ such that

$$\#\bar{I}_* = \operatorname{rank} \tilde{X}_{\bar{I}_*} = \operatorname{rank} \tilde{X}_{I_*} \tag{46}$$

and $(\hat{w}, 0)$ is a local solution of one of the two optimization problems

$$\max_{w,t} / \min_{w,t} \left\{ t \;\middle|\; \left\langle w, x^i \right\rangle = t \quad \left(i \in \bar{I}_*\right); \langle w, w \rangle = 1 \right\}. \tag{47}$$

By (46), these problems satisfy the assumption of Lemma 3 with $X_* := X_{\bar{I}_*}$ and $k := \#\bar{I}_*$. According to that Lemma (last statement) we have that $\gamma_{\bar{I}_*} = 1$ with γ_I as introduced in the statement of this Theorem. Repeating the argument below (43) for $\hat{w} \in \operatorname{Ker} X_{I_*}^T \cap \mathbb{S}^{n-1}$, we observe that

$$\operatorname{rank} \tilde{X}_{I_*} \leq \min \left\{ n, \operatorname{rank} \tilde{X} \right\}.$$

On the other hand, since $\hat{I} \subseteq I_*$ by (44) and by definition of I_*, the rank relation in (44) leads to

$$\operatorname{rank} \tilde{X}_{I_*} \geq \operatorname{rank} \tilde{X}_{\hat{I}} = \min \left\{ n, \operatorname{rank} \tilde{X} \right\},$$

whence, along with (46)

$$\#\bar{I}_* = \operatorname{rank} \tilde{X}_{\bar{I}_*} = \min \left\{ n, \operatorname{rank} \tilde{X} \right\}. \tag{48}$$

Summarizing, we have shown that $\bar{I}_* \in \Phi_0$.

If in (48) $\operatorname{rank} \tilde{X}_{\bar{I}_*} = \operatorname{rank} \tilde{X}$, then there exist coefficients λ_j^i such that

$$\begin{pmatrix} x^i \\ -1 \end{pmatrix} = \sum_{j \in \bar{I}_*} \lambda_j^i \begin{pmatrix} x^j \\ -1 \end{pmatrix} \quad \forall i = 1, \ldots, N.$$

Therefore, we have for all $i = 1, \ldots, N$ and all $w \in \operatorname{Ker} X_{\bar{I}_*}^T$ that

$$\langle x^i, w \rangle = \left\langle \begin{pmatrix} x^i \\ -1 \end{pmatrix}, \begin{pmatrix} w \\ 0 \end{pmatrix} \right\rangle = \sum_{j \in \bar{I}_*} \lambda_j^i \left\langle \begin{pmatrix} x^j \\ -1 \end{pmatrix}, \begin{pmatrix} w \\ 0 \end{pmatrix} \right\rangle = \sum_{j \in \bar{I}_*} \lambda_j^i \langle x^j, w \rangle = 0.$$

This amounts to saying that $\operatorname{Ker} X_{\bar{I}_*}^T \subseteq \operatorname{Ker} X^T$, whence $\mu^{emp}(w, 0) = 1$ and so $\Delta(w, 0) = 1/2$ for all $w \in \operatorname{Ker} X_{\bar{I}_*}^T \cap \mathbb{S}^{n-1}$. Since $\hat{w} \in \operatorname{Ker} X_{\bar{I}_*}^T \cap \mathbb{S}^{n-1}$, we conclude from (45) that

$$\Delta = \left| \mu^{emp}(\hat{w}, 0) - \tfrac{1}{2} \right| = \Delta(\hat{w}, 0) = 1/2 = \Delta(w, 0) \quad \forall w \in \operatorname{Ker} X_{\bar{I}_*}^T \cap \mathbb{S}^{n-1},$$

whence $\Delta \leq \Delta_0$ with Δ_0 as introduced in the statement of this Theorem.

