On Statistical Learning of Simplices: Unmixing Problem Revisited

Preprint · October 2018

CITATIONS
0

READS
202

6 authors, including:

Amir Najafi
Institute for Research in Fundamental Sciences (IPM)
21 PUBLICATIONS **663** CITATIONS
SEE PROFILE

Seyed amir hossein Saberi
Sharif University of Technology
4 PUBLICATIONS **1** CITATION
SEE PROFILE

Abolfazl Motahari
Sharif University of Technology
56 PUBLICATIONS **442** CITATIONS
SEE PROFILE

Babak Khalaj
Sharif University of Technology
253 PUBLICATIONS **2,505** CITATIONS
SEE PROFILE

On Statistical Learning of Simplices: Unmixing Problem Revisited

Amir Najafi * Saeed Ilchi * Amir H. Saberi † Seyed Abolfazl Motahari *
Babak H. Khalaj † Hamid R. Rabiee ‡

*Data Analytics Lab (DAL), Computer Engineering Dept.,
‡Data science and Machine learning Lab (DML), Computer Engineering Dept.,
†Electrical Engineering Dept.,
Sharif University of Technology, Tehran, Iran

Abstract

We study the sample complexity of learning a high-dimensional simplex from a set of points uniformly sampled from its interior. Learning of simplices is a long studied problem in computer science and has applications in computational biology and remote sensing, mostly under the name of 'spectral unmixing'. We theoretically show that a sufficient sample complexity for reliable learning of a K-dimensional simplex up to a total-variation error of ϵ is $O\left(\frac{K^2}{\epsilon}\log\frac{K}{\epsilon}\right)$, which yields a substantial improvement over existing bounds. Based on our new theoretical framework, we also propose a heuristic approach for the inference of simplices. Experimental results on synthetic and real-world datasets demonstrate a comparable performance for our method on noiseless samples, while we outperform the state-of-the-art in noisy cases.

1 Introduction

High-dimensional measurements from a physical system can sometimes be thought as convex combinations of a number of unknown sources, where profiles associated with each source might not be easy to assess, separately. Even though both the source profiles and combination weights are assumed to be unknown, however, one can still attempt to infer them through collection and careful analysis of a large number of i.i.d. measurements. This computational problem, generally known as the 'unmixing problem', arises in many areas including hyper-spectral remote sensing [1, 2], mixture modeling, and analysis of tumor heterogeneity in computational biology and bioinformatics [3, 4, 5]. In all the above cases, unmixing the data refers to inferring all the unknown source profiles as well as the weights for each observed sample.

From a geometric point of view, this problem can be formulated as learning of a high-dimensional simplex. A K-dimensional simplex is defined as the set of all convex combinations of $K+1$ points in $\mathbb{R}^K$, also called vertices. For example, simplices with $K=2$ and $K=3$ correspond to triangles and tetrahedrons, respectively. Learning of a simplex refers to inferring its vertices through observing randomly chosen points from its interior. This problem is also closely related to solving the following

An implementation of our method is available at: `github.com/seyedsaberi/simplex`

E-mails: *{najafy,silchi}@ce.sharif.edu, †sah.saberi@ee.sharif.edu, {*motahari,†khalaj,‡rabiee}@sharif.edu

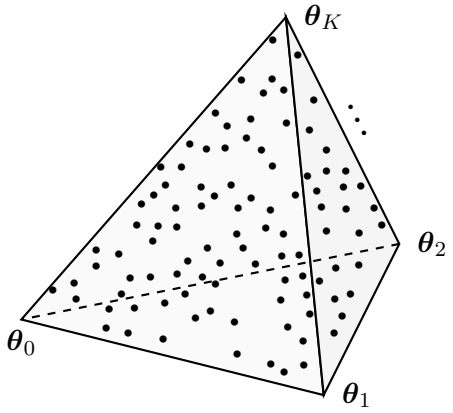


Figure 1: Learning of simplices from a geometric viewpoint: Assume a K-dimensional simplex specified by $K+1$ latent vertices $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \in \mathbb{R}^K$ which correspond to the columns of the vertex matrix $\boldsymbol{\Theta}$. Given n random points uniformly sampled from the interior of the simplex, our aim is to find an estimator for $\boldsymbol{\Theta}$, say $\hat{\boldsymbol{\Theta}}$, such that the uniform probability measures over the simplices specified by $\boldsymbol{\Theta}$ and $\hat{\boldsymbol{\Theta}}$ have a total variation distance of at most ϵ with probability at least $1-\zeta$, for any given $\epsilon, \zeta > 0$.

set of equations:

$$\boldsymbol{\Theta}\boldsymbol{p}_i = \boldsymbol{X}_i \quad , \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{X}_i \in \mathbb{R}^K$ represents the ith observed point, each column of the $K \times (K+1)$ matrix $\boldsymbol{\Theta}$ denotes a source profile (or alternatively a vertex of the simplex), and $\boldsymbol{p}_i \in [0,1]^{K+1}$ is the ith weight vector with $\mathbf{1}^T \boldsymbol{p}_i = 1$. Here, the matrix $\boldsymbol{\Theta}$ and all the weight vectors $\boldsymbol{p}_i$ are assumed to be unknown. In order to transform the problem into a statistical setting, we assume $\boldsymbol{p}_i$s are generated independently from a uniform Dirichlet distribution which means data points are distributed uniformly inside the simplex. It is easy to show that (1), for any n, has an infinite number of solutions. However, as we show it in proceeding sections, the solution which corresponds to the minimum volume of the simplex specified by $\boldsymbol{\Theta}$ is the maximum-likelihood estimator of the above-mentioned statistical inference problem.

Assume a uniform probability distribution in $\mathbb{R}^K$ which is supported over the simplex specified by $\boldsymbol{\Theta}$. Then, our aim is to design an algorithm that with probability at least $1-\zeta$ estimates this probability distribution with a total variation error of at most ϵ, for any given $\epsilon, \zeta > 0$, given that the number of observations satisfies $n \geq \text{poly}\left(\epsilon^{-1}, \log\left(1/\zeta\right)\right)$. In this setting, ϵ also coincides with the normalized Nikodym distance, or the relative volume of the symmetric difference between the estimated simplex and the true one [6]. The geometric interpretation of this problem is illustrated in Figure 1. Some interesting questions in this setting would be: How many samples are required for the above-mentioned estimation problem (sample complexity)? or, How well does any particular algorithm perform if observations become noisy? In this study, we show that a sufficient sample complexity to ensure a total variation error of at most ϵ with probability at least $1-\zeta$ is

$$n \geq O\left(\frac{1}{\epsilon}\left[K^2 \log\frac{K}{\epsilon} + \log\frac{1}{\zeta}\right]\right), \tag{2}$$

where achievability is guaranteed by the Maximum Likelihood Estimator (MLE). Our approach is based on tools from statistical learning theory and high-dimensional geometry. To the best of our knowledge, this result yields a significant improvement over the only existing bound of [7], which is $n \geq O\left(K^{22}/\epsilon^2\right)$. However, finding the MLE requires solving a combinatoric optimization problem and is proved to have an exponential-time computational complexity [8], while the method of [7] is polynomial-time. Motivated by this fact, we also propose a novel continuously-relaxed version of the original ML problem, followed by a heuristic Gradient Descent (GD)-based algorithm for its numerical optimization. Theoretical sample complexity for the proposed relaxed-MLE is $O\left(\epsilon^{-2}\log\frac{1}{\epsilon}\right)$, while maintaining the same polynomial dependence on dimension K as that of MLE. We have tested our method on a number of synthetic and real-world datasets. Experimental results

demonstrate an acceptable performance for our method in the noiseless case, while they show a considerable superiority over rival techniques when data samples are noisy.

The paper is organized as follows: Section 1.1 reviews the related works. Section 2 explains the mathematical notation and also formally defines our problem. Our proposed method and theoretical results are presented in Section 3. Section 4 is devoted to experimental results on both synthetic and real-world datasets. Finally, conclusions are made in Section 5.

1.1 Related Works

Previous works in this domain can be divided into two separate categories: papers that follow a theoretical approach, and those trying to solve a real-world problem via heuristics.

From a theoretical viewpoint, in a seminal work on learning linear transformations [9], authors have proved the efficient learnability[1] of *hyper-cubes* via $O\left(K^8\right)$ samples. They suggested a possibly similar result for simplices. However, authors in [7] have recently proved that a sufficient sample size for efficient learnability of simplices is $n \geq O\left(K^{22}\right)$, which is the only existing bound with respect to dimension K prior to this paper. Regarding the estimation error ϵ, it is known that learning of high-dimensional polytopes requires $\tilde{O}\left(\epsilon^{-1}\right)$ [2] samples [11, 12]. In particular, Theorem 1 and Corollary 1 of [11] shows that the following bound holds for the expected total-variation error (or equivalently, the expected relative Nikodym error) of estimating a K-dimensional polytope with r vertices ($r \geq K+1$):

$$\epsilon \leq_K \frac{r \log n}{n}. \tag{3}$$

Here, $\leq_K$ indicates that the inequality holds up to a factor that only depends on dimension K. The bound is achievable via the ML estimator which in this case would be the minimum-volume polytope with at most r vertices that includes all the data points. To the best of our knowledge, the (seemingly) optimal dependence of sample complexity on K has been left as an open problem. In this paper, we introduce a new approach to solve this problem for the special case of simplices, i.e. when $r = K+1$. Our method is based on statistical analysis of the $K+1$ facets of a K-dimensional simplex; However, this technique may not work for the broader case of K-dimensional r-polytopes with $r > K+1$, due to the rapidly increasing number of facets as a function of vertex number r in high dimensions. To the best of our knowledge, the best sample complexity results for learning of r-polytopes when r is strictly larger than $K+1$ is still (3).

A nice line of research focuses on estimating convex bodies in general (and simplices, in particular) through the *convex hull* of n uniformly sampled data points, denoted by $\hat{K}_n$. For example, it has been shown that the random polytope $\hat{K}_n$ is rate optimal over the class of all convex bodies in $\mathbb{R}^K$ in a minimax sense, when Nikodym metric is used [11]. There are also other studies on the estimation power and/or the expected number of vertices of Convex Hull Estimators (CHE) for different classes of geometric shapes such as ellipsoids and polytopes; See [13, 14, 15]. As long as one is concerned about estimating the data generating distribution as a whole, without any constraint on the shape of its support, CHE is shown to attain the following total-variation error bound in an efficient (polynomial) time (Theorem 2 of [14]):

$$\epsilon \lesssim \frac{K^2}{n}\left(\frac{\log n}{K+1}\right)^{K-1} + \frac{(\log n)^{K-1}}{n}. \tag{4}$$

[1] In this paper, the term *learnable* refers to a learning task with a polynomial sample complexity w.r.t. ϵ and ζ. This notation is consistent with that of [10]. A learnable case which also has a polynomial computational complexity is referred to as *efficiently learnable*.

[2] We use $\tilde{O}(\cdot)$ to hide possible logarithmic dependencies

We show that, for the case of simplices, MLE achieves a better sample complexity bound by removing the exponential terms w.r.t. K in the r.h.s. of (4). However, unlike CHE, MLE is NP-hard. Once again, we should reiterate that CHE does not necessarily output a simplex and thus is not directly linked to the approach that we follow in this paper. In this work, we are interested in estimators in the class of all K-dimensional *simplices* and not the far richer class of all polytopes in $\mathbb{R}^K$.

Learning of simplices also has close ties with Nonnegative Matrix Factorization (NMF) research [16, 17]. In fact, with proper normalization, one can transform many NMF problems into learning of a simplex (or more generally, a polytope) as long as there exists a minimum-volume constraint or related regularizers in the formulation of their corresponding objective functions. A nice survey on NMF methods used for data clustering can be found in [18]. Most of the existing works on NMF focus on its real-world applications, while fewer interesting works have considered its theoretical identifiability issues [19, 20]. In addition, a nice study on the generalization theory of dictionary learning and other matrix factorizations can be found in [21]. In proceeding sections, we propose a relaxed scheme for learning of simplices that has some mathematical similarities to the regularization techniques that are known to be useful in this line of research; See, for example, the volume regularizer proposed in [16]. It is worth mentioning that the output of NMF, in general, may span a larger vector space than that of the data samples, which is another subtle difference between learning of simplices and NMF research. In any case, not much work has been done so far on the sample complexity of NMF problems, which means our work could be helpful for researchers working in this area.

From an algorithmic point of view, the ML estimator in the noiseless setting is equivalent to finding the minimum volume simplex that contains all the data points. This task is shown to be NP-hard, irrespective of the input representation which could be either via facets or vertices [8]. In this regard, [22] introduced an LP-relaxation that computes the determinant of a modified vertex matrix instead of the volume. Determinant can then be written as the sum of matrix co-factors and consequently optimized in a row-wise manner. However, authors do not provide any theoretical guarantees, while experiments are used for justification of their method. Authors of [23] studied a similar problem in lattice-based cryptography, where inputs are generated from a parallelepiped instead of a simplex. Our problem also shares some similarities to the Blind Source Separation (BSS) research [21]. In fact, [9] has already shown that at least one particular problem from Independent Component Analysis (ICA) area can be transformed into learning of a high-dimensional hyper-cube from uniform samples. However, assumptions on the statistical generation of weight vectors $\boldsymbol{p}_i$ are still crucially different between a hyper-cube and a simplex. As a result, in BSS research one usually employs high-order statistical moments for capturing the source signals [24], while we follow a different approach which is described in Section 3.

From a more applied perspective, learning of simplices is of practical interest in a variety of applications, including hyper-spectral imaging [25]. Hyper-spectral cameras capture electromagnetic energy scatters from a scene in multiple wavelengths, where the spectrum of each pixel is a combination of reflected (or radiated) patterns of its basic elements [2, 26]. Recently, this field has attracted an intensive research interest, which are mostly centered on analysis of the minimum volume simplex, c.f. [1, 2, 27, 28]. Another tightly related application to our paper is the analysis of common patterns in cancer tumors [4]. Tumors can have a high cell-type heterogeneity, and a sufficient knowledge of the genetic characteristics that correspond to each cell-type is vital for recommending an effective treatment. However, biological datasets are mostly in *bulk format* which means each sample is an aggregation of cell populations from all the different cell-types with unknown combination weights. Again, the idea of finding the smallest inclusive simplex for capturing these hidden characteristics is exploited in several recent articles [3, 4, 5, 29].

2 Notation and Definitions

For $K \in \mathbb{N}$, let us denote the number of vertices in our model by $K+1$. Without loss of generality[3], the dimensionality of data points can be assumed to be K. Let us denote $\boldsymbol{\Phi}$ as the set of all discrete $(K+1)$-dimensional probability mass functions, i.e. $\boldsymbol{\Phi} \triangleq \left\{ \boldsymbol{p} \in \mathbb{R}^{K+1} \middle| \ \sum_k p_k = 1,\ p_k \geq 0 \right\}$. $\boldsymbol{\Phi}$ is generally referred to as the *standard simplex*. Assume $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_K \in \mathbb{R}^K$, and let $\boldsymbol{\Theta} \triangleq [\boldsymbol{\theta}_0 | \cdots | \boldsymbol{\theta}_K]$ to denote the vertex matrix, where the kth column of $\boldsymbol{\Theta}$ represents the kth vertex. We define $\mathcal{S}$ as a K-simplex with the vertex matrix $\boldsymbol{\Theta}$ as

$$\mathcal{S} = \mathcal{S}(\boldsymbol{\Theta}) \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}^K \middle| \ \boldsymbol{x} = \sum_k p_k \boldsymbol{\theta}_k,\ \boldsymbol{p} \in \boldsymbol{\Phi} \right\}. \tag{5}$$

Also, $\mathbb{S}_K$ represents the set of all K-simplices in $\mathbb{R}^K$. For a K-simplex $\mathcal{S} \in \mathbb{S}_K$, $\mathrm{Vol}(\mathcal{S})$ denotes the Lebesgue measure (or volume) of $\mathcal{S}$. For $\mathcal{S} \in \mathbb{S}_K$, let $\mathcal{S}_{-k}$ be the kth polygonal facet of $\mathcal{S}$, or equivalently, the $(K-1)$-simplex obtained by removing the kth vertex of $\mathcal{S}$.

In Section 3, we argue that in continuously-relaxed regimes, learning of a simplex may heavily depend on its level of geometric regularity. In other words, simplices with more or less equally-sized facets are much easier to learn, compared to those which include very acute corners. Inspired by the *isoperimetric* inequality in geometry [30], let us define the $(\underline{\lambda}, \bar{\lambda})$-isoperimetricity property for a K-simplex as follows:

Definition 1. *A K-simplex $\mathcal{S}(\boldsymbol{\Theta}) \in \mathbb{S}_K$ for $\boldsymbol{\Theta} \triangleq [\boldsymbol{\theta}_0 | \cdots | \boldsymbol{\theta}_K]$ is said to be $(\underline{\lambda}, \bar{\lambda})$-isoperimetric for some $\underline{\lambda}, \bar{\lambda} > 0$, if*

$$\max_{k,k'} \ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'}\|_2 \ \leq \ \underline{\lambda} K \mathrm{Vol}(\mathcal{S})^{\frac{1}{K}} \quad , \quad \max_k \ \mathrm{Vol}(\mathcal{S}_{-k}) \ \leq \ \bar{\lambda} \mathrm{Vol}(\mathcal{S})^{\frac{K-1}{K}}.$$

The essence of $(\underline{\lambda}, \bar{\lambda})$-isoperimetricity property is to ensure that a given K-simplex is comparably stretched along all the K dimensions of $\mathbb{R}^K$. According to Lemma 7 (see Appendix B), for a perfectly regular simplex with equal side lines and all $K \in \mathbb{N}$, $\underline{\lambda}$ and $\bar{\lambda}$ can be chosen to be as small as 1 and e, respectively.