Otherwise, if in (48) $\operatorname{rank} \tilde{X}_{\bar{I}_*} = n$, then $\dim \operatorname{Ker} \tilde{X}_{\bar{I}_*}^T = 1$ and so, due to $\bar{I}_* \subseteq I_*$,

$$\begin{pmatrix} \hat{w} \\ 0 \end{pmatrix} \in \operatorname{Ker} \tilde{X}_{I_*}^T \cap (\mathbb{S}^{n-1} \times \{0\}) \subseteq \operatorname{Ker} \tilde{X}_{\bar{I}_*}^T \cap (\mathbb{S}^{n-1} \times \{0\}) = \left\{ \begin{pmatrix} \tilde{w} \\ 0 \end{pmatrix}, \begin{pmatrix} -\tilde{w} \\ 0 \end{pmatrix} \right\}$$

for some $\tilde{w} \in \operatorname{Ker} X_{\bar{I}_*}^T \cap \mathbb{S}^{n-1}$. Thus, $\hat{w} \in \{\tilde{w}, -\tilde{w}\}$. Let $w \in \operatorname{Ker} X_{\bar{I}_*}^T \cap \mathbb{S}^{n-1}$ be arbitrary. Then,

$$\begin{pmatrix} w \\ 0 \end{pmatrix} \in \operatorname{Ker} \tilde{X}_{\bar{I}_*}^T \cap (\mathbb{S}^{n-1} \times \{0\})$$

and, hence, $w = \pm \tilde{w}$. Therefore,

$$\begin{aligned} \Delta &= \left| \mu^{emp}(\hat{w}, 0) - \tfrac{1}{2} \right| = \Delta(\hat{w}, 0) \\ &\leq \max \left\{ \Delta(\tilde{w}, 0), \Delta(-\tilde{w}, 0) \right\} \\ &= \max \left\{ \Delta(w, 0), \Delta(-w, 0) \right\} \quad \forall w \in \operatorname{Ker} X_{\bar{I}_*}^T \cap \mathbb{S}^{n-1}. \end{aligned}$$

As in the previous case we conclude that $\Delta \leq \Delta_0$.

Summarizing, our proof has shown by case distinction that necessarily $\Delta \leq \Delta_1$ (see (36)) or $\Delta \leq \Delta_0$. Therefore, $\Delta \leq \max\{\Delta_1, \Delta_0\}$. On the other hand, each of the quantities Δ_1, Δ_0 is either zero or corresponds to a concrete value $\Delta(w, t)$ for some $w \in \mathbb{S}^{n-1}$ and $t \in [-1, 1]$. Hence, in any case $\max\{\Delta_1, \Delta_0\} \leq \Delta$. This finishes the proof. □

We observe that the cardinality of index sets to be checked in the proven formula is at most

$$\sum_{i=1}^{\min\{n,\operatorname{rank}\tilde{X}\}} \binom{N}{i}.$$

Whether calculating the spherical cap discrepancy is NP-hard (or W[1]-hard) is left open for future work. Clearly, this aspect of complexity limits the application of the formula to low-dimensional spheres and moderate sample sizes. Hence, it will not be suitable for verifying asymptotic aspects of sampling schemes. On the other hand, it may be used to correctly calibrate the efficiency of sampling schemes within a certain range of the sample size.

4 Numerical Illustration

In this section we illustrate the application of the derived formula for the spherical cap discrepancy to spheres $\mathbb{S}^2$ to $\mathbb{S}^5$ with sample sizes reaching from 2000 to 100 depending on dimension. Samples were generated by normalizations of Monte Carlo simulated independent Gaussian distributions which are approximations of the uniform distribution on the sphere. For the sake of comparison, we oppose the results to the application of an easily computable lower estimate of the discrepancy as it was used, e.g., in [1]: Given a sample $\{x^1, \ldots, x^N\}$, we clearly have that

$$\tilde{\Delta} := \max_{i=1,\ldots,N} \sup_{t\in[-1,1]} \left|\mu^{emp}\left(x^i, t\right) - \mu^{cap}\left(x^i, t\right)\right| \leq \Delta.$$

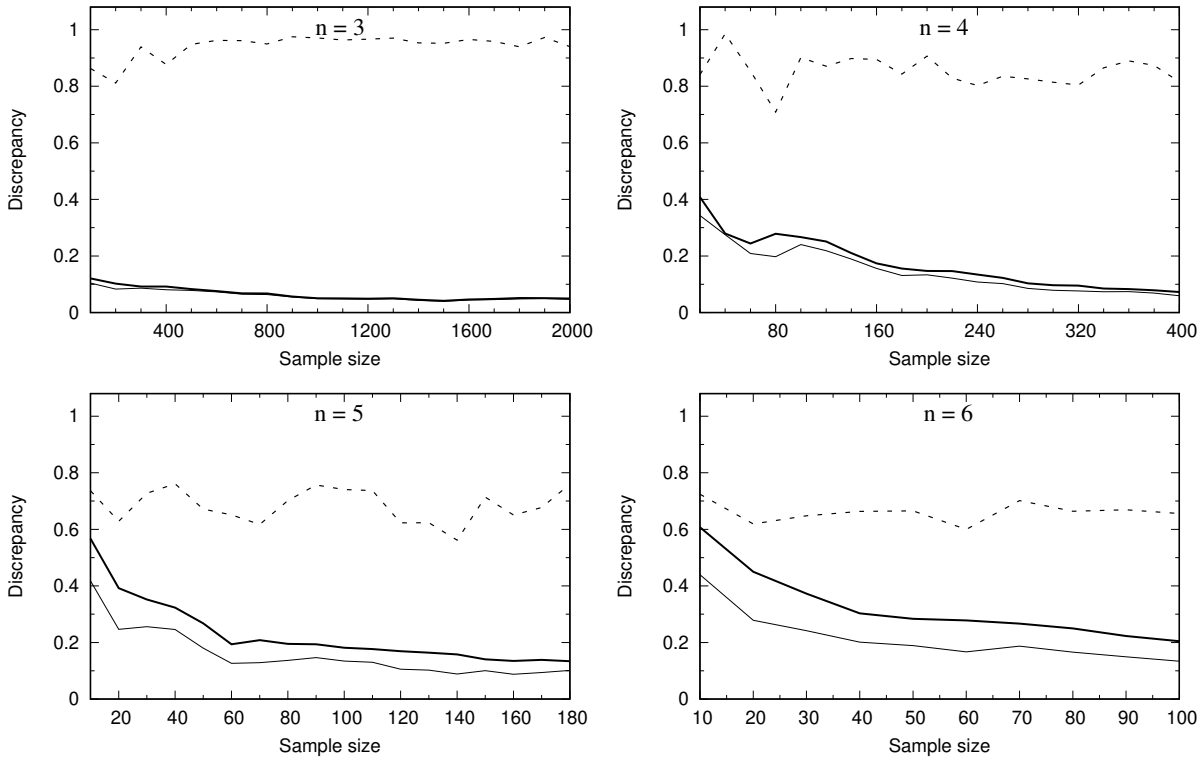

Figure 1: Plot of discrepancy Δ (thick line), of its lower estimate $\tilde{\Delta}$ (thin line) and of the ratio $\tilde{\Delta}/\Delta$ (dashed line) for different dimensions (n=3,4,5,6) and sample sizes.

Fig. 1 shows the numerical results. We observe the following trends:

- Both, Δ and $\tilde{\Delta}$ are decreasing with increasing sample size.
- The absolute difference between Δ and $\tilde{\Delta}$ decreases with the sample size.
- The absolute difference between Δ and $\tilde{\Delta}$ increases with the dimension of the sphere.
- The ratio between $\tilde{\Delta}$ and Δ is basically constant for variable sample size in each dimension of the sphere (with different values of the constant).
- The constant itself is decreasing with the dimension of the sphere.

In particular, it seems that the discrepancy can be replaced by its lower estimate without loss of information in $\mathbb{S}^2$ starting from a sample size of approximately 500. For larger dimension or sample size, it appears that at least the decay rate with respect to the sample size is well reflected by the lower estimate (approximately constant ratio with the true discrepancy), while the deviation from the true discrepancy becomes significant.

All computations are performed on a standard computer with single CPU (3.2 GHz). Table 1 displays the CPU time for selected instances of the tested range of N and n.

N	10	20	40	100	180	400	1000	2000
$\mathbb{S}^2$	0	0	0	3	30	499	15924	252313
$\mathbb{S}^3$	0	0	2	139	1911	77449	-	-
$\mathbb{S}^4$	0	0	14	1932	48134	-	-	-
$\mathbb{S}^5$	0	1	92	32658	-	-	-	-

Table 1: CPU time (in seconds) of the enumeration formula for selected instances.

In order to illustrate even more directly the application of the proven formula, we provide a comparison of 4 sampling schemes on $\mathbb{S}^2$ for small sample sizes (≤ 1000).