For any simplex $\mathcal{S} \in \mathbb{S}_K$, let $\mathcal{H}_k \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}^K | \ \boldsymbol{w}_k^T \boldsymbol{x} + b_k = 0 \right\}$ for $k = 0, \ldots, K$ represent the hyper-plane which encompasses the kth facet of $\mathcal{S}$. This way, $\boldsymbol{w}_k \in \mathbb{R}^K$ denotes a normal vector (with $\|\boldsymbol{w}_k\|_2 = 1$) and $b_k \in \mathbb{R}$ is a corresponding bias value. Vectors $\boldsymbol{w}_k$ are assumed to be outward w.r.t. $\mathcal{S}$. Then, the following defines the *planar distance* of a point in $\mathbb{R}^K$ from the simplex $\mathcal{S}$:

Definition 2. *For any point $\boldsymbol{x} \in \mathbb{R}^K$, let us define the planar distance of $\boldsymbol{x}$ from $\mathcal{S}$, denoted by $d_{\mathcal{S}}(\boldsymbol{x})$, as*

$$d_{\mathcal{S}}(\boldsymbol{x}) \triangleq \max \left\{ 0, \ \max_k \ \boldsymbol{w}_k^T \boldsymbol{x} + b_k \right\}. \tag{6}$$

We always have $d_{\mathcal{S}}(\boldsymbol{x}) \geq 0$, and since each $\boldsymbol{w}_k$ is an outward normal vector w.r.t. $\mathcal{S}$, $d_{\mathcal{S}}(\boldsymbol{x}) = 0$ holds if and only if $\boldsymbol{x} \in \mathcal{S}$. Also, it is easy to check that for any fixed $d_0 \geq 0$, the set $\left\{ \boldsymbol{x} \in \mathbb{R}^K | \ d_{\mathcal{S}}(\boldsymbol{x}) \leq d_0 \right\}$ is also a simplex which encompasses $\mathcal{S}$ (see Figure 2). For a simplex $\mathcal{S} \in \mathbb{S}_K$, $\mathbb{P}_{\mathcal{S}}$ denotes the uniform probability distribution which is supported over $\mathcal{S}$. Let $\rho_{\mathcal{S}}$ be the probability density function associated to $\mathbb{P}_{\mathcal{S}}$, then

$$\rho_{\mathcal{S}}(\boldsymbol{x}) \triangleq \frac{\mathbf{1}_{\mathcal{S}}(\boldsymbol{x})}{\mathrm{Vol}(\mathcal{S})} \quad \text{for} \quad \forall \boldsymbol{x} \in \mathbb{R}^K, \tag{7}$$

[3] Points within a K-simplex lie on a K-dimensional linear subspace, which is almost surely identifiable and efficiently learnable as long as $n \geq K+1$.

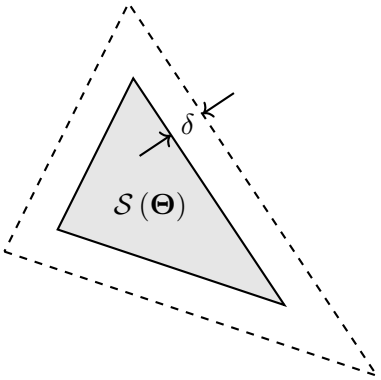


Figure 2: Graphical illustration of the *planar distance* $d_{\mathcal{S}(\boldsymbol{\Theta})}(\cdot)$ in Definition 2. The gray triangle shows an arbitrary two-dimensional simplex $\mathcal{S}(\boldsymbol{\Theta}) \in \mathbb{S}_2$. For some $\delta > 0$, the dashed contour corresponds to the set $\left\{\boldsymbol{x} \in \mathbb{R}^2 |\ d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{x}) = \delta\right\}$. As can be seen, the set of points in $\mathbb{R}^K$ with the same positive planar distance from a simplex $\mathcal{S} \in \mathbb{S}_K$ is in fact the boundary of another simplex whose facets are parallel to those of $\mathcal{S}$.

where $\mathbf{1}_{\mathcal{S}}(\boldsymbol{x})$ is the indicator function of the K-simplex $\mathcal{S}$ which returns 1 if $\boldsymbol{x} \in \mathcal{S}$, and zero otherwise. Also, $\mathcal{D}_{\mathrm{TV}}(\cdot,\cdot)$ represents the total-variation distance between two probability distributions.

Now, we can formally define the problem that has been tackled in this paper. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^K$ to be n i.i.d. samples which are generated uniformly from $\mathcal{S}_T \in \mathbb{S}_K$, i.e. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim \mathbb{P}_{\mathcal{S}_T}$. The problem is to find an approximation of $\mathcal{S}_T$, denoted by $\mathcal{S}^*$, from the dataset $\boldsymbol{D} \triangleq \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ such that with probability at least $1-\zeta$ the total variation between $\mathbb{P}_{\mathcal{S}^*}$ and $\mathbb{P}_{\mathcal{S}_T}$ is less than ϵ.

3 Statistical Learning of Simplices: Method and Results

In this section, we first formulate the MLE for estimation of $\mathcal{S}_T$ from the data data points in $\boldsymbol{D}$. We then give a new sample complexity bound for MLE which seems to be optimal with respect to both K and ϵ. Then, we propose a novel relaxation for the ML estimator that has the same sample complexity bound as that of MLE, at least with respect to dimension K.

Given a dataset $\boldsymbol{D}$, the ML estimator for $\mathcal{S}_T$, which we denote by $\mathcal{S}^*_{\mathrm{ML}} = \mathcal{S}^*_{\mathrm{ML}}(\boldsymbol{D})$ would be the maximizer of the log-likelihood function $\log \mathbb{P}_{\mathcal{S}}(\boldsymbol{D})$ according to the following formulation:

$$\mathcal{S}^*_{\mathrm{ML}} \triangleq \underset{\mathcal{S} \in \mathbb{S}_K}{\operatorname{argmax}} \left\{ \log \rho_{\mathcal{S}}(\boldsymbol{D}) = \log \prod_{i=1}^{n} \rho_{\mathcal{S}}(\boldsymbol{X}_i) = \sum_{i=1}^{n} \log \mathbf{1}_{\mathcal{S}}(\boldsymbol{X}_i) - n \log \mathrm{Vol}(\mathcal{S}) \right\}. \tag{8}$$

It is easy to see that the optimization in (8) aims to find the smallest simplex (in terms of volume), which contains all the data points in $\boldsymbol{D}$: For a simplex $\mathcal{S} \in \mathbb{S}_K$, even if a single data point happens to be outside of $\mathcal{S}$, then the log-likelihood of $\mathcal{S}$ being the true simplex becomes $-\infty$. Moreover, among those simplices that contain all the data points, the one(s) with the smallest volume would be the maximizer(s) of (8). Regardless of the computational burden that is required to find the MLE, one can still be interested in a sufficient number of samples in $\boldsymbol{D}$ which guarantees the suitability of MLE in a PAC sense. In this regard, Theorem 1 states one of our main contributions in this paper, where we derive a new sample complexity bound for MLE as a function of both dimension K and estimation error ϵ.

Theorem 1 (Sample Complexity of MLE)**.** *Assume a K-simplex $\mathcal{S}_T \in \mathbb{S}_K$ and let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n i.i.d. samples drawn from $\mathbb{P}_{\mathcal{S}_T}$. Assume there exist $\epsilon, \zeta > 0$, such that*

$$n \geq O\left(\frac{1}{\epsilon}\left[K^2 \log\left(\frac{K}{\epsilon}\right) + \log\frac{1}{\zeta}\right]\right).$$

Then, with probability at least $1-\zeta$*, the maximum likelihood estimator of* $\mathcal{S}_T$*, denoted by* $\mathcal{S}^*_{\mathrm{ML}}$*, satisfies* $\mathcal{D}_{\mathrm{TV}}\left(\mathbb{P}_{\mathcal{S}_T}, \mathbb{P}_{\mathcal{S}^*_{\mathrm{ML}}}\right) \leq \epsilon$*.*

The proof of Theorem 1 is given in Appendix A. Sample complexity is shown to be $O\left(K^2 \log K\right)$ w.r.t. the dimension K, while the dependence on estimation error ϵ has remained optimal [12]. Interestingly, the given guarantees on the accuracy of MLE hold regardless of the shape of the simplex and does not impose any geometric constraints on $\mathcal{S}_T$.

However, solving for (8) is proved to be NP-hard, and thus impractical in real-world situations [8]. In the remainder of this section, we try to propose a heuristic framework for the estimation of simplices, which is practically amenable and at the same time enjoys from a number of weaker theoretical guarantees. A solution for the computational hardness of ML estimator is to replace the objective function in (8) with a continuously-relaxed surrogate. In order to do so, first it should be noted that the maximization in (8) is equivalent to the following constrained minimization problem:

$$\begin{aligned} \mathcal{S}^*_{\mathrm{ML}} \;=\; & \underset{\mathcal{S}\in\mathbb{S}_K}{\operatorname{argmin}} \quad \mathrm{Vol}\left(\mathcal{S}\right) \\ & \text{subject to} \quad d_{\mathcal{S}}\left(\boldsymbol{X}_i\right) = 0 \;,\; \forall i. \end{aligned} \tag{9}$$

The above formulation directly follows from the explanations that we gave earlier on how ML algorithm attempts to find the optimal simplex. It should be noted that any other distance measure rather than $d_{\mathcal{S}}\left(\cdot\right)$ can also be used in (9); However, this particular choice significantly helps us to derive our sample complexity bounds in the proceeding parts of the paper. Motivated by the formulation in (9) and also the idea of Lagrangian relaxation, let us propose the Continuously-Relaxed Risk (CRR) in order to approximate $\mathcal{S}_T$ as follows:

Definition 3 (Continuously-Relaxed Risk)**.** *Assume a dataset* $\boldsymbol{D} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ *in* $\mathbb{R}^K$*, parameter* $\gamma \geq 0$*, and an increasing and integrable function* $\ell : \mathbb{R} \to \mathbb{R}$*. Then, the empirical Continuously-Relaxed Risk* $\hat{R}_{\mathrm{CRR}} : \mathbb{S}_K \to \mathbb{R}$ *is defined as:*

$$\hat{R}_{\mathrm{CRR}}\left(\mathcal{S}; \boldsymbol{D}, \gamma, \ell\right) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell\left(d_{\mathcal{S}}\left(\boldsymbol{X}_i\right)\right) + \gamma \mathrm{Vol}\left(\mathcal{S}\right). \tag{10}$$

Also, let us define

$$\mathcal{S}^* = \mathcal{S}^*\left(\boldsymbol{D}, \gamma, \ell\right) \;\triangleq\; \underset{\mathcal{S}\in\mathbb{S}_K}{\operatorname{argmin}} \; \hat{R}_{\mathrm{CRR}}\left(\mathcal{S}; \boldsymbol{D}, \gamma, \ell\right), \tag{11}$$

as the Continuously-Relaxed Estimator (CRE) of $\mathcal{S}_T$*.*

Parameter $\gamma > 0$ and function $\ell\left(\cdot\right)$ can be freely chosen, except that $\ell\left(\cdot\right)$ must be increasing and integrable; However, ℓ does not need to be strictly increasing or even continuous in general. In any case, choosing a differentiable ℓ makes $\hat{R}_{\mathrm{CRR}}\left(\mathcal{S}\left(\boldsymbol{\Theta}\right); \boldsymbol{D}, \gamma, \ell\right)$ to be differentiable w.r.t. the vertex matrix $\boldsymbol{\Theta}$ almost everywhere[4]. In Section 3.1, we derive an efficient technique to numerically compute the derivatives of $\hat{R}_{\mathrm{CRR}}$ w.r.t. the vertex matrix $\boldsymbol{\Theta}$ which works for all differentiable functions $\ell\left(\cdot\right)$. We then propose a heuristic GD-based scheme to find a local minimizer of (11). The following theorem gives us a sample complexity bound for learning of S_T via (11). Our results hold for any increasing and integrable function $\ell : \mathbb{R} \to \mathbb{R}$.

Theorem 2 (Sample Complexity for General ℓ)**.** *Assume a* K*-simplex* $\mathcal{S}_T$ *with Lebesgue measure* $V_T \triangleq \mathrm{Vol}\left(\mathcal{S}_T\right)$*, which is* $\left(\underline{\lambda}, \bar{\lambda}\right)$*-isoperimetric for some* $\underline{\lambda}, \bar{\lambda} \geq 0$*. Also, assume* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ *to*

[4]As shown in Section 3.1, the subset of simplices in $\mathbb{S}_K$ for which the gradient of $\hat{R}_{\mathrm{CRR}}$ does not exist has a zero Lebesgue measure.

be n i.i.d. samples drawn from $\mathbb{P}_{\mathcal{S}_T}$. Assume for $\zeta, \epsilon > 0$, the following condition holds for n:

$$n \geq \left(\frac{6\ell\left(3\underline{\lambda} K V_T^{\frac{1}{K}}\right)\left(\sqrt{K^2 \log \frac{ne}{K}} + \sqrt{\log \frac{1}{\zeta}}\right) + \gamma V_T \epsilon}{\epsilon L\left(\frac{\epsilon V_T^{1/K}}{(K+1)\bar{\lambda}}\right)} \right)^2,$$

where $L(x) \triangleq \frac{1}{x}\int_0^x \ell(u)\,\mathrm{d}u - \ell(0)$. Then, with probability at least $1-\zeta$ we have $\mathcal{D}_{\mathrm{TV}}(\mathbb{P}_{\mathcal{S}_T}, \mathbb{P}_{\mathcal{S}^*}) \leq \epsilon$, where $\mathcal{S}^*$ is an optimizer of (11).

Proof of Theorem 2 is given in Appendix A. Similar to Theorem 1, the proof includes mathematical techniques from high-dimensional geometry and calculus, as well as tools from Vapnik-Chervonenkis (VC) theory of statistical learning [10]. The general statement of Theorem 2, for particular choices of the smoothing function ℓ and/or under certain asymptotic regimes can be extended to some interesting side results. For example, with appropriate choices of γ, one can always bound the total-variation error below $O\left(\sqrt{\log n/n}\right)$. Another interesting issue is the mathematical dependence of sample complexity on the geometric regularity of $\mathcal{S}_T$. In fact, searching for functions ℓ that decrease (or even obviate) the need to put geometric restrictions on the true simplex could be a nice research direction which goes beyond the scope of this paper. The following corollary is perhaps the most notable consequence of Theorem 2.

Corollary 1 (Sample Complexity of Soft-ML)**.** *Assume a K-simplex $\mathcal{S}_T \in \mathbb{S}_K$ and let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ to be n i.i.d. samples drawn from $\mathbb{P}_{\mathcal{S}_T}$. Also, assume $\mathcal{S}_T$ is $(\underline{\lambda}, \bar{\lambda})$-isoperimetric for some bounded $\underline{\lambda}, \bar{\lambda} > 0$. For $\epsilon, \zeta > 0$ and parameter $\gamma > 0$, let function $\ell : \mathbb{R} \to \mathbb{R}$ be*

$$\ell(u) \triangleq 1 - e^{-bu}, \quad \forall u \in \mathbb{R},$$

with $b \triangleq \frac{K}{\epsilon}$, and also assume

$$n \geq_{\gamma, \bar{\lambda}, \underline{\lambda}} O\left(\frac{1}{\epsilon^2}\left[K^2 \log\left(\frac{K}{\epsilon}\right) + \log\frac{1}{\zeta}\right]\right),$$

where $\geq_{\gamma, \bar{\lambda}, \underline{\lambda}}$ means the inequality holds up to a factor that only depends on the mentioned parameters. Then, with probability at least $1-\zeta$ the minimizer of (11), denoted by $\mathcal{S}^*$, satisfies the inequality $\mathcal{D}_{\mathrm{TV}}(\mathbb{P}_{\mathcal{S}_T}, \mathbb{P}_{\mathcal{S}^*}) \leq \epsilon$.

Proof of Corollary 1 is given in Appendix B. In Corollary 1, we show that for at least one class of differentiable functions ℓ, the sample complexity for acquiring a reliable solution to (11) has the same order with that of MLE in terms of parameter ζ and dimension K. However, the optimal dependence of sample complexity w.r.t. ϵ, which is $\tilde{O}\left(\epsilon^{-1}\right)$, has been deteriorated to $\tilde{O}\left(\epsilon^{-2}\right)$. We conjecture that this sub-optimal dependence cannot be improved, since prior works on learning of other geometric shapes, such as hyper-cubes in [9], have also concluded that by relaxing the combinatoric and exponential-time algorithms the dependence on ϵ would be sacrificed.

By using the Soft-ML setting of Corollary 1, the optimization problem of Definition 3 becomes differentiable w.r.t. $\boldsymbol{\Theta}$ almost everywhere and thus one can apply any heuristic continuous optimization tool, such as Gradient Descent (GD), to solve it. It should be noted that this result is not in contrast with the computational hardness of ML for two reasons: first, Corollary 1 does not say anything about the convexity of (10). Second, even if (10) happens to be convex with the above choice of $\ell(\cdot)$, we are still tackling a different problem compared to finding the *smallest inclusive simplex* which is what ML estimation attempts to do.

A thorough convexity analysis of our method goes beyond the scope of this paper. As a result, we make no claims regarding the computational complexity of our continuous-relaxation in Definition

Algorithm 1 Learning of simplices via Gradient Descent

procedure SIMPLEX INFERENCE($\boldsymbol{D} = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}, K, \ell(\cdot), \gamma, T, \alpha$)
  Select $\{i_0, i_1, \dots, i_K\} \subset [n]$ uniformly at random.
  Initialize $\boldsymbol{\Theta}^{(0)} = [\boldsymbol{X}_{i_0} | \dots | \boldsymbol{X}_{i_K}]$
  for $t = 0 : \cdots : T-1$ **do**
    $\boldsymbol{\Theta}^{(t+1)} \leftarrow \boldsymbol{\Theta}^{(t)} - \alpha \nabla_{\boldsymbol{\Theta}} \left[\hat{R}_{\text{CRR}}(\mathcal{S}(\boldsymbol{\Theta}); \boldsymbol{D}, \ell, \gamma) \right]$ ▷ According to (16) and Theorem 3
  end for
end procedure

3. Instead, we derive an efficient scheme to compute the gradient of (11) w.r.t. almost any vertex matrix $\boldsymbol{\Theta}$, and then show that a simple heuristic approach, such as GD, can achieve a very good performance in practice.

3.1 Numerical Optimization

We show how to numerically compute the gradient of (10) in order to employ it in the GD-based optimization of Algorithm 1. Later in Section 4, we have empirically investigated the convergence and effectiveness of Algorithm 1 through several computer simulations.

According to (10), $\hat{R}_{\text{CRR}}(\mathcal{S})$ is naturally broken into two separate terms: one controls the distance of data points from $\mathcal{S}$, and the second one is a regularization term on the volume of $\mathcal{S}$. Gradient of $\gamma \text{Vol}(\mathcal{S})$ is easy to compute, since volume can be rewritten as

$$\text{Vol}(\mathcal{S}(\boldsymbol{\Theta})) = \frac{1}{K!} \left| \det \left[\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0 \middle| \cdots \middle| \boldsymbol{\theta}_K - \boldsymbol{\theta}_0 \right] \right|. \tag{12}$$

On the other hand, for any matrix $\boldsymbol{A} \in \mathrm{R}^{K \times K}$ we have $\nabla_{\boldsymbol{A}} \det(\boldsymbol{A}) = \text{adj}^T(\boldsymbol{A})$, where $\text{adj}(\cdot)$ denotes the *adjugate* function. To ensure that the gradient is well-defined, assume $\boldsymbol{\Theta}$ to be non-degenerate, i.e. $\mathcal{S}(\boldsymbol{\Theta})$ has a positive volume. This assumption only excludes a zero-measure (according to the Lebesgue measure) subset of solutions from $\mathbb{S}_K$. For the remaining part of (10), one needs to compute the gradient vectors $\boldsymbol{g}_{ik} = \nabla_{\boldsymbol{\theta}_k} \ell\left(d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X}_i)\right),\ i \in [n],\ k \in [K]$, or alternatively, the gradient matrices $\boldsymbol{G}_i \triangleq [\boldsymbol{g}_{i0} | \cdots | \boldsymbol{g}_{iK}]\ ,\ i \in [n]$. The following theorem gives a straightforward procedure for numerical computation of $\boldsymbol{G}_i$s.

Theorem 3 (Gradient of the Planar Distance)**.** *Assume $\boldsymbol{X} \in \mathbb{R}^K$, $\mathcal{S}(\boldsymbol{\Theta}) \in \mathbb{S}_K$ with $\text{Vol}(\mathcal{S}(\boldsymbol{\Theta})) > 0$, and an increasing function $\ell : \mathbb{R} \to \mathbb{R}$ which is differentiable in $\mathbb{R}^+$.*

i) If $d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X}) = 0$, assume $\boldsymbol{w}_k^T \boldsymbol{X} + b_k < 0$ for all $k = 0, \dots, K$, where $(\boldsymbol{w}_k, b_k)$ represent the parameters of hyper-plane $\mathcal{H}_k$ which encompasses the kth facet of $\mathcal{S}(\boldsymbol{\Theta})$. Then, the gradient w.r.t. $\boldsymbol{\Theta}$ is zero.

ii) If $d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X}) > 0$, assume $k^* \triangleq \text{argmax}_k\ \boldsymbol{w}_k^T \boldsymbol{X} + b_k$ is unique. Next, one needs to compute the normal outward vector $\boldsymbol{w}_{k^*}$. Let the $K \times K$ matrix $\boldsymbol{\Theta}_{-k^*}$ represent $\boldsymbol{\Theta}$ with its k^*th column removed. Then, $\boldsymbol{w}_{k^*}$ is the eigenvector corresponding to the unique zero eigenvalue of $\left(\boldsymbol{I} - \frac{1}{K}\boldsymbol{1}\boldsymbol{1}^T\right)\boldsymbol{\Theta}_{-k^*}^T$. Without loss of generality, assume $\|\boldsymbol{w}_{k^*}\|_2 = 1$ and $\boldsymbol{w}_{k^*}^T\left(\boldsymbol{\theta}_{k^*} - \frac{1}{K}\boldsymbol{\Theta}_{-k^*}\boldsymbol{1}\right) \leq 0$ which means $\boldsymbol{w}_{k^*}$ is outward w.r.t. $\mathcal{S}(\boldsymbol{\Theta})$. Let

$$\boldsymbol{p}^* \triangleq \boldsymbol{\Theta}_{-k^*}^{\dagger} \left(\boldsymbol{X} - \boldsymbol{w}_{k^*} \boldsymbol{w}_{k^*}^T \left(\boldsymbol{X} - \frac{1}{K} \boldsymbol{\Theta}_{-k^*} \boldsymbol{1} \right) \right) \in \mathbb{R}^K, \tag{13}$$

where † denotes the pseudo-inverse operator. Then, we have:

$$\nabla_{\boldsymbol{\Theta}} \ell \left(d_{\mathcal{S}(\boldsymbol{\Theta})} \left(\boldsymbol{X} \right) \right) = -\ell' \left(d_{\mathcal{S}(\boldsymbol{\Theta})} \left(\boldsymbol{X} \right) \right) \boldsymbol{w}_{k^*} \boldsymbol{p}^{*T} \boldsymbol{J}_{k^*} \ \in \ \mathbb{R}^{K \times (K+1)}, \tag{14}$$

where $\ell' (\cdot)$ denotes the derivative of $\ell (\cdot)$, and $\boldsymbol{J}_{k^*}$ is a $K \times (K+1)$ zero-padded identity matrix as follows:

$$\boldsymbol{J}_{k^*} \triangleq \left[\boldsymbol{e}_1 | \cdots | \boldsymbol{e}_{k^*-1} | \boldsymbol{0} | \boldsymbol{e}_{k^*} | \cdots | \boldsymbol{e}_K \right]. \tag{15}$$

Here, $\boldsymbol{e}_k \in \mathbb{R}^K$ refers to a binary vector where the kth component is 1 and the rest are zero.

Proof of Theorem 3 is given in Appendix A. It should be noted that conditions i) and ii) of Theorem 3 hold for all $\mathcal{S} \in \mathbb{S}_K$, except for a subset with a zero Lebesgue measure. Total gradient of the continuously-relaxed risk $\hat{R}_{\text{CRR}} \left(\mathcal{S} \left(\boldsymbol{\Theta} \right) ; \boldsymbol{D}, \ell, \gamma \right)$ w.r.t. vertex matrix $\boldsymbol{\Theta}$ can then be computed as

$$\nabla_{\boldsymbol{\Theta}} \hat{R}_{\text{CRR}} \left(\mathcal{S} \left(\boldsymbol{\Theta} \right) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{G}_i + \frac{\gamma s}{K!} \left[\boldsymbol{0} \middle| \text{adj}^T \left(\boldsymbol{\Theta}_{1:K} - \boldsymbol{\theta}_0 \boldsymbol{1}^T \right) \right] \left(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}^T}{K+1} \right), \tag{16}$$

where $\boldsymbol{0}$ is an all-zero vector, $\boldsymbol{1}$ is an all-one vector and $\boldsymbol{I}$ denotes the identity matrix, all with corresponding appropriate sizes. Variable $s \in \{\pm 1\}$ denotes the sign of $\det \left[\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0 | \cdots | \boldsymbol{\theta}_K - \boldsymbol{\theta}_0 \right]$, and $\boldsymbol{\Theta}_{1:K}$ represents the $K \times K$ matrix $\left[\boldsymbol{\theta}_1 | \cdots | \boldsymbol{\theta}_K \right]$. Also, we have $\boldsymbol{G}_i \triangleq \nabla_{\boldsymbol{\Theta}} \ell \left(d_{\mathcal{S}(\boldsymbol{\Theta})} \left(\boldsymbol{X}_i \right) \right)$. In order to avoid the zero-measure set of ill-conditioned simplices in $\mathbb{S}_K$, as described by conditions i and ii of Theorem 3, one can add an infinitesimally small noise to $\boldsymbol{\Theta}$ before computing the gradient matrix at each iteration. For example, assuming the numerical computations are performed by a digital processor with B precision bits, the addition of a Gaussian noise drawn from $\mathcal{N} \left(\boldsymbol{0}, 2^{-B} \boldsymbol{I} \right)$ implies that the probability of facing an ill-conditioned simplex in Algorithm 1 is zero.

4 Experimental Results

This section is devoted to testing the performance of Algorithm 1 on a number of synthetic and real-world datasets. In particular, we report the results of applying our method on three different tasks: 1) analysis of convergence, noise robustness and error performance on synthetic data, 2) cell-type extraction in computational biology, and 3) spectral unmixing of remote sensing data. We also compare our method with a number of popular rival frameworks. Throughout this section, *estimation error* is measured in terms of the average Euclidean distance between the vertices of the true simplex $\mathcal{S}_T$ and those of an estimated simplex $\hat{\boldsymbol{S}}$ [5]. For simplicity, assume $\boldsymbol{\Theta} = \left[\boldsymbol{\theta}_0 | \cdots | \boldsymbol{\theta}_K \right]$ and $\hat{\boldsymbol{\Theta}} = \left[\hat{\boldsymbol{\theta}}_0 | \cdots | \hat{\boldsymbol{\theta}}_K \right]$ to be the vertex matrices of $\mathcal{S}_T$ and $\hat{\mathcal{S}}$, respectively. Then, *error* can be mathematically formulated as follows:

$$error^2 \triangleq \min_{(i_0, \ldots, i_K)} \frac{1}{K \left(K+1 \right)} \sum_{k=0}^{K} \left\| \boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_{i_k} \right\|_2^2, \tag{17}$$

where minimization is taken over all permutations $(i_0, \ldots, i_K)$ of the numbers $\{0, 1, \ldots, K\}$. The main reason for using *error* instead of the total-variation distance $\mathcal{D}_{\text{TV}} (\cdot, \cdot)$ is due to the numerical hardness of computing TV distance, specially in higher dimensions. Motivated by Corollary 1, we have chosen $\ell (u) \triangleq 1 - e^{-bu}$ as our smoothing loss function in Definition 3, where $b \triangleq \text{diam}^{-1} (\boldsymbol{D})$ is chosen as the inverse diameter of the input database $\boldsymbol{D} = \{ \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \}$. Here, by $\text{diam} (\boldsymbol{D})$ we simply mean $\max_{i,j} \| \boldsymbol{X}_i - \boldsymbol{X}_j \|_2$. Parameters γ and the learning rate α in Algorithm 1 are manually adjusted to optimize the performance and convergence rate. Noticeably, we did not observe a considerable sensitivity to any of these parameters.

[5] We use the notation $\hat{\mathcal{S}}$ for the output of Algorithm 1 instead of $\mathcal{S}^*$, since $\mathcal{S}^*$ is already defined as the global optimizer of (11). However, Algorithm 1 may converge to a local minimizer which could be different from $\mathcal{S}^*$.

4.1 Initialization and Acceleration

We have empirically observed that by setting an appropriate initial simplex $\mathcal{S}_{\text{init}} \in \mathbb{S}_K$ for Algorithm 1, one can significantly improve its convergence rate. Moreover, numerical evaluation of the gradient in (16) is $O\left(nK^3\right)$ which could be computationally intensive in some applications. In this part, we propose a simple heuristic approach to circumvent the above-mentioned problems. Let $\text{conv}\left(\boldsymbol{D}\right) \subseteq \boldsymbol{D}$ represent the vertices of the convex hull of data points $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \boldsymbol{D}$. In other words, $\text{conv}\left(\boldsymbol{D}\right)$ is the smallest subset of $\boldsymbol{D}$ with the following property:

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \left\{ \sum_i q_i \boldsymbol{Z}_i \middle| \ \forall \boldsymbol{Z}_i \in \text{conv}\left(\boldsymbol{D}\right), \ \sum_i q_i = 1, \ \forall q_i \geq 0 \right\}. \tag{18}$$

This way, one can choose $\mathcal{S}_{\text{init}} \in \mathbb{S}_K$ such that its vertices are randomly selected from $\text{conv}\left(\boldsymbol{D}\right)$. We have empirically observed that such initialization leads to a substantially faster convergence. Moreover, one can accelerate the execution of Algorithm 1 by approximating the gradient in (16) via computing it only for data points in $\text{conv}\left(\boldsymbol{D}\right)$ instead of the whole dataset $\boldsymbol{D}$.

The effect of the proposed approximation for the gradient remains negligible on the performance of Algorithm 1 in low noise regimes. In other words, when γ is chosen to be relatively low, which means the objective function of Algorithm 1 becomes more similar to that of ML estimator, the proposed algorithm tries to find a simplex that encompasses almost all of the data points. This is, however, equivalent to encompassing the convex hull of data points which suggests that the proposed approximation of gradient becomes significantly close to (16) after a number of initial iterations. We already know that the convex hull of a set of n points can be determined in $g_{n,K} \leq \tilde{O}\left(n^{\lfloor K/2 \rfloor}\right)$ arithmetic operations, which grows exponentially w.r.t. K [31]. However, the mentioned bound corresponds to a worst-case analysis which might not be the case in many real-world or randomly formed databases. As a result, we have faced a far smaller computational burden in our experiments, i.e. $g_{n,K} \ll \tilde{O}\left(n^{\lfloor K/2 \rfloor}\right)$ even for moderate K. A detailed experimental or analytic investigation of this issue goes beyond the scope of this paper.

Therefore, the proposed strategy reduces the expected computational complexity of approximating the gradient to $O\left(g_{n,K} + f_{n,K} K^3\right)$, where $f_{n,K}$ denotes the expected number of vertices of the convex hull of n randomly chosen points inside a K-simplex. It can be at least empirically confirmed that for large n, we have $f_{n,K} \ll n$. As a result of the mentioned initialization and acceleration strategies for moderately small K, our method runs fairly faster than its rival frameworks with a negligible harm to its performance. Detailed experiments on the performance of Algorithm 1 are given in the next sub-section. In all the following simulations we have always used the discussed initialization technique unless we say otherwise.

One more issue that should be discussed before proceeding to the next sub-section, is that finding the convex hull of a dataset $\boldsymbol{D}$ might be very sensitive to the presence of outliers in the dataset. However, analysis of the effect of outliers goes beyond the scope of this paper and we leave this interesting line of work for future researches in this area.

4.2 Synthetic Data

Figure 3 shows four snapshots from the outputs of Algorithm 1 running on a set of $n = 100$ noiseless data points in $\mathbb{R}^2$, which have been uniformly sampled from a typical two-dimensional simplex. We have chosen $K = 2$ in order to facilitate the visualization of our results. For this particular simulation, we have chosen a completely random initial simplex instead of the suggested

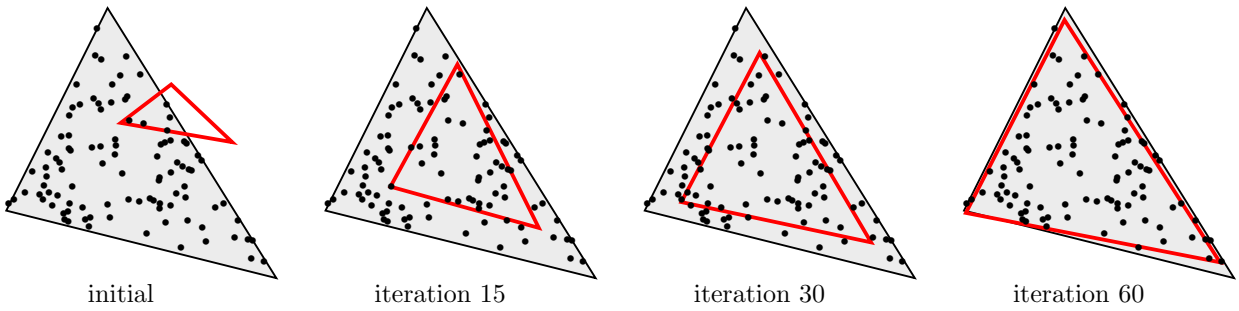


Figure 3: Snapshots from running Algorithm 1 on a set of $n = 100$ noiseless samples drawn uniformly from a two-dimensional simplex. The original triangle is drawn in black and the outputs of the proposed method for four different iteration steps are shown in red.

technique in the previous sub-section. The reason behind this choice is to show the convergence performance of Algorithm 1 and the fact that it does not necessarily need a good initialization. Here, the shaded simplex denotes the true $\mathcal{S}_T$, while the "red" simplex is the output of Algorithm 1 for four different iteration numbers. As it can be seen, despite the fact that algorithm has started from an inappropriate initial point, it converges to an acceptable solution in a reasonable number of iterations. Figure 4 depicts *error* as a function of iteration number for the experiment of Figure 3. Results have been averaged over 100 trials. According to the curve, sample mean and the standard deviation of *error* decay as the number of iterations increases.