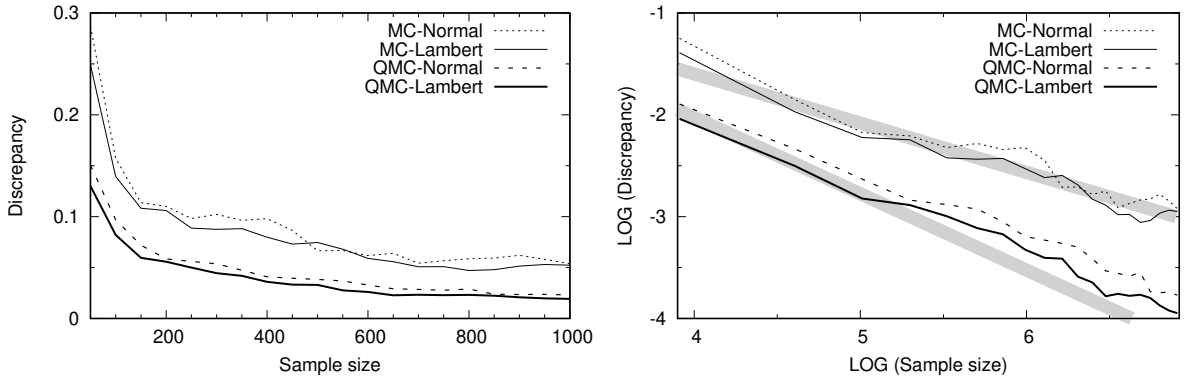

Figure 2: Discrepancy as a function of sample size for 4 different sampling schemes in original form (left) and Log-Log Plot (right). For details see text.

The first two methods are based on the already mentioned fact that the normalization to unit length of a standard Gaussian distribution $\mathcal{N}(0_m, I_m)$ yields a uniform distribution on $\mathbb{S}^{m-1}$. Therefore, we may simulate the Gaussian distribution via Monte Carlo (MC) or via Quasi-Monte Carlo (QMC). As an alternative, we follow the proposal in [1], to use the equal-area Lambert transform from the unit square to $\mathbb{S}^2$, again for MC and QMC. For QMC, we applied in both cases Sobol' sequences as a special case of low-discrepancy sequences.

Figure 2 (left) shows the corresponding plots of the discrepancy as a function of the sample size for 20 steps à 50 sampling points. Not surprisingly, the QMC-based samples clearly outperform their MC counterparts. Moreover, in both classes, the Lambert transformation yields slightly better results than the normalization of Gaussians. The Log-Log plot (right) incorporates two gray strips with slopes identical to -1/2 (upper strip) and -3/4 (lower strip) with empirically shifted intercepts. It can be seen that the MC-based methods are closely tied with the expected decay rate of -1/2, whereas the QMC counterparts get a slope slightly above the optimal rate of -3/4.

References

[1] C. Aistleitner, J.S. Brauchart, and J. Dick. Point Sets on the Sphere $\mathbb{S}^2$ with Small Spherical Cap Discrepancy. *Discrete Comput. Geom.*, 48:990–1024, 2012.

[2] William Chen, Anand Srivastav, and Giancarlo Travaglini. *A Panorama of Discrepancy Theory*. Springer, Heidelberg, 2014.

[3] Carola Doerr, Michael Gnewuch, and Magnus Wahlström. Calculation of Discrepancy Measures and Applications. In William Chen, Anand Srivastav, and Giancarlo Travaglini, editors, *A Panorama of Discrepancy Theory*, pages 621–678. Springer, Heidelberg, 2014.

[4] M. Drmota and R.F. Tichy. *Sequences, Discrepancies and Application*. Springer, Berlin, 1997.

[5] P. Giannopoulos, C. Knauer, M. Wahlström, and D. Werner. Hardness of discrepancy computation and ε-net verification in high dimension. *J. Complexity*, 28:162–176, 2012.

[6] M. Gnewuch, A. Srivastav, and C. Winzen. Finding optimal volume subintervals with k points and calculating the star discrepancy are NP-hard problems. *J. Complexity*, 25:115–127, 2009.

[7] J. Grabner and R.F. Tichy. Spherical Designs, Discrepancy and Numerical Integration. *Math. Comp.*, 60:327–336, 1993.

[8] R. Henrion, C. Küchler, and W. Römisch. Discrepancy distances and scenario reduction in two-stage stochastic integer programming. *J. Ind. Manag. Optim.*, 4:363–384, 2008.

[9] R. Henrion, C. Küchler, and W. Römisch. Scenario reduction in stochastic programming with respect to discrepancy distances. *Comput. Optim. Appl.*, 43:67–93, 2009.

[10] J. Matoušek. *Geometric Discrepancy*. Springer, Berlin, 1999.

[11] W. Römisch. Stability of Stochastic Programming Problems. In A. Ruszczynski and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 483–554. Elsevier, Amsterdam, 2003.

[12] W. van Ackooij and R. Henrion. (Sub-) Gradient formulae for probability functions of random inequality systems under Gaussian distribution. *SIAM/ASA J. Uncertainty Quantification*, 5:63–87, 2017.