Figure 5 aims to investigate the robustness of our method to noise and also the dimension of the space. We have used synthetic datasets for this purpose. In particular, Figure 5a illustrates the performance of Algorithm 1 when samples are contaminated with additive white Gaussian noise. In this regard, $n = 100$ uniformly-sampled data points have been drawn from a two-dimensional simplex. Each sample has been contaminated with a noise vector drawn independently from a zero-mean Gaussian distribution $\mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$, where σ denotes the standard deviation of the noise components. Let us define the *noise strength* ρ as the normalized standard deviation of noise as follows:

$$\rho \triangleq \sigma \left[\frac{1}{K\left(K+1\right)} \sum_{k,k'} \left\| \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'} \right\|_2 \right]^{-1} . \tag{19}$$

According to the above definition, value of ρ is invariant to the size of the original simplex $\mathcal{S}_T$. Figure 5a depicts the *error* of Algorithm 1 as a function of noise strength ρ. The proposed method

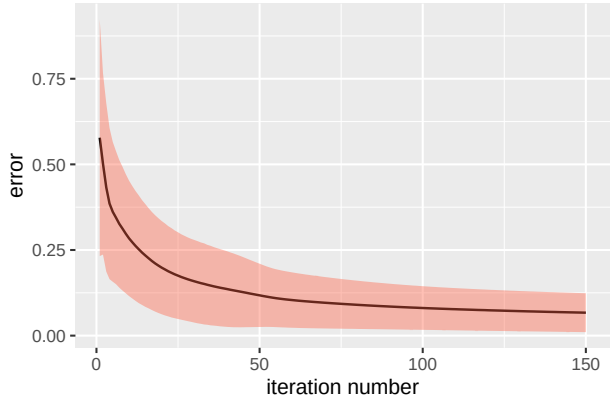
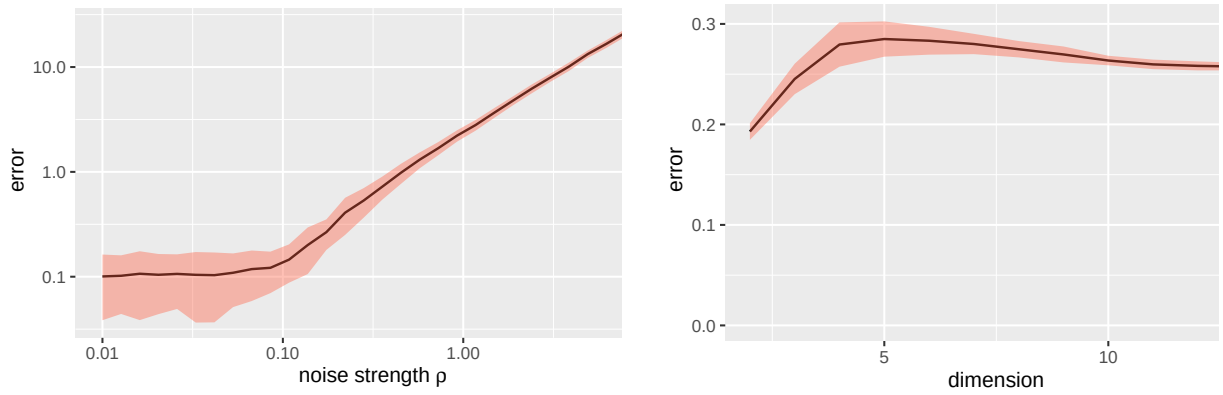


Figure 4: Depiction of *error* in (17) as a function of iteration number for Algorithm 1. The experiment has been performed on $n = 100$ data points uniformly sampled from a two-dimensional simplex. Parameters γ and optimization step α have been adjusted to optimize the performance. According to the curve, sample mean and the standard deviation of *error* decay as the number of iterations is increased.



(a) (b)

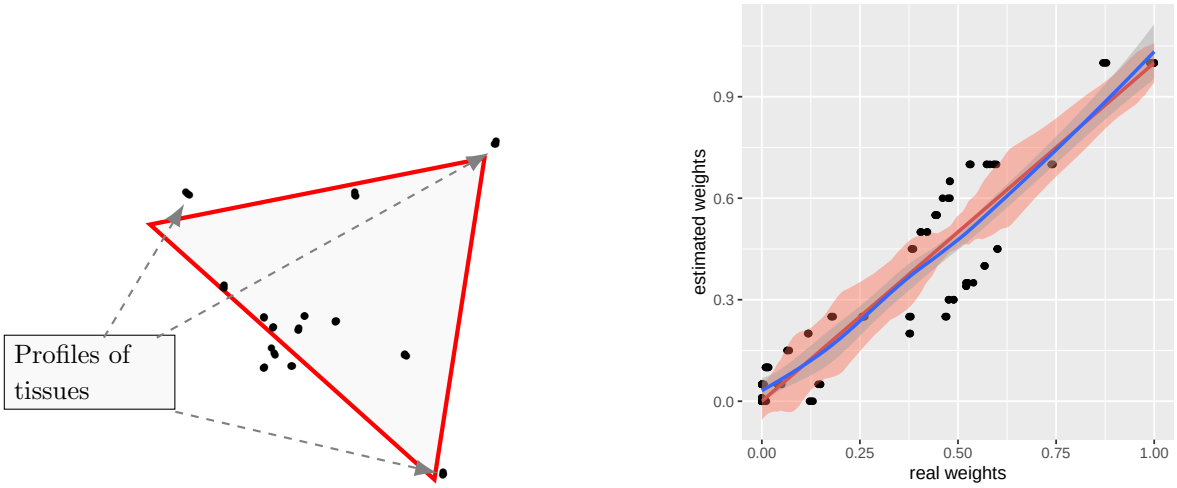
Figure 5: Estimation *error* as a function of noise strength ρ and dimension. In 5a, $n = 100$ data samples are drawn from a two-dimensional simplex and then contaminated with additive white Gaussian noise. However, for 5b data samples are noiseless and n has been increased proportional to $K^2 \log K$, where K indicates the dimension.

is considerably robust to noise roughly for $\rho \leq 0.2$. However, the performance is degraded when noise strength is being increased. On the other hand, Figure 5b illustrates the *error* as a function of dimension K. For each point of the curve, dataset size has been increased proportional to $K^2 \log K$ in order to match the result of Theorem 2 and Corollaries 1 and 1. Surprisingly, we have empirically validated that as long as n is increased proportional to $K^2 \log K$, the *error* remains constant. This implies that Algorithm 1 has the same sample complexity (order-wise) as that of ML, while unlike the ML estimator it converges to a solution in a practically reasonable time.

In Table 1, we have compared the *error* performance of our method with a number of well-known computational techniques for learning of simplices, namely MVSA [1], SISAL [32], VCA [33], and UNMIX [3]. We have used three datasets for the experiments which are i) Plain dataset: $n = 100$ data points generated from a simple two-dimensional simplex, ii) Noisy dataset: a noisy version of the "Plain" dataset with $\rho = 0.5$ (highly noisy), and iii) High-Dimensional (HD) dataset: $n = 1000$ data points generated uniformly from a noiseless simplex with $K = 9$. According to Table 1, our method has a comparable performance on "Plain" and "HD" datasets, while it outperforms all the rival strategies on Noisy dataset. The reason behind this result might be due to the continuously-relaxed nature of our method which is in contrast to the majority of existing frameworks in this area. In fact, the inherent flexibility of the objective function in (11) achieves more robustness to noisy data points that fall out of the true simplex.

	Plain	Noisy	HD
Proposed	0.20	**0.51**	0.74
MVSA	0.14	1.84	0.76
SISAL	0.16	1.65	0.77
VCA	1.09	1.006	5.93
UNMIX	0.14	1.83	-

Table 1: Comparison of the proposed method with MVSA [1], SISAL [32], VCA [33], and UNMIX [3]. Methods have been tested on three different datasets. The values of *error* have been averaged over several runs, such that all relative standard deviations become less than 10%. UNMIX did not execute on "HD" dataset in a reasonable time.



(a) (b)

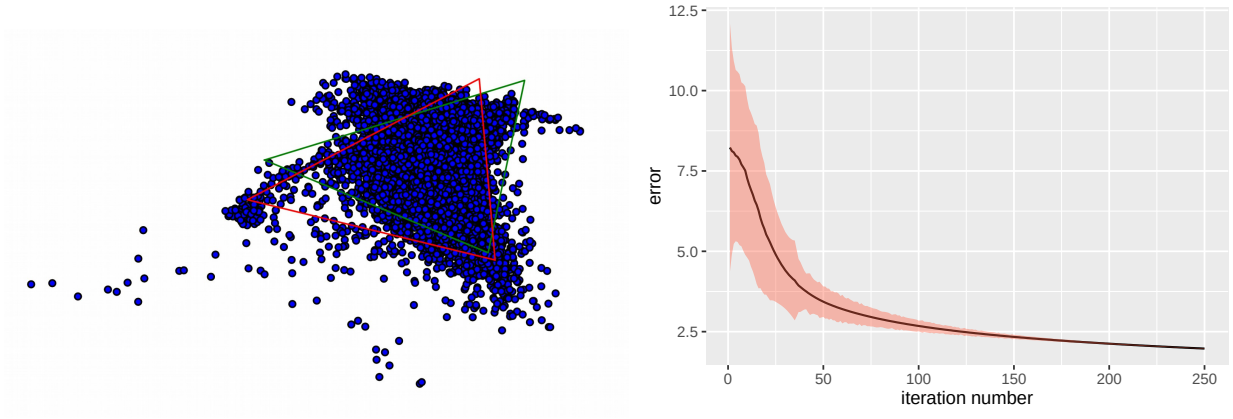
Figure 6: Cell-type identification from micro-array data given in [34]. 6a: Visualization of data points, as well as the estimated simplex. Vertices of the estimated simplex highly resemble the expression levels of the ground truth tissues. 6b: Estimated weights for the samples as a function of real weights reported in the dataset. Data points are scattered around the $X = Y$ curve (red). Also, the result of a LOESS regression of the samples (blue) falls very close to the $X = Y$ curve.

4.3 Computational Biology

Unmixing of cell-types from biological samples has become a commonplace task in bioinformatics, and in particular, in cancer therapy. In a recent work [34], researchers have synthetically merged cells from three different tissues (brain, liver and kidney) with several different percentages, where combination weights were known prior to each merging. The aim of this experiment is to investigate whether one can deconvolve the gene-expression profiles of different tissues from a synthetic bulk dataset. However, the dataset in [34] is also a potentially appropriate target for Algorithm 1. In fact, we can neglect the knowledge of the combination weights, and try to blindly infer both the unknown gene-expression profiles of the three tissues, as well as the combination weights, directly from the gene-expression levels of samples.

In this regard, we have used all the $n = 42$ samples of the dataset in [34], where the initial dimensionality of data is $K_{\text{init}} = 31100$. Due to the presence of redundant information in the dataset[6], samples have been linearly projected onto a corresponding 41-dimensional subspace via Principal Component Analysis (PCA), and then the proposed method has been applied. Figure 6 illustrates the final results, where Figure 6a shows a qualitative two-dimensional visualization of the final simplex (via the first two principal components) that matches the ground truth gene-expression profiles of the tissues with a relatively high accuracy. From a more quantitative perspective, Figure 6b shows the inferred weights for all the $n = 42$ samples as a function of real weights which are already given in the dataset. As can be seen, points are scattered around the $X = Y$ curve which is depicted in red. We have also shown the result of a LOESS (Locally Estimated Scatterplot Smoothing) regression of the samples via a blue curve, which is very close to the $X = Y$ curve. This suggests that the real weights and the ones that are estimated via our proposed method are

[6] Any n points in $\mathbb{R}^{K_{\text{init}}}$, for $n \leq K_{\text{init}}$, lie on a $(n-1)$-dimensional linear subspace which is almost surely identifiable and can be attained by Principal Component Analysis. This way, the reduction of dimensionality in this case does not result into any information loss.



(a) (b)

Figure 7: (a): Estimation of dominant elements in a hyper-spectral imaging dataset (red simplex) compared to the ground truth (green simplex). Three major elements, called *Andradite*, *Dumortierite* and *Sphene* have been identified which considerably match with our experimental findings. Data points have been visualized via the first two PCs. (b): Estimation error as a function of iterations for the *Cuprit* dataset.

similar to each other.

4.4 Hyper-Spectral Remote Sensing

Another major application of the "unmixing problem" is in hyper-spectral remote sensing, where one aims to reveal the chemical composition of materials in a remote area by using hyper-spectral imaging devices. Each pixel in a hyper-spectral image is a vector including radiation intensities (in different wave-lengths) from a relatively small region of a remote location. This way, pixels can be considered as weighted averages of a limited number of radiation profiles which correspond to the chemical elements that are likely to be found in that location. The *Cuprit* dataset [35] includes $n = 47500$ pixels, where each pixel has been measured in effectively $m = 188$ spectral wavelengths. In [35], it has been suggested that $K = 12$ basic elements might be present in the remote location that corresponds to the image pixels, however, a smaller subset of them might be dominant.

We have randomly chosen $n = 10000$ points from this dataset and reduced the dimensionality of data to $\dim = 10$ via PCA. The latter is performed in order to eliminate the majority of linear dependencies and also to suppress the effect of noise. According to a simple eigen-analysis of the data, we chose the number of vertices for the simplex as $K + 1 = 3$. Based on the above setting, Figure 7 illustrates the performance of the proposed method on the data compared to the ground truth profiles presented in the literature. Similar to the previous sub-section, Figure 7a depicts a two-dimensional visualization, obtained from the first two principal components of the data samples and the estimated simplex via Algorithm 1 (shown in red). For the sake of comparison, the ground truth profiles have been shown by a green simplex. Evidently, our method finds the dominant elements with an acceptable accuracy. Figure 7b shows the approximation *error* versus the number of iterations for Algorithm 1.

5 Conclusions

This paper aims to develop a computational framework for learning of simplices in arbitrary dimensions. In order to deal with the combinatoric structure of this problem, a continuously-relaxed optimization scheme has been proposed which comes with new theoretical sample complexity bounds. In particular, we have shown that a sufficient sample complexity for both ML and the proposed Soft-ML surrogate are $n \geq O\left(K^2 \log K\right)$, which is a significant improvement of the existing bounds for this problem, i.e. $O\left(K^{22}\right)$. In addition, a heuristic, yet computationally tractable algorithm is proposed for numerical optimization of our scheme. Numerous experiments magnify the applicability of our method on synthetic and real-world datasets. Our method has shown a comparable performance to a number of well-known rival strategies on noiseless data, while shows a considerable superiority in noisy regimes. For future works, one may attempt to provide similar sample complexity bounds for the noisy case, which has not been tackled yet. Assuming non-uniform distributions over the simplices to generate the data could be another interesting problem to consider. Also, deriving the computational complexity of this problem, as well as finding lower bounds on the sample complexity, are examples of other existing research directions in this area.

References

[1] J. Li, A. Agathos, D. Zaharie, J. M. Bioucas-Dias, A. Plaza, and X. Li, "Minimum volume simplex analysis: A fast algorithm for linear hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 9, pp. 5067–5082, 2015.

[2] A. Ambikapathi, T.-H. Chan, W.-K. Ma, and C.-Y. Chi, "Chance-constrained robust minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4194–4209, 2011.

[3] D. Tolliver, C. Tsourakakis, A. Subramanian, S. Shackney, and R. Schwartz, "Robust unmixing of tumor states in array comparative genomic hybridization data," *Bioinformatics*, vol. 26, no. 12, pp. i106–i114, 2010.

[4] G. Satas and B. J. Raphael, "Tumor phylogeny inference using tree-constrained importance sampling," *Bioinformatics*, vol. 33, no. 14, pp. i152–i160, 2017.

[5] N. S. Zuckerman, Y. Noam, A. J. Goldsmith, and P. P. Lee, "A self-directed method for cell-type identification and separation of gene expression microarrays," *PLoS computational biology*, vol. 9, no. 8, p. e1003189, 2013.

[6] V.-E. Brunel *et al.*, "Adaptive estimation of convex polytopes and convex sets from noisy data," *Electronic Journal of Statistics*, vol. 7, pp. 1301–1327, 2013.

[7] J. Anderson, N. Goyal, and L. Rademacher, "Efficient learning of simplices," in *Conference on Learning Theory*, 2013, pp. 1020–1045.

[8] A. Packer, "Np-hardness of largest contained and smallest containing simplices for v-and h-polytopes," *Discrete and Computational Geometry*, vol. 28, no. 3, pp. 349–377, 2002.

[9] A. Frieze, M. Jerrum, and R. Kannan, "Learning linear transformations," in *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on.* IEEE, 1996, pp. 359–368.

[10] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[11] V.-E. Brunel, "Adaptive estimation of convex and polytopal density support," *Probability Theory and Related Fields*, vol. 164, no. 1-2, pp. 1–16, 2016.

[12] V.-E. Brunel *et al.*, "Methods for estimation of convex sets," *Statistical Science*, vol. 33, no. 4, pp. 615–632, 2018.

[13] H. Groemer, "On the mean value of the volume of a random polytope in a convex set," *Archiv der Mathematik*, vol. 25, no. 1, pp. 86–90, 1974.

[14] I. Bárány and C. Buchta, "Random polytopes in a convex polytope, independence of shape, and concentration of vertices," *Mathematische Annalen*, vol. 297, no. 1, pp. 467–497, 1993.

[15] I. Bárány and D. G. Larman, "Convex bodies, economic cap coverings, random polytopes," *Mathematika*, vol. 35, no. 2, pp. 274–291, 1988.

[16] M. Ang and N. Gillis, "Algorithms and comparisons of non-negative matrix factorization with volume regularization for hyperspectral unmixing," *arXiv preprint arXiv:1903.04362*, 2019.

[17] R. Wu, Q. Li, and W.-K. Ma, "Stochastic ml simplex-structured matrix factorization under the dirichlet mixture model," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5561–5565.

[18] T. Li and C.-c. Ding, "Nonnegative matrix factorizations for clustering: A survey," in *Data Clustering*. Chapman and Hall/CRC, 2018, pp. 149–176.

[19] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 328–332, 2018.

[20] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications." *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, 2019.

[21] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3469–3486, 2015.

[22] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4418–4432, 2009.

[23] P. Q. Nguyen and O. Regev, "Learning a parallelepiped: Cryptanalysis of ggh and ntru signatures," *Journal of Cryptology*, vol. 22, no. 2, pp. 139–160, 2009.

[24] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[25] S. Zhang, A. Agathos, and J. Li, "Robust minimum volume simplex analysis for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6431–6439, 2017.

[26] A. Agathos, J. Li, J. M. Bioucas-Dias, and A. Plaza, "Robust minimum volume simplex analysis for hyperspectral unmixing," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 1582–1586.

[27] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[28] C.-H. Lin, A. Ambikapathi, W.-C. Li, and C.-Y. Chi, "On the endmember identifiability of craig's criterion for hyperspectral unmixing: A statistical analysis for three-source case," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2139–2143.

[29] R. Schwartz and S. E. Shackney, "Applying unmixing to gene expression data for tumor phylogeny inference," *BMC bioinformatics*, vol. 11, no. 1, p. 42, 2010.

[30] R. Osserman, "The isoperimetric inequality," *Bulletin of the American Mathematical Society*, vol. 84, no. 6, pp. 1182–1238, 1978.

[31] B. Chazelle, "An optimal convex hull algorithm in any fixed dimension," *Discrete & Computational Geometry*, vol. 10, no. 4, pp. 377–409, 1993.

[32] J. M. Bioucas-Dias, "A variable splitting augmented lagrangian approach to linear spectral unmixing," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS'09. First Workshop on*. IEEE, 2009, pp. 1–4.

[33] J. M. Nascimento and J. M. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.

[34] S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis, and A. J. Butte, "Cell type–specific gene expression differences in complex tissues," *Nature methods*, vol. 7, no. 4, p. 287, 2010.

[35] F. Zhu, Y. Wang, B. Fan, G. Meng, S. Xiang, and C. Pan, "Spectral unmixing via data-guided sparsity," *CoRR*, vol. abs/1403.3155, 2014.

[36] V. Vapnik, "Estimation of dependencies based on empirical data springer," *Information and Control*, 1982.

[37] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the vapnik-chervonenkis dimension," *Journal of the ACM (JACM)*, vol. 36, no. 4, pp. 929–965, 1989.

[38] S. Hanneke, "The optimal sample complexity of pac learning," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1319–1333, 2016.

A Proof of Theorems

Proof of Theorem 1. We show that the problem of estimating the true simplex $\mathcal{S}_T$ up to a total-variation error of ϵ can be transformed into a binary classification task, where one aims to learn a

classifier up to a same classification error rate. Mathematically speaking, for $\mathcal{S} \in \mathbb{S}_K$ and $\boldsymbol{X} \in \mathbb{R}^K$, let us define the parametric binary classifier $c(\cdot;\mathcal{S})$ as

$$c(\boldsymbol{X};\mathcal{S}) \triangleq \begin{cases} 0 & \boldsymbol{X} \in \mathcal{S} \\ 1 & \text{O.W.} \end{cases}, \tag{20}$$

and the corresponding classifier family $\mathscr{C}$ as

$$\mathscr{C} \triangleq \left\{ c(\cdot;\mathcal{S}) : \mathbb{R}^K \to \{1,0\} \;\middle|\; \mathcal{S} \in \mathbb{S}_K \right\}. \tag{21}$$

Having the input dataset $\boldsymbol{D} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ drawn independently from distribution $\mathbb{P}_{\mathcal{S}_T}$, let us consider an augmented supervised dataset $\tilde{\boldsymbol{D}}$ which includes the n data points $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, all labeled with $c(\cdot;\mathcal{S}_T)$; Thus, all data points in $\tilde{\boldsymbol{D}}$ are labeled with 0. Assume there exists algorithm $\mathscr{A}$, which given the dataset $\tilde{\boldsymbol{D}}$, is able to output a *consistent* classifier $c^* \in \mathscr{C}$ and a corresponding simplex $\mathcal{S}^* \in \mathbb{S}_K$, i.e., $c^*(\cdot) = c(\cdot;\mathcal{S}^*)$, where the volume of $\mathcal{S}^*$ is strictly bounded. Here, by consistent, we mean c^* classifies all the data points in $\tilde{\boldsymbol{D}}$ correctly. Mathematically speaking, let us assume c^* and $\mathcal{S}^*$ have the following two properties:

$$\begin{aligned} &\text{Consistency :} && c^*(\boldsymbol{X}_i) = 0, \text{ for all } i = 1, \ldots, n, \\ &\text{Bounded Volume Property :} && \text{Vol}(\mathcal{S}^*) \leq \text{Vol}(\mathcal{S}_T), \end{aligned}$$

where $\mathcal{S}_T$ denotes the true simplex underlying the input data samples. We first claim that the simplex $\mathcal{S}^*$ which is found by algorithm $\mathscr{A}$ and has the above properties approximates the true simplex $\mathcal{S}_T$ up to a tolerable total-variation error. Then, we show that the ML algorithm has both of the above-mentioned properties and thus MLE is a desired estimator in the total-variation (Nikodym distance) sense. In order to do so, first we show the following:

Lemma 1. *For any $\epsilon \in [0,1]$, if we have $\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}_{\mathcal{S}_T}}\{c^*(\boldsymbol{X})\} \leq \epsilon$, then $\mathcal{D}_{\text{TV}}(\mathbb{P}_{\mathcal{S}_T}, \mathbb{P}_{\mathcal{S}^*}) \leq \epsilon$.*

Proof. It should be noted that

$$\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}_{\mathcal{S}_T}}\{c(\boldsymbol{X};\mathcal{S}^*)\} = \mathbb{P}\{\boldsymbol{X} \sim \mathbb{P}_{\mathcal{S}_T} \text{ and } \boldsymbol{X} \in \mathcal{S}^*\} \leq \epsilon, \tag{22}$$

which means $\text{Vol}(\mathcal{S}^* \cap \mathcal{S}_T) \geq V_T(1-\epsilon)$, where $V_T = \text{Vol}(\mathcal{S}_T)$. Also, we already know that $\text{Vol}(\mathcal{S}^*) \leq V_T$. These facts lead us to the following three inequalities:

$$\begin{aligned} \text{Vol}(\mathcal{S}^*) &\geq V_T(1-\epsilon), \\ \text{Vol}(\mathcal{S}^* - \mathcal{S}_T) &\leq \text{Vol}(\mathcal{S}^*) - V_T(1-\epsilon) \\ \text{and } \text{Vol}(\mathcal{S}_T - \mathcal{S}^*) &\leq V_T\epsilon. \end{aligned} \tag{23}$$

On the other hand, based on the definition of the total variation distance between $\mathbb{P}_{\mathcal{S}^*}$ and $\mathbb{P}_{\mathcal{S}_T}$, we have

$$\begin{aligned} \mathcal{D}_{\text{TV}}(\mathbb{P}_{\mathcal{S}^*}, \mathbb{P}_{\mathcal{S}_T}) &\triangleq \frac{1}{2}\left[\int_{\mathcal{S}^*-\mathcal{S}_T} \rho_{\mathcal{S}^*} + \int_{\mathcal{S}_T-\mathcal{S}^*} \rho_{\mathcal{S}_T} + \int_{\mathcal{S}^*\cap\mathcal{S}_T} |\rho_{\mathcal{S}^*} - \rho_{\mathcal{S}_T}|\right] \\ &\leq \frac{1}{2}\left[\frac{\text{Vol}(\mathcal{S}^*) - V_T(1-\epsilon)}{\text{Vol}(\mathcal{S}^*)} + \epsilon \cdot \frac{V_T}{V_T} + (1-\epsilon)\left(\frac{V_T}{\text{Vol}(\mathcal{S}^*)} - 1\right)\right] \\ &= \epsilon, \end{aligned} \tag{24}$$

where for any $\mathcal{S} \in \mathbb{S}_K$, $\rho_{\mathcal{S}}$ denotes the probability density function associated to $\mathbb{P}_{\mathcal{S}}$. This completes the proof. □

Next, we take advantage of a well-known generalization bound for consistent binary classifiers that has been originally implied from a series of theorems in [36]. We use a modified version of this bound which has been formulated in Theorem A2.1 of [37] and, for example, can also be found in Lemma 4 of [38]. Formal statement of the above-mentioned generalization bound is rewritten in Lemma 2 for the sake of the readability of the paper. Before proceeding to the lemma, let us define the "growth function" of a set of binary classifiers:

Definition 4. *Assume $\mathcal{X}$ is an arbitrary set of feature vectors and $\mathscr{C} \subseteq \{0,1\}^{\mathcal{X}}$ denotes a corresponding set of binary classifiers. The growth function of $\mathscr{C}$, denoted by $\Pi_{\mathscr{C}}(\cdot) : \mathbb{N} \to \mathbb{N}$, is defined such that for all $m \in \mathbb{N}$, $\Pi_{\mathscr{C}}(m)$ represents the maximum number of distinct ways that any set of m points in $\mathcal{X}$ can be labeled by classifiers in $\mathscr{C}$. Mathematically speaking, we have*

$$\Pi_{\mathscr{C}}(m) \triangleq \sup_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_m \in \mathcal{X}} \left| \left\{ \left(c(\boldsymbol{X}_1), \ldots, c(\boldsymbol{X}_m) \right) \;\middle|\; c \in \mathscr{C} \right\} \right|,$$

where $|\{\cdot\}|$ denotes the cardinality of a set.

Lemma 2. *Assume an arbitrary and unknown distribution $\mathbb{P}$ supported over a set $\mathcal{X}$, and a learnable class of binary classifiers $\mathscr{C} \subseteq \{0,1\}^{\mathcal{X}}$. Let $\Pi_{\mathscr{C}} : \mathbb{N} \to \mathbb{N}$ be the "growth function" of $\mathscr{C}$. Assume a dataset $\tilde{\boldsymbol{D}}$ consisting of n i.i.d. samples drawn from $\mathbb{P}$, where each sample has been labeled by a classifier $c_T \in \mathscr{C}$.*

Assume there exists algorithm $\mathscr{A}$ which given $\tilde{\boldsymbol{D}}$ outputs a consistent classifier $c^* \in \mathscr{C}$. Then, for any $\zeta > 0$, the following generalization bound holds with probability at least $1 - \zeta$:

$$\mathbb{P}\{c^*(\boldsymbol{X}) \neq c_T(\boldsymbol{X})\} \leq O\left(\frac{\log \Pi_{\mathscr{C}}(2n) + \log \frac{1}{\zeta}}{n} \right).$$

Proof can be found in any of the above-mentioned references. Moreover, Lemma 3 proves that for the particular class of K-dimensional simplices, the growth function associated to the classifier set $\mathscr{C}$ can be bounded as $\log \Pi_{\mathscr{C}}(n) \leq O\left(K^2 \log(n/K)\right)$.

Lemma 3. *For $K \in \mathbb{N}$, assume the classifier set $\mathscr{C}$ which consists of all the simplex-based binary classifiers in $\mathbb{R}^K$ as described in (21). Then, the following upper-bound holds for the growth function $\Pi_{\mathscr{C}}$, for $n \in \mathbb{N}$:*

$$\Pi_{\mathscr{C}}(n) \leq \left(\frac{ne}{K+1} \right)^{(K+1)^2}.$$

See Appendix B for the proof. Substituting into the generalization bound of Lemma 2, we have the following bound for the total-variation error of the output of algorithm $\mathscr{A}$, denoted by $\mathcal{S}^*$:

$$\mathcal{D}_{\mathrm{TV}}(\mathbb{P}_{\mathcal{S}_T}, \mathbb{P}_{\mathcal{S}^*}) \leq O\left(\frac{K^2 \log \frac{n}{K} + \log \frac{1}{\zeta}}{n} \right). \tag{25}$$

Following some simple algebra, one can show that the minimum number of samples n which guarantees a maximum total-variation error of $\epsilon > 0$ with probability at least $1 - \zeta$ for algorithm $\mathscr{A}$ must satisfy the following sample complexity bound:

$$n \geq O\left(\frac{K^2 \log \frac{K}{\epsilon} + \log \frac{1}{\zeta}}{\epsilon} \right). \tag{26}$$

What remains to prove is that the ML algorithm has the same *consistency* and *bounded volume property* as the imaginary algorithm $\mathscr{A}$. First, remember that all the data samples in the dataset $\tilde{\boldsymbol{D}}$ are labeled with 0. On the other hand, MLE always finds a simplex that encapsulates all the data points $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Therefore, the classifier in $\mathscr{C}$ which corresponds to the MLE of $\mathcal{S}_T$, say $c\left(\cdot; \mathcal{S}^*_{\mathrm{ML}}\right)$, is always a consistent classifier. This proves the consistency property. In addition, ML always finds the smallest (in terms of Lebesgue measure or volume) simplex that encapsulates all the samples. It should be noted that the true simplex $\mathcal{S}_T$ also corresponds to a consistent classifier in $\mathscr{C}$, which means the volume of $\mathcal{S}^*_{\mathrm{ML}}$ must be smaller or equal to the volume of $\mathcal{S}_T$ which is denoted by V_T, i.e.

$$\mathrm{Vol}\left(\mathcal{S}^*_{\mathrm{ML}}\right) \leq V_T. \tag{27}$$

Therefore, ML algorithm satisfies the two properties mentioned above and the proof is complete. □

Proof of Theorem 2. Recall that $\mathcal{S}^*$ is obtained as

$$\mathcal{S}^* \triangleq \underset{\mathcal{S}\in\mathbb{S}_K}{\operatorname{argmin}} \left\{ \hat{R}_{\mathrm{CRR}}\left(\mathcal{S}\right) \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi\left(\boldsymbol{X}_i, \mathcal{S}\right) + \gamma \mathrm{Vol}\left(\mathcal{S}\right) \right\}, \tag{28}$$

where for $\boldsymbol{X} \in \mathbb{R}^K$, $\phi\left(\boldsymbol{X}, \mathcal{S}\right) \triangleq \sqrt{n}\ell\left(d_{\mathcal{S}}\left(\boldsymbol{X}\right)\right)$. Instead of working directly with $\mathcal{S}^*$ and obtaining an upper bound on $\mathcal{D}_{\mathrm{TV}}\left(\mathbb{P}_{\mathcal{S}^*}, \mathbb{P}_{\mathcal{S}_T}\right)$, we define the following set

$$\begin{aligned} \mathcal{Q}\left(\epsilon\right) \triangleq \Bigg\{ \mathcal{S} \in \mathbb{S}_K \Bigg| \; & \mathrm{Vol}\left(\mathcal{S}\right) \leq V_T \;, \\ & \mathrm{Vol}\left(\mathcal{S} \cap \mathcal{S}_T\right) \geq (1-\epsilon) V_T \;, \\ & \max_{\boldsymbol{u}\in\mathcal{S}_T} d_{\mathcal{S}}\left(\boldsymbol{u}\right) \leq 3\mathrm{Diam}\left(\mathcal{S}_T\right) \Bigg\}, \end{aligned} \tag{29}$$

where the diameter of a simplex $\mathcal{S} \in \mathbb{S}_K$, denoted by $\mathrm{Diam}\left(\mathcal{S}\right)$, is defined as

$$\mathrm{Diam}\left(\mathcal{S}\right) \triangleq \max_{\boldsymbol{u},\boldsymbol{u}'\in\mathcal{S}} \left\|\boldsymbol{u} - \boldsymbol{u}'\right\|_2. \tag{30}$$

Then, we prove the following two claims:
Claim 1: For all $\mathcal{S} \in \mathcal{Q}\left(\epsilon\right)$, we have $\mathcal{D}_{\mathrm{TV}}\left(\mathbb{P}_{\mathcal{S}}, \mathbb{P}_{\mathcal{S}_T}\right) \leq \epsilon$. This is essentially a direct consequence of Lemma 1 which has been stated and subsequently proved in Appendix A.
Claim 2: The probability that $\mathcal{S}^* \in \mathcal{Q}\left(\epsilon\right)$ is at least $1-\zeta$, if the number of samples satisfies the following bound:

$$n \geq \left(\frac{6\ell\left(3\underline{\lambda} K V_T^{\frac{1}{K}}\right)\left(\sqrt{K^2 \log\frac{ne}{K}} + \sqrt{\log\frac{1}{\zeta}}\right) + \gamma V_T \epsilon}{L\left(\frac{\epsilon V_T^{1/K}}{(K+1)\overline{\lambda}}\right)\epsilon} \right)^2. \tag{31}$$

In fact, combining the two claims will prove the theorem. Next, we prove each claim separately.

Proof of Claim 2. Claim 2 can be shown by proving that the following three conditions are held with probability at least $1-\zeta$:
C1: $V^* \triangleq \mathrm{Vol}\left(\mathcal{S}^*\right) \leq V_T$, and consequently, the volume of the solution to (11) is bounded.
C2: $\max_{\boldsymbol{u}\in\mathcal{S}_T} d_{\mathcal{S}^*}\left(\boldsymbol{u}\right) \leq 3\mathrm{Diam}\left(\mathcal{S}_T\right)$, which means $\mathcal{S}^*$ cannot be very far from $\mathcal{S}_T$.
C3: $\mathrm{Vol}\left(\mathcal{S}^* \cap \mathcal{S}_T\right) \geq \left(1-\epsilon\right) V_T$.
The first two conditions **C1** and **C2** do hold with certainty as the following lemma presents.

Lemma 4. *With certainty, we have* $V^* \leq V_T$ *and* $max_{\boldsymbol{u}\in\mathcal{S}_T} d_{\mathcal{S}^*}(\boldsymbol{u}) \leq 3\mathrm{Diam}(\mathcal{S}_T)$.

See Appendix B for the proof. From Lemma 4, one can deduce that $\mathcal{S}^*$ belongs to either $\mathcal{Q}(\epsilon)$ or its partial complement $\mathcal{Q}^c(\epsilon)$ which is defined as

$$\mathcal{Q}^c(\epsilon) \triangleq \Bigg\{ \mathcal{S} \in \mathbb{S}_K \Bigg| \begin{array}{l} \mathrm{Vol}(\mathcal{S}) \leq V_T \ , \\ \mathrm{Vol}(\mathcal{S} \cap \mathcal{S}_T) < (1-\epsilon) V_T \ , \\ \max\limits_{\boldsymbol{u}\in\mathcal{S}_T} d_{\mathcal{S}}(\boldsymbol{u}) \leq 3\mathrm{Diam}(\mathcal{S}_T) \end{array} \Bigg\}, \tag{32}$$

For a given ϵ, we are interested in finding n such that $\mathcal{S}^* \in \mathcal{Q}(\epsilon)$ with probability at least $1-\zeta$. To this end, we seek an appropriate function $u(n,\epsilon)$ such that

$$\mathbb{P}\left\{ \inf_{\mathcal{S}\in\mathcal{Q}^c(\epsilon)} \hat{R}_{\mathrm{CRR}}(\mathcal{S}) - \hat{R}_{\mathrm{CRR}}(\mathcal{S}_T) > u(n,\epsilon) \right\} > 1-\zeta. \tag{33}$$

Next, we find n such that $u(n,\epsilon)$ is non-negative. In this way, one can deduce that $\mathcal{S}_T$ is preferable over all the simplices in $\mathcal{Q}^c(\epsilon)$ with probability at least $1-\zeta$. Since $\mathcal{S}^*$ is the optimal solution, it is in $\mathcal{Q}(\epsilon)$ with probability at least $1-\zeta$. To this aim, we obtain a candidate for $u(n,\epsilon)$ by considering two facts for each $\mathcal{S} \in \mathcal{Q}^c(\epsilon)$. First, $\hat{R}_{\mathrm{CRR}}(\mathcal{S})$ is sufficiently close to its statistical average $\mathbb{E}_{\mathcal{S}_T}\hat{R}_{\mathrm{CRR}}(\mathcal{S})$. Second, $\mathbb{E}_{\mathcal{S}_T}\hat{R}_{\mathrm{CRR}}(\mathcal{S})$ is sufficiently larger than $\hat{R}_{\mathrm{CRR}}(\mathcal{S}_T)$. The main challenge is to show that these two facts hold for all simplices in $\mathcal{Q}^c(\epsilon)$, with high probability. To prove the first fact, we define $\phi_{\max}$ as

$$\phi_{\max} \triangleq \sup_{\boldsymbol{X}\in\mathcal{S}_T} \ \sup_{\mathcal{S}\in\mathcal{Q}(\epsilon)\cup\mathcal{Q}^c(\epsilon)} \phi(\boldsymbol{X},\mathcal{S}), \tag{34}$$

and use it to normalize the function ϕ as

$$\bar{\phi}(\boldsymbol{X},\mathcal{S}) \triangleq \frac{\sqrt{n}\ell(d_{\mathcal{S}}(\boldsymbol{X}))}{\phi_{\max}}. \tag{35}$$

For any $\boldsymbol{X} \in \mathcal{S}_T$ and $\mathcal{S} \in \mathcal{Q}(\epsilon) \cup \mathcal{Q}^c(\epsilon)$, we have:

$$\phi_{\max} = \sup_{\boldsymbol{X},\mathcal{S}} \sqrt{n}\ell(d_{\mathcal{S}}(\boldsymbol{X})) \leq \sqrt{n}\ell\left(3 \max_{\boldsymbol{u},\boldsymbol{u}'\in\mathcal{S}_T} \left\|\boldsymbol{u}-\boldsymbol{u}'\right\|_2\right) \leq \sqrt{n}\ell\left(3\underline{\lambda}KV_T^{\frac{1}{K}}\right), \tag{36}$$

which holds due to both the $(\underline{\lambda},\bar{\lambda})$-isoperimetricity assumption on $\mathcal{S}_T$, and ℓ being an increasing function. In more details, we already know from Definition 1 that

$$\max_{\boldsymbol{u},\boldsymbol{u}'\in\mathcal{S}_T} \left\|\boldsymbol{u}-\boldsymbol{u}'\right\|_2 \leq \max_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\boldsymbol{\Theta_T}} \left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|_2 \leq \underline{\lambda}KV_T^{\frac{1}{K}}. \tag{37}$$

Also, since ℓ is increasing, the maximum of its output occurs when the input argument is maximized. Defining the function set Φ as

$$\Phi \triangleq \left\{ \bar{\phi}(\cdot,\mathcal{S}) : \mathcal{S}_T \to [0,1] \ \middle| \ \mathcal{S} \in \mathcal{Q}^c(\epsilon) \right\},$$

and using one of the fundamental theorems in Vapnik-Chervonenkis (VC) theory of statistical learning gives us the following bound, for all $\zeta > 0$:

$$\mathbb{P}\left\{ \sup_{\mathcal{S}\in\mathcal{Q}^c(\epsilon)} \left| \frac{\hat{R}_{\mathrm{CRR}}(\mathcal{S}) - \mathbb{E}_{\mathcal{S}_T}\hat{R}_{\mathrm{CRR}}(\mathcal{S})}{\phi_{\max}} \right| \overset{(a)}{\leq} 2\mathfrak{R}_n\{\Phi\} + \sqrt{\frac{\log(2/\zeta)}{2n}} \right\} \overset{(b)}{\geq} 1-\zeta, \tag{38}$$

where $\mathfrak{R}_n\{\Phi\}$ represents the *Rademacher complexity* of the function set Φ, for sample size n and dataset $\boldsymbol{D} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \sim \mathbb{P}_{\mathcal{S}_T}$ [10].

One of the main challenges is to find an appropriate upper bound on $\mathfrak{R}_n\{\Phi\}$. In the following lemma, we obtain an upper bound showing that both terms in the r.h.s. of inequality (a) in (38) are $\tilde{O}\left(n^{-1/2}\right)$.

Lemma 5. *Rademacher complexity of the function set Φ, for all $\epsilon > 0$ and increasing functions $\ell : \mathbb{R} \to \mathbb{R}$ can be bounded as*

$$\mathfrak{R}_n\{\Phi\} \leq \sqrt{\frac{4K^2}{n} \log\left(\frac{ne}{K}\right)},$$

for $n, K \in \mathbb{N}$

See Appendix B for the proof. The second fact, i.e. for all $\mathcal{S} \in \mathcal{Q}^c(\epsilon)$, the value of $\mathbb{E}_{\mathcal{S}_T}\hat{R}_{\mathrm{CRR}}(\mathcal{S})$ is sufficiently larger than $\hat{R}_{\mathrm{CRR}}(\mathcal{S}_T)$, is formalized by the following lemma.

Lemma 6. *Let $\mathcal{S}_T \in \mathbb{S}_K$ be $\left(\underline{\lambda}, \bar{\lambda}\right)$-isoperimetric and assume $\boldsymbol{D} \triangleq \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ are n i.i.d. samples drawn uniformly from $\mathcal{S}_T$. For any increasing and integrable function ℓ, all $\gamma \geq 0$ and $0 < \epsilon \leq 1$, assume we have*

$$n \geq \left(\frac{\gamma V_T}{\ell\left(\frac{\epsilon V_T^{1/K}}{(K+1)\bar{\lambda}}\right)}\right)^2. \tag{39}$$

Then, the following bound holds:

$$\inf_{\mathcal{S}\in\mathcal{Q}^c(\epsilon)} \mathbb{E}_{\mathcal{S}_T}\left\{\hat{R}_{\mathrm{CRR}}(\mathcal{S}; \boldsymbol{D}, \ell, \gamma)\right\} \geq \left(\sqrt{n}L\left(\frac{\epsilon V_T^{1/K}}{(K+1)\bar{\lambda}}\right) - \gamma V_T\right)\epsilon + \hat{R}_{\mathrm{CRR}}(\mathcal{S}_T; \boldsymbol{D}, \ell, \gamma), \tag{40}$$

where $L(x) \triangleq \int_0^x \ell(u)\,\mathrm{d}u - \ell(0)$.

See Appendix B for the proof. From (38) and Lemma 5, the following lower-bound holds for all $\mathcal{S} \in \mathcal{Q}^c(\epsilon)$ with probability at least $1 - \zeta$:

$$\hat{R}_{\mathrm{CRR}}(\mathcal{S}) \geq \mathbb{E}_{\mathcal{S}_T}\hat{R}_{\mathrm{CRR}}(\mathcal{S}) - \ell\left(3\underline{\lambda}KV_T^{\frac{1}{K}}\right)\left(4\sqrt{K^2\log\frac{ne}{K}} + \sqrt{\log\frac{2}{\zeta}}\right). \tag{41}$$

Substituting from Lemma 6, for all $\mathcal{S} \in \mathcal{Q}^c(\epsilon)$ and with probability at least $1 - \zeta$, we have

$$\begin{aligned}\hat{R}_{\mathrm{CRR}}(\mathcal{S}) - \hat{R}_{\mathrm{CRR}}(\mathcal{S}_T) \geq\ & \left(\sqrt{n}L\left(\frac{\epsilon V_T^{1/K}}{(K+1)\bar{\lambda}}\right) - \gamma V_T\right)\epsilon - \\ & \ell\left(3\underline{\lambda}KV_T^{\frac{1}{K}}\right)\left(4\sqrt{K^2\log\frac{ne}{K}} + \sqrt{\log\frac{2}{\zeta}}\right) \triangleq u(n,\epsilon),\end{aligned} \tag{42}$$

subject to n satisfies (39). In this way, we have obtained the desired $u(n,\epsilon)$. One can show that for

$$n \geq \left(\frac{6\ell\left(3\underline{\lambda}KV_T^{\frac{1}{K}}\right)\left(\sqrt{K^2\log\frac{ne}{K}} + \sqrt{\log\frac{1}{\zeta}}\right) + \gamma V_T\epsilon}{L\left(\frac{\epsilon V_T^{1/K}}{(K+1)\bar{\lambda}}\right)\epsilon}\right)^2, \tag{43}$$

both (39) and $u(n,\epsilon) > 0$ are satisfied. Therefore,

$$\mathbb{P}\left\{\inf_{\mathcal{S}\in\mathcal{Q}^c(\epsilon)} \hat{R}_{\mathrm{CRR}}(\mathcal{S}) - \hat{R}_{\mathrm{CRR}}(\mathcal{S}_T) > 0\right\} > 1-\zeta, \tag{44}$$

which implies $\mathbb{P}\{\mathcal{S}^* \in \mathcal{Q}(\epsilon)\} > 1-\zeta$ and completes the proof of Claim 2. □

As mentioned earlier, combining Claims 1 and 2 proves Theorem 2. □

Proof of Theorem 3. According to the definition of $d_{\mathcal{S}}(\cdot)$, having $d_{\mathcal{S}}(\boldsymbol{X}) = 0$ immediately implies that $\boldsymbol{w}_k^T\boldsymbol{X} + b_k \leq 0,\ k = 0,\dots,K$. However, we have already assumed that $\boldsymbol{w}_k^T\boldsymbol{X} + b_k < 0$ for all $k = 0,\dots,K$, which means $\boldsymbol{X}$ is inside $\mathcal{S}$ with a positive margin from its boundary. Therefore, an infinitesimally small perturbation of $\boldsymbol{\Theta}$ (regardless of the direction) cannot throw $\boldsymbol{X}$ out of $\mathcal{S}(\boldsymbol{\Theta})$ and thus the gradient of $d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X})$ is zero.

For the case of $d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X}) > 0$, one can deduce that $\boldsymbol{w}_k^T\boldsymbol{X} + b_k > 0$ for at least one $k \in \{0,\dots,K\}$. Due to the assumption of lemma, $k^* \triangleq \operatorname{argmax}_k\ \boldsymbol{w}_k^T\boldsymbol{X} + b_k$ is unique. This alternatively implies that there exists $\delta > 0$ such that for all $\boldsymbol{u} \in \mathbb{R}^K$ with $\|\boldsymbol{u} - \boldsymbol{X}\|_2 \leq \delta$ we have $d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{u}) = \boldsymbol{w}_{k^*}^T\boldsymbol{u} + b_{k^*}$. Therefore, we have

$$\nabla_{\boldsymbol{\Theta}} d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X}) = \nabla_{\boldsymbol{\Theta}}\left(\min_{\boldsymbol{p}\in\mathbb{R}^K|\ \boldsymbol{p}^T\mathbf{1}=1} f(\boldsymbol{X},\boldsymbol{p}) \triangleq \|\boldsymbol{X} - \boldsymbol{\Theta}_{-k^*}\boldsymbol{p}\|_2\right), \tag{45}$$

which is based on the fact that $d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X})$ is the minimum Euclidean distance of $\boldsymbol{X}$ from $\mathcal{H}_{k^*} = \left\{\boldsymbol{\Theta}_{-k^*}\boldsymbol{p}|\ \boldsymbol{p}\in\mathbb{R}^K\ ,\ \boldsymbol{p}^T\mathbf{1} = 1\right\}$ which represents the hyper-plane including the k^*th facet of $\mathcal{S}$. Due to the continuity of f, one can use the envelope theorem and obtain the following formulation for the gradient:

$$\nabla_{\boldsymbol{\Theta}_{-k^*}} d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X}) = \nabla_{\boldsymbol{\Theta}_{-k^*}} \|\boldsymbol{X} - \boldsymbol{\Theta}_{-k^*}\boldsymbol{p}^*\|_2 = \frac{(\boldsymbol{\Theta}_{-k^*}\boldsymbol{p}^* - \boldsymbol{X})\,\boldsymbol{p}^{*T}}{\|\boldsymbol{\Theta}_{-k^*}\boldsymbol{p}^* - \boldsymbol{X}\|_2}, \tag{46}$$

where

$$\boldsymbol{p}^* \triangleq \operatorname*{argmin}_{\boldsymbol{p}\in\mathbb{R}^K|\ \boldsymbol{p}^T\mathbf{1}=1} \|\boldsymbol{X} - \boldsymbol{\Theta}_{-k^*}\boldsymbol{p}\|_2. \tag{47}$$

It should not be surprising that derivatives w.r.t. $\boldsymbol{\theta}_{k^*}$ is zero since the k^*th vertex does not contribute to the planar distance from $\mathcal{H}_{k^*}$. Also, note that $\boldsymbol{p^*}$ can be computed analytically. More precisely, $\boldsymbol{\Theta}_{-k^*}\boldsymbol{p}^*$ is the projection of $\boldsymbol{X}$ into the $(K-1)$-dimensional subspace $\mathcal{H}_{k^*}$. Thus, we have

$$\begin{aligned} \boldsymbol{\Theta}_{-k^*}\boldsymbol{p}^* &= \boldsymbol{X} - \left(\boldsymbol{w}_{k^*}^T\boldsymbol{X} + b_{k^*}\right)\boldsymbol{w}_{k^*}, \\ \text{with}\quad b_{k^*} &= \frac{-1}{K}\boldsymbol{w}_{k^*}^T\sum_{k\neq k^*}\boldsymbol{\theta}_k = -\frac{1}{K}\boldsymbol{\Theta}_{-k^*}\mathbf{1}, \end{aligned} \tag{48}$$

where $\boldsymbol{w}_{k^*}$, as mentioned in the lemma, is the normal outward vector of the k^*th facet of $\mathcal{S}$. In order to compute $\boldsymbol{w}_{k^*}$, it should be noted that $\boldsymbol{w}_{k^*}$ must be orthogonal to all $\boldsymbol{\theta}_k - \frac{1}{K}\boldsymbol{\Theta}_{-k^*}\mathbf{1}$ for $k \neq k^*$. Thus, we have

$$\left(\boldsymbol{\Theta}_{-k^*} - \frac{1}{K}\boldsymbol{\Theta}_{-k^*}\mathbf{1}\mathbf{1}^T\right)^T\boldsymbol{w}_{k^*} = \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{K}\right)\boldsymbol{\Theta}_{-k^*}^T\boldsymbol{w}_{k^*} = 0. \tag{49}$$

Since the columns of $\boldsymbol{\Theta}_{-k^*} - \frac{1}{K}\boldsymbol{\Theta}_{-k^*}\mathbf{1}\mathbf{1}^T$, or equivalently $\boldsymbol{\theta}_k - \frac{1}{K}\boldsymbol{\Theta}_{-k^*}\mathbf{1}$ for $k \neq k^*$, all lie on the hyper-plane $\mathcal{H}_{k^*}$, then $\left(\boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right)\boldsymbol{\Theta}_{-k^*}^T$ must have at least one zero eigenvalue. On the other

hand, since we have already assumed that $\mathcal{S}(\boldsymbol{\Theta})$ has a non-zero volume, the projections of the mentioned columns on $\mathcal{H}_{k^*}$ are linearly independent in $\mathbb{R}^{K-1}$, and therefore the above-mentioned zero eigenvalue is unique. Thus, $\boldsymbol{w}_{k^*}$ can be obtained as described in the statement of lemma. In this regard, we have

$$\boldsymbol{\Theta}_{-k^*}\boldsymbol{p}^* = \boldsymbol{X} - \left(\boldsymbol{w}_{k^*}^T\boldsymbol{X} + b_{k^*}\right)\boldsymbol{w}_{k^*} = \boldsymbol{X} - \boldsymbol{w}_{k^*}\boldsymbol{w}_{k^*}^T\left(\boldsymbol{X} - \frac{1}{K}\boldsymbol{\Theta}_{-k^*}\mathbf{1}\right), \tag{50}$$

which means $\boldsymbol{p}^* \triangleq \boldsymbol{\Theta}_{-k^*}^{\dagger}\left(\boldsymbol{X} - \boldsymbol{w}_{k^*}\boldsymbol{w}_{k^*}^T\left(\boldsymbol{X} - \frac{1}{K}\boldsymbol{\Theta}_{-k^*}\mathbf{1}\right)\right) \in \mathbb{R}^K$. Note that using the pseudo-inverse operator does not result into any information loss here since $\boldsymbol{X} - \left(\boldsymbol{w}_{k^*}^T\boldsymbol{X} + b_{k^*}\right)\boldsymbol{w}_{k^*}$ lies on the hyper-plane specified by $\boldsymbol{\Theta}_{-k^*}$. Finally, we have $\nabla_{\boldsymbol{\Theta}_{-k^*}}(\cdot) = \nabla_{\boldsymbol{\Theta}}(\cdot)\,\boldsymbol{J}_{k^*}$, and $\nabla_{\boldsymbol{\Theta}}\ell(\cdot) = \ell'(\cdot)\,\nabla_{\boldsymbol{\Theta}}(\cdot)$. Considering the fact that

$$\boldsymbol{w}_{k^*} = -\frac{\boldsymbol{\Theta}_{-k^*}\boldsymbol{p}^* - \boldsymbol{X}}{\left\|\boldsymbol{\Theta}_{-k^*}\boldsymbol{p}^* - \boldsymbol{X}\right\|_2}, \tag{51}$$

and substituting into (46), completes the proof. □

B Proofs of Lemmas and Corollaries

Proof of Corollary 1. Proof is similar to that of Corollary 1. First, let us compute the integral function $L(\cdot)$:

$$L(x) = \frac{1}{x}\int_0^x \left(1 - e^{-bu}\right)\mathrm{d}u = 1 - \left(\frac{1 - e^{-bx}}{bx}\right), \quad x \triangleq \frac{\epsilon V_T^{1/K}}{(K+1)\,\bar{\lambda}}. \tag{52}$$

Considering the fact that $b = K/\epsilon$, it can be easily verified that $L(x)$ has a positive lower-bound which is independent of K, ϵ and ζ. On the other hand, we have $\ell(\cdot) \leq 1$ due to the particular choice of ℓ. Substituting the above results into the sample complexity of Theorem 2, and using the same techniques which we already used for the proof Theorem 1 will complete the proof. □

Proof of Lemma 3. The proof is motivated by the fact that any K-simplex $\mathcal{S} \in \mathbb{S}_K$ can be viewed as a set of points in $\mathbb{R}^K$ satisfying the following set of constraints:

$$\mathcal{S} = \left\{\boldsymbol{u} \in \mathbb{R}^K \,\middle|\, \boldsymbol{w}_k^T\boldsymbol{u} + b_k \leq 0,\ \forall k = 0, 1, \ldots, K\right\}, \tag{53}$$

where for each k, the couple $H_k \triangleq (\boldsymbol{w}_k, b_k) \in \mathbb{R}^K \times \mathbb{R}$ represent a $(K-1)$-dimensional linear hyper-plane that includes the kth facet of $\mathcal{S}$. Not to mention that $\boldsymbol{w}_k$ are considered to be outwards w.r.t. $\mathcal{S}$. We can then write $\mathcal{S} = \mathcal{S}(H_0, \ldots, H_K)$.

Assume one aims to use a K-simplex as a binary classifier, where for any $\mathcal{S} \in \mathbb{S}_K$ and $\boldsymbol{x} \in \mathbb{R}^K$, $\boldsymbol{x}$ is labeled as $+1$ if $\boldsymbol{x} \in \mathcal{S}$ and is labeled -1, otherwise. Then, we are interested to bound the maximum number of ways that a set of n generally-positioned points in $\mathbb{R}^K$ can be labeled via simplices in $\mathbb{S}_K$, i.e. the growth function of $\mathscr{C}$ or $\Pi_{\mathscr{C}}(n)$. Fortunately, this goal can be achieved by using the growth function of hyper-planes in $\mathbb{R}^K$, which is already derived.

Let us define $\mathcal{L}_{K+1} \triangleq \{-1, +1\}^{K+1}$ as the set of all binary *super labels* of dimension $K+1$ and the function family $\mathcal{H}$, such that $h = h(\boldsymbol{w}_{0:K}, b_{0:K}) \in \mathcal{H} : \mathbb{R}^K \to \mathcal{L}_{K+1}$ is defined as:

$$h(\boldsymbol{x}) \triangleq \left[\mathrm{sign}\left(\boldsymbol{w}_0^T\boldsymbol{x} + b_0\right)\ ,\ \ldots\ ,\ \mathrm{sign}\left(\boldsymbol{w}_K^T\boldsymbol{x} + b_K\right)\right]^T,\ \forall \boldsymbol{x} \in \mathbb{R}^K, \tag{54}$$

where we define $\text{sign}(0) = -1$. Then, by letting $\mathcal{I}_K$ to be the set of all hyper-plane binary classifiers in $\mathbb{R}^K$ and according to the *rule of product*, the growth function of $\mathcal{H}$ can be computed as

$$\Pi_{\mathcal{H}}(n) = \underbrace{\Pi_{\mathcal{I}_K}(n) \times \cdots \times \Pi_{\mathcal{I}_K}(n)}_{K+1 \text{ times}}. \tag{55}$$

It is easy to see that $\Pi_{\mathscr{C}}(n) \leq \Pi_{\mathcal{H}}(n)$. In fact, for any $\mathcal{S} = \mathcal{S}(H_0, \ldots, H_K) \in \mathbb{S}_K$, the hyper-plane parameter sets $H_0, \ldots, H_K$ uniquely identify a function in $\mathcal{H}$ that can assign super labels to $\mathbb{R}^K$, while only a subset of them can be generated by $\mathcal{S}$. On the other hand, considering the fact that VC dimension of hyper-planes in $\mathbb{R}^K$ is exactly $K+1$, and using the Sauer-Shelah lemma (see, for example, Theorem 3.5 of [10]), we have the following bound for $\Pi_{\mathcal{I}_K}(n)$:

$$\Pi_{\mathcal{I}_K}(n) \leq \sum_{i=0}^{K+1} \binom{n}{i} \leq \left(\frac{ne}{K+1}\right)^{K+1}, \tag{56}$$

which, according to the previous inequality, means: $\Pi_{\mathscr{C}}(n) \leq \left(\frac{ne}{K+1}\right)^{(K+1)^2}$ and completes the proof. □

Proof of Lemma 4. First, we prove $V^* \triangleq \text{Vol}(\mathcal{S}^*) \leq V_T$; The proof is by contradiction. Assume $V^* > V_T$. Since $d_{\mathcal{S}_T}(\boldsymbol{X}_i) = 0$ for $i = 1, \ldots, n$, and the fact that ℓ is an increasing function, it can be readily seen that

$$\hat{R}_{\text{CRR}}(\mathcal{S}^*) \geq \sqrt{n}\ell(0) + \gamma V^* > \sqrt{n}\ell(0) + \gamma V_T = \hat{R}_{\text{CRR}}(\mathcal{S}_T), \tag{57}$$

which contradicts the fact that $\mathcal{S}^*$ is the minimizer.

Next, we show that

$$\max_{\boldsymbol{u} \in \mathcal{S}_T} d_{\mathcal{S}^*}(\boldsymbol{u}) \leq 3 \max_{\boldsymbol{u}, \boldsymbol{u}' \in \mathcal{S}_T} \left\|\boldsymbol{u} - \boldsymbol{u}'\right\|_2, \tag{58}$$

where $\mathcal{S}^*$ is the minimizer of (11) for the dataset $\boldsymbol{D} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ that includes an arbitrary set of n points $\boldsymbol{X}_i \in \mathcal{S}_T,\ i \in [n]$. It should be noted that for this lemma, we do not need to assume any particular stochastic model for the generation of $\boldsymbol{X}_i$s, and these points can be arbitrarily chosen from $\mathcal{S}_T$. The main implication of the above inequality is the fact that $\mathcal{S}^*$ cannot be very far from the the points in $\boldsymbol{D}$, and the maximum planar distance between each $\boldsymbol{X}_i$ and $\mathcal{S}^*$ is bounded by a factor of the diameter of $\mathcal{S}_T$ ($\text{Diam}(\mathcal{S}_T) \triangleq \max_{\boldsymbol{u}, \boldsymbol{u}' \in \mathcal{S}_T} \|\boldsymbol{u} - \boldsymbol{u}'\|_2$).

Before attempting to prove (58), we first show that for any $\mathcal{S}^* \in \mathbb{S}_K$ and $\boldsymbol{Z} \in \mathbb{R}^K$, the following useful bound holds for planar distance $d_{\mathcal{S}^*}(\boldsymbol{Z})$:

$$d_{\mathcal{S}^*}(\boldsymbol{Z}) \leq \min_{\boldsymbol{u} \in \mathcal{S}^*} \|\boldsymbol{Z} - \boldsymbol{u}\|_2. \tag{59}$$

Let us denote the r.h.s. of (59) as the *spherical distance* of $\boldsymbol{Z}$ from $\mathcal{S}^*$. For $\boldsymbol{Z} \in \mathcal{S}^*$, both the planar distance and spherical distance of $\boldsymbol{Z}$ from $\mathcal{S}^*$ are zero and thus the inequality holds. For $\boldsymbol{Z} \notin \mathcal{S}^*$, first consider $\boldsymbol{w}_k \in \mathbb{R}^K$ (with $\|\boldsymbol{w}_k\|_2 = 1$) and $b_k \in \mathbb{R}$, and then let $\mathcal{H}_k \triangleq \left\{\boldsymbol{x} \in \mathbb{R}^K \middle|\ \boldsymbol{w}_k^T \boldsymbol{x} + b_k = 0\right\}$ represent the hyper-plane that encompasses the kth facet of $\mathcal{S}^*$. This way, the planar distance can be rewritten as $d_{\mathcal{S}^*}(\boldsymbol{Z}) = \min_{\boldsymbol{u} \in \mathcal{H}_{k^*}} \|\boldsymbol{Z} - \boldsymbol{u}\|_2$, where $k^* \triangleq \text{argmax}_k\ \boldsymbol{w}_k^T \boldsymbol{Z} + b_k$. Also, for $\boldsymbol{Z} \notin \mathcal{S}^*$ we have $\boldsymbol{w}_{k^*}^T \boldsymbol{Z} + b_{k^*} > 0$, while for all $\boldsymbol{u} \in \mathcal{S}^*$, we have $\boldsymbol{w}_{k^*}^T \boldsymbol{u} + b_{k^*} \leq 0$, based on the fact that $\boldsymbol{w}_k$ vectors are defined as outward unitary vectors w.r.t. $\mathcal{S}^*$. Therefore, the line segment which connects $\boldsymbol{Z}$ to $\boldsymbol{u}$ must cross $\mathcal{H}_{k^*}$ at some point, which we denote by $\boldsymbol{Z}^*(\boldsymbol{u}) \in \mathcal{H}_{k^*}$. Note that we have:

$$\|\boldsymbol{Z} - \boldsymbol{u}\|_2 = \|\boldsymbol{Z} - \boldsymbol{Z}^*(\boldsymbol{u})\|_2 + \|\boldsymbol{Z}^*(\boldsymbol{u}) - \boldsymbol{u}\|_2 \geq \|\boldsymbol{Z} - \boldsymbol{Z}^*(\boldsymbol{u})\|_2. \tag{60}$$

Therefore, the following relations hold and (59) is proved:

$$\min_{\boldsymbol{u}\in\mathcal{S}^*}\|\boldsymbol{Z}-\boldsymbol{u}\|_2\geq\min_{\boldsymbol{u}\in\mathcal{S}^*}\|\boldsymbol{Z}-\boldsymbol{Z}^*(\boldsymbol{u})\|_2\geq\min_{\boldsymbol{u}\in\mathcal{H}_{k^*}}\|\boldsymbol{Z}-\boldsymbol{u}\|_2=d_{\mathcal{S}^*}(\boldsymbol{Z}). \tag{61}$$

Back to proving (58), first let us define $\bar{\boldsymbol{X}}\triangleq\frac{1}{n}\sum_i\boldsymbol{X}_i$ as the center of mass for data points in $\boldsymbol{D}$. Then, we show the following set of inequalities hold with certainty:

$$\boldsymbol{w}_k^T\bar{\boldsymbol{X}}+b_k\leq 2\text{Diam}(\mathcal{S}_T),\ k=0,\ldots,K. \tag{62}$$

Based on the definition of diameter $\text{Diam}(\cdot)$ for a simplex, the following inequality hold for any $\boldsymbol{Z}\in\mathcal{S}_T$:

$$\max_{\|\boldsymbol{w}\|_2=1}\boldsymbol{w}^T\left(\boldsymbol{Z}-\bar{\boldsymbol{X}}\right)\leq\max_{\boldsymbol{u},\boldsymbol{u}'}\left\|\boldsymbol{u}-\boldsymbol{u}'\right\|_2=\text{Diam}(\mathcal{S}_T), \tag{63}$$

which also takes into account the fact that due to the convexity of a simplex, we have $\bar{\boldsymbol{X}}\in\mathcal{S}_T$. In this regard, we can prove (62) by contradiction: assume there exists k such that $\boldsymbol{w}_k^T\bar{\boldsymbol{X}}+b_k>2\text{Diam}(\mathcal{S}_T)$. Then, one can show that for all $i\in[n]$ we have

$$2\text{Diam}(\mathcal{S}_T)<\boldsymbol{w}_k^T\bar{\boldsymbol{X}}+b_k=\boldsymbol{w}_k^T\left(\bar{\boldsymbol{X}}-\boldsymbol{X}_i\right)+\boldsymbol{w}_k^T\boldsymbol{X}_i+b_k\leq\text{Diam}(\mathcal{S}_T)+\boldsymbol{w}_k^T\boldsymbol{X}_i+b_k, \tag{64}$$

or alternatively, $\boldsymbol{w}_k^T\boldsymbol{X}_i+b_k>\text{Diam}(\mathcal{S}_T)$. This way, due to the definition of the planar distance in Section 2, we end up with the following lower-bound for $d_{\mathcal{S}^*}(\boldsymbol{X}_i)$ for all $i\in[n]$:

$$d_{\mathcal{S}^*}(\boldsymbol{X}_i)=\max\left\{0\ ,\ \max_t\ \boldsymbol{w}_t^T\boldsymbol{X}_i+b_t\right\}>\text{Diam}(\mathcal{S}_T). \tag{65}$$

The above result also gives a lower-bound for the continuously-relaxed risk in (11) as $\hat{R}(\mathcal{S}^*;\boldsymbol{D})>\sqrt{n}\ell(\text{Diam}(\mathcal{S}_T))+\gamma V^*$, where $V^*\triangleq\text{Vol}(\mathcal{S}^*)$. On the other hand, assume one shifts the simplex $\mathcal{S}^*$ to obtain a new simplex $\hat{\mathcal{S}}^*$ which includes the point $\bar{\boldsymbol{X}}$. Note that the volume of $\hat{\mathcal{S}}^*$ is also V^*. However, the following relations hold w.r.t. $\hat{S}^*$ and for all $i\in[n]$:

$$d_{\hat{S}^*}(\boldsymbol{X}_i)\leq\min_{\boldsymbol{u}\in\hat{S}^*}\|\boldsymbol{X}_i-\boldsymbol{u}\|_2\leq\left\|\boldsymbol{X}_i-\bar{\boldsymbol{X}}\right\|_2\leq\text{Diam}(\mathcal{S}_T), \tag{66}$$

which also gives the following upper-bound for the continuously-relaxed risk in (11): $\hat{R}\left(\hat{\mathcal{S}}^*;\boldsymbol{D}\right)\leq\sqrt{n}\ell(\text{Diam}(\mathcal{S}_T))+\gamma V^*$. This yields that $\hat{\mathcal{S}}^*$ results in a smaller objective value for (11) which contradicts the assumption that $\mathcal{S}^*$ is a global minimizer. Therefore, we must have $\boldsymbol{w}_k^T\bar{\boldsymbol{X}}+b_k\leq 2\text{Diam}(\mathcal{S}_T),\ \forall k=0,\ldots,K$, which results in the following inequality for each $k=0,\ldots,K$ and all $\boldsymbol{Z}\in\mathcal{S}_T$:

$$\boldsymbol{w}_k^T\boldsymbol{Z}+b_k=\boldsymbol{w}_k^T\left(\boldsymbol{Z}-\bar{\boldsymbol{X}}\right)+\boldsymbol{w}_k^T\bar{\boldsymbol{X}}+b_k\leq 3\text{Diam}(\mathcal{S}_T). \tag{67}$$

This also implies that $d_{\mathcal{S}^*}(\boldsymbol{Z})\leq 3\text{Diam}(\mathcal{S}_T)$ and completes the proof. □

Proof of Lemma 5. Rademacher complexity of hypothesis set Φ measures its capability to (in some sense) correlate with a set of n randomly labeled and generally positioned points in the input space. Mathematically speaking, it has the following definition:

$$\mathfrak{R}_n\{\Phi\}\triangleq\mathbb{E}_{\boldsymbol{X}_{1:n}\sim\mathbb{P}_{\mathcal{S}_T}}\left\{\mathbb{E}_{\boldsymbol{\sigma}}\left(\sup_{\mathcal{S}\in\mathcal{Q}^c(\epsilon)}\frac{1}{n}\sum_{i=1}^n\sigma_i\bar{\phi}(\boldsymbol{X}_i,\mathcal{S})\right)\right\}, \tag{68}$$

where $\boldsymbol{\sigma}\in\{-1,+1\}^n$ is a vector of *i.i.d.* Rademacher random variables, and $\bar{\phi}(\cdot,\cdot)$ has been already defined in (35). Also, remember the set of binary classifiers $\mathscr{C}$ which was defined in the

proof of Theorem 1 in Appendix A; Let us rewrite the definition of the classifier set $\mathscr{C}$, this time using the planar distance $d_{\mathcal{S}}(\cdot)$:

$$\mathscr{C} \triangleq \left\{ f\left(d_{\mathcal{S}}(\cdot)\right) : \mathbb{R}^K \to \{0,1\} \;\middle|\; \mathcal{S} \in \mathbb{S}_K \right\}, \tag{69}$$

where function $f : \mathbb{R} \to \{0,1\}$ is defined as below:

$$f(x) \triangleq \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}. \tag{70}$$

The binary classifier set $\mathscr{C}$ is different from Φ in two important aspects. First, functions in $\mathscr{C}$ assign binary labels $\{0,1\}$ instead of real values in $[0,1]$. Also, $\mathcal{S}$ is chosen from $\mathbb{S}_K$ instead of $\mathcal{Q}^c(\epsilon)$ in the case of Φ. This way, we can proceed to the proof by first showing that

$$\mathfrak{R}_n\{\Phi\} \leq \mathfrak{R}_n\{\mathscr{C}\}. \tag{71}$$

The reason behind this approach is the fact that r.h.s. of (71) is much easier to handle in terms of statistical complexity. In order to show (71), we prove a stronger claim. More precisely, we prove that for each $\boldsymbol{\sigma} \in \{-1,+1\}^n$, the following inequality holds for any dataset $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ in $\mathcal{S}_T$:

$$J \triangleq \sup_{\mathcal{S} \in \mathcal{Q}^c(\epsilon)} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\phi}(\boldsymbol{X}_i, \mathcal{S}) \leq \sup_{\mathcal{S} \in \mathbb{S}_K} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\bar{\phi}(\boldsymbol{X}_i, \mathcal{S})) \triangleq J_{(\text{ext})}. \tag{72}$$

It should be reminded that $\mathcal{Q}^c(\epsilon) \subseteq \mathbb{S}_K$, and also $f\left(d_{\mathcal{S}}(\boldsymbol{X}_i)\right) = f\left(\bar{\phi}(\boldsymbol{X}_i, \boldsymbol{S})\right)$. Let $\mathcal{S}_{\max} \in \mathbb{S}_K$ represent the maximizer of the l.h.s. of (72). Without loss of generality, assume that:

$$\begin{aligned} 0 = \bar{\phi}(\boldsymbol{X}_1, \mathcal{S}_{\max}) = \cdots = \bar{\phi}(\boldsymbol{X}_{n_0}, \mathcal{S}_{\max}) \\ < \bar{\phi}(\boldsymbol{X}_{n_0+1}, \mathcal{S}_{\max}) \leq \cdots \leq \bar{\phi}(\boldsymbol{X}_n, \mathcal{S}_{\max}), \end{aligned} \tag{73}$$

where n_0 shows the number of data in the dataset such that $\bar{\phi}(\boldsymbol{X}_i, \mathcal{S}_{\max}) = 0$. For the sake of simplicity, assume that for every $i < j$, if $\bar{\phi}(\boldsymbol{X}_i, \mathcal{S}_{\max}) = \bar{\phi}(\boldsymbol{X}_j, \mathcal{S}_{\max})$, then $\sigma_i \geq \sigma_j$. In this regard, to prove (72) one only needs to run the following procedure:

Marking algorithm Let the integer numbers between n_0 and n to be either *marked* or *unmarked*. Also, assume all numbers $\{n_0, \ldots, n\}$ are unmarked in the beginning, and then increase the index i from n_0 to n. For each i, if $\sigma_i = -1$ define j as the largest unmarked index such that $n_0 \leq j < i$, and $\sigma_j = +1$. If there are no index j which satisfies these conditions, then just pass to the next i. Otherwise, match σ_i with σ_j and set both i and j as marked.

Note that according to the above algorithm, each matched pair contribute a negative value to J. Let n^* be the largest index satisfying the conditions: $n^* > n_0$ and $\sigma_{n^*} = -1$. If there is no index with these conditions, set $n^* = n_0 - 1$. Then, the following facts hold for n^*:

- For all matched pairs (σ_i, σ_j), we have either $\min(i,j) < n^*$ or $n^* < \max(i,j)$.
- For all unmarked $i \leq n^*$, we have $\sigma_i = -1$.
- For all unmarked $i > n^*$, we have $\sigma_i = +1$.

Next, let us define the set $\mathcal{S}^*_{\max} \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}^K \middle| \; d_{\mathcal{S}_{\max}}(\boldsymbol{x}) \leq \varepsilon \triangleq d_{\mathcal{S}_{\max}}(\boldsymbol{X}_{n^*}) \right\}$. Surprisingly, and based on the definition of the planar distance in Section 2, $\mathcal{S}^*_{\max}$ is also a simplex in $\mathbb{S}_K$. Similar to the notation introduced in Section 2, let $(\boldsymbol{w}_k, b_k)$ represent the parameters associated to the hyper-plane which encompasses the kth facet of $\mathcal{S}_{\max}$. This way, $\boldsymbol{w}_k \in \mathbb{R}^K$ (with $\|\boldsymbol{w}_k\|_2 = 1$) denotes the normal outward vector associated to the kth facet, while $b_k \in \mathbb{R}$ represents the corresponding bias value. Then, it can be easily seen that $\mathcal{S}^*_{\max}$ is also a simplex whose hyper-plane parameters are $(\boldsymbol{w}_k, b_k - \varepsilon)$, i.e. for each k, the kth facet is ε-shifted toward its outward normal vector $\boldsymbol{w}_k$. Therefore, and based on the definition of $J_{(\text{ext})}$, we can write:

$$J_{(\text{ext})} \geq \frac{1}{n} \sum_{i=1}^{n} \sigma_i f\left(\bar{\phi}\left(\boldsymbol{X}_i, \mathcal{S}^*_{\max}\right)\right). \tag{74}$$

Thanks to the definition of $\mathcal{S}^*_{\max}$, we can now investigate the contribution of each data sample to both J and at least a lower-bound of $J_{(\text{ext})}$.

- Any unmatched data point $\boldsymbol{X}_i$ with $n_0 \leq i \leq n^*$, contributes negatively to J, while its contribution to $J_{(\text{ext})}$ is at least zero.
- Any matched pair contribute to J non-positively, while their contribution to $J_{(\text{ext})}$ is at least zero.
- For any unmatched data point $\boldsymbol{X}_i$ with $i > n^*$, contribution to J is at most 1, while contribution to $J_{(\text{ext})}$ is exactly 1.

In all the above-mentioned cases, contribution of each $\boldsymbol{X}_i$ to J is less than or equal to that of $J_{(\text{ext})}$ which proves (72). A direct consequence of these arguments is the following:

$$\mathfrak{R}_n\{\Phi\} \leq \mathfrak{R}_n\{\mathscr{C}\}, \tag{75}$$

since the expectation of $J_{(\text{ext})}$ w.r.t. $\boldsymbol{\sigma}$ and data points $\boldsymbol{X}_i$ is in fact the Rademacher complexity of $\mathscr{C}$. For a function class such as $\mathscr{C}$, we can bound the Rademacher complexity via a number of previously-established tools and measures, e.g. VC-dimension and Massart's lemma. However, we first need to bound the *growth functions* of $\mathcal{C}$, which has been already carried away in Lemma 3. In this regard, according to Massart's lemma (see, for example, Theorem 3.3 of [10]), and also using (72), we have

$$\mathfrak{R}_n\{\Phi\} \leq \sqrt{\frac{2\log \Pi_{\mathscr{C}}(n)}{n}}, \tag{76}$$

Substitution from Lemma 3 into the above inequality completes the proof. □

Proof of Lemma 6. Lemma roughly states that whenever the overlap between a K-simplex $\mathcal{S}$ and $\mathcal{S}_T$ is strictly smaller than $1-\epsilon$, then with high probability $\mathcal{S}$ would not cover all the training data points when n becomes sufficiently large. Thus, the expected value of cost function increases accordingly. First, let us take a closer look at $\mathbb{E}_{\mathcal{S}_T}\hat{R}_{\text{CRR}}$ and analyze its minimum over $\mathcal{S} \in \mathcal{Q}^c(\epsilon)$:

$$\min_{\mathcal{S}\in\mathcal{Q}^c(\epsilon)} \mathbb{E}_{\mathcal{S}_T}\hat{R}_{\text{CRR}}(\mathcal{S}) = \min_{\mathcal{S}\in\mathcal{Q}^c(\epsilon)} \left\{\sqrt{n}\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}_{\mathcal{S}_T}}\left[\ell\left(d_{\mathcal{S}}(\boldsymbol{X})\right)\right] + \gamma \text{Vol}(\mathcal{S})\right\}, \tag{77}$$

where throughout the proof we refer to $\text{Vol}(\mathcal{S})$ simply as V for the sake of simplicity.

The planar distance $d_{\mathcal{S}}(\boldsymbol{X})$ is zero when $\boldsymbol{X} \in \mathcal{S} \cap \mathcal{S}_T$. Thus, any nonzero term in the above expectation corresponds to a point in the differential set $\mathcal{S}_T - \mathcal{S}$, which due to the minimization

over $\mathcal{S} \in \mathcal{Q}^c(\epsilon)$, has a Lebesgue measure of at least ϵV_T. In this regard, let us define the following sets:

$$\mathcal{A}_k = \mathcal{A}_k(\mathcal{S}, \mathcal{S}_T) \triangleq \left\{ \boldsymbol{X} \in \mathcal{S}_T - \mathcal{S} \;\middle|\; d_{\mathcal{S}}(\boldsymbol{X}) = \boldsymbol{w}_k^T \boldsymbol{X} + b_k \right\}, \quad \forall k = 0, 1, \ldots, K, \tag{78}$$

where, similar to Section 2, $(\boldsymbol{w}_k, b_k)$ represent the hyper-plane which encompasses the kth facet of $\mathcal{S}$. In other words, $\mathcal{A}_k$ shows the set of points inside $\mathcal{S}_T$ and outside of $\mathcal{S}$, which have a smaller planar distance to the kth facet of $\mathcal{S}$ compared to its other facets. For each point that has an equal planar distance from more than one facet, k can be chosen as any of those facets, arbitrarily. In this regard, it is easy to see that

$$\mathcal{A}_k \cap \mathcal{A}_l = \emptyset \;\Leftrightarrow\; k \neq l \quad , \quad \bigcup_{k=0}^{K} \mathcal{A}_k = \mathcal{S}_T - \mathcal{S}. \tag{79}$$

Also, let

$$\alpha_k \triangleq \frac{\mathrm{Vol}(\mathcal{A}_k)}{\sum_{l=0}^{K} \mathrm{Vol}(\mathcal{A}_l)} \;,\; k = 0, 1, \ldots, K, \tag{80}$$

with $\sum_k \alpha_k = 1$. Consider the parameteric sets $\mathcal{B}_\delta$, $\epsilon \leq \delta \leq 1$, defined as

$$\mathcal{B}_\delta \triangleq \left\{ \mathcal{S} \in \mathcal{Q}^c(\epsilon) \;\middle|\; \frac{\mathrm{Vol}(\mathcal{S} \cap \mathcal{S}_T)}{V_T} = 1 - \delta \right\}. \tag{81}$$

Obviously, we have $\bigcup_{\epsilon \leq \delta \leq 1} \mathcal{B}_\delta = \mathcal{Q}^c(\epsilon)$. Then, one can show that the following chain of relations hold for all $\epsilon \leq \delta \leq 1$:

$$\begin{aligned}
\min_{\mathcal{S} \in \mathcal{B}_\delta} \left\{ \mathbb{E}_{\mathcal{S}_T} \hat{R}_{\mathrm{CRR}}(\mathcal{S}) - \gamma V \right\} &= \sqrt{n} \min_{\mathcal{S} \in \mathcal{B}_\delta} \mathbb{E}_{\mathcal{S}_T} \left[\ell\left(d_{\mathcal{S}}(\boldsymbol{X}) \right) \right] \\
&\geq \delta \sqrt{n} \min_{\mathcal{S} \in \mathcal{B}_\delta} \; \sum_{k=0}^{K} \alpha_k \mathbb{E}_{\mathcal{A}_k} \left[\ell\left(d_{\mathcal{S}}(\boldsymbol{X}) \right) \right] + (1 - \delta) \sqrt{n} \ell(0) \\
&\geq \delta \sqrt{n} \min_{\boldsymbol{\alpha} \in \boldsymbol{\Phi}} \; \sum_{k=0}^{K} \left(\frac{\alpha_k}{\Delta_k} \int_0^{\Delta_k} \ell(u) \, \mathrm{d}u \right) + (1 - \delta) \sqrt{n} \ell(0) \\
&= \delta \sqrt{n} \min_{\boldsymbol{\alpha} \in \boldsymbol{\Phi}} \; \sum_{k=0}^{K} \alpha_k L(\Delta_k) + \sqrt{n} \ell(0),
\end{aligned} \tag{82}$$

where $\boldsymbol{\Phi}$ denotes the standard simplex, or the set of all $(K+1)$-dimensional discrete probability mass functions. $L(x) \triangleq \frac{1}{x} \int_0^x (\ell(u) - \ell(0)) \, \mathrm{d}u$ for $x \geq 0$, and $\Delta_k \triangleq \mathrm{Vol}(\mathcal{A}_k) / (\max_k \; \mathrm{Vol}(\mathcal{S}_{-k}))$. The main intuition behind the above inequalities is the fact that the term

$$\mathbb{E}_{\mathcal{A}_k} \left[\ell\left(d_{\mathcal{S}}(\boldsymbol{X}) \right) \right] \tag{83}$$

is minimized when $\mathcal{A}_k$ is concentrated in the form of a thin cylindrical structure above the largest facet of $\mathcal{S}$. The volume of the largest facet is $\max_k \; \mathrm{Vol}(\mathcal{S}_{-k})$, which makes Δ_k to be the height of this imaginary cylinder. We have also taken advantage of the fact that ℓ is increasing and integrable. Based on the definition, we have $\mathrm{Vol}(\mathcal{A}_k) = \delta \alpha_k V_T$. Also, due to the $(\underline{\lambda}, \bar{\lambda})$-isoperimetricity of $\mathcal{S}_T$, we have

$$\max_k \; \mathrm{Vol}(\mathcal{S}_{-k}) \leq \bar{\lambda} K V_T^{\frac{K-1}{K}} \quad \text{which implies} \quad \Delta_k \geq \left(\frac{\delta \alpha_k}{\bar{\lambda}} \right) V_T^{1/K}. \tag{84}$$

Also, note that since $\ell(\cdot)$ is assumed to be increasing, $L(\cdot)$ is both increasing and convex. Therefore, the minimization over $\boldsymbol{\alpha} \in \boldsymbol{\Phi}$ in (82) results in $\boldsymbol{\alpha}^* = \left(\frac{1}{K+1}, \cdots, \frac{1}{K+1}\right)$. Substitution into (82) yields the following lower-bound for the expected risk minimized over $\mathcal{B}_\delta$:

$$\begin{aligned}\min_{\mathcal{S}\in\mathcal{B}_\delta} \left\{\mathbb{E}_{\mathcal{S}_T} \hat{R}_{\mathrm{CRR}}(\mathcal{S})\right\} &\geq \delta\sqrt{n} L\left(\frac{\delta V_T^{1/K}}{(K+1)\,\bar{\lambda}}\right) + \gamma \min_{\mathcal{S}\in\mathcal{B}_\delta} \mathrm{Vol}(\mathcal{S}) + \sqrt{n}\ell(0) \\ &\geq \delta\sqrt{n} L\left(\frac{\delta V_T^{1/K}}{(K+1)\,\bar{\lambda}}\right) - \gamma V_T \delta + \hat{R}_{\mathrm{CRR}}(\mathcal{S}_T), \end{aligned} \tag{85}$$

where we have used the following simple facts: $\min_{\mathcal{S}\in\mathcal{B}_\delta} \mathrm{Vol}(\mathcal{S}) = V_T(1-\delta)$, and also $\hat{R}_{\mathrm{CRR}}(\mathcal{S}_T) = \sqrt{n}\ell(0) + \gamma V_T$. The rest of the proof is straightforward. Finding a lower-bound for $\mathbb{E}_{\mathcal{S}_T}\hat{R}_{\mathrm{CRR}}(\mathcal{S})$ when $\mathcal{S} \in \mathcal{Q}^c(\epsilon)$ can be accomplished through the following set of relations and inequalities:

$$\begin{aligned}\min_{\mathcal{S}\in\mathcal{Q}^c(\epsilon)} \left\{\mathbb{E}_{\mathcal{S}_T} \hat{R}_{\mathrm{CRR}}(\mathcal{S})\right\} &= \min_{\epsilon\leq\delta\leq 1} \min_{\mathcal{S}\in\mathcal{B}_\delta} \mathbb{E}_{\mathcal{S}_T} \hat{R}_{\mathrm{CRR}}(\mathcal{S}) \\ &\geq \min_{\epsilon\leq\delta\leq 1} \left(\delta\sqrt{n} L\left(\frac{\delta V_T^{1/K}}{(K+1)\,\bar{\lambda}}\right) - \gamma V_T \delta\right) + \hat{R}_{\mathrm{CRR}}(\mathcal{S}_T). \end{aligned} \tag{86}$$

Again, remember that $L(\cdot)$ is an increasing function. Therefore, given that the derivative of the r.h.s. of the above inequality remains positive w.r.t. δ, and for all $\epsilon \leq \delta \leq 1$, then the minimum occurs at $\delta^* = \epsilon$. It can be easily verified that given the condition on n in the lemma, the derivative remains positive and thus the proof is complete. □

Lemma 7 (Restricted Isoperimetry)**.** *For $K \in \mathbb{N}$, assume $\mathcal{S}_{\mathrm{reg}} \in \mathbb{S}_K$ to be a perfectly regular simplex with equal side lines. Then for all K, $\mathcal{S}_{\mathrm{reg}}$ is $(\underline{\lambda}, \bar{\lambda})$-isoperimetric where $\underline{\lambda}$ and $\bar{\lambda}$ can be chosen to be as small as 1 and e, respectively.*

Proof. For a perfectly regular simplex $\mathcal{S}_{\mathrm{reg}}(\boldsymbol{\Theta})$ with all side lines equal to $d \geq 0$, i.e. $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'}\|_2 = d,\ \forall k \neq k'$, the Lebesgue measure $\mathrm{Vol}(\mathcal{S}_{\mathrm{reg}})$ is already shown to be

$$\mathrm{Vol}(\mathcal{S}_{\mathrm{reg}}) = \frac{\sqrt{K+1}}{K!\sqrt{2^K}} d^K. \tag{87}$$

Thus, we have

$$\max_{k,k'} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'}\|_2 = d = \left(\frac{K!\sqrt{2^K}}{\sqrt{K+1}}\right)^{1/K} \mathrm{Vol}^{\frac{1}{K}}(\mathcal{S}_{\mathrm{reg}}). \tag{88}$$

On the other hand, for all $K \in \mathbb{N}$ we have the following upper-bound for $K!$: $K! \leq Ke\left(\frac{K}{e}\right)^K$, which leads us to the following relations:

$$\max_{k,k'} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'}\|_2 \leq \left[\left(\frac{Ke}{\sqrt{K+1}}\right)^{1/K} \frac{\sqrt{2}}{e}\right] K\mathrm{Vol}(\mathcal{S}_{\mathrm{reg}})^{1/K} \leq K\mathrm{Vol}(\mathcal{S}_{\mathrm{reg}})^{1/K}, \tag{89}$$

and means $\underline{\lambda}$ can be chosen to be as small as 1.

For the other claim, we should note that the maximal facet of $\mathcal{S}_{\mathrm{reg}}$ (noting the fact that for a perfectly regular simplex all facets are equal), can be attained by computing the volume of a

perfectly regular $(K-1)$-simplex with all side lines equal to d. In other words:

$$\max_k \ \mathrm{Vol}\left(\mathcal{S}_{\mathrm{reg}_{-k}}\right) = \frac{\sqrt{K}d^{K-1}}{(K-1)!\sqrt{2^{K-1}}} = \frac{\sqrt{K2^{1-K}}}{(K-1)!}\left(\frac{K!\sqrt{2^K}}{\sqrt{K+1}}\right)^{\frac{K-1}{K}} \mathrm{Vol}\left(\mathcal{S}_{\mathrm{reg}}\right)^{\frac{K-1}{K}}. \tag{90}$$

which, again by using the previously-mentioned upper-bound on $K!$, can be bounded as

$$\begin{aligned}
\max_k \ \mathrm{Vol}\left(\mathcal{S}_{\mathrm{reg}_{-k}}\right) &= \left[\sqrt{\frac{K}{(K+1)^{\frac{K-1}{K}}}}\,(K!)^{-1/K}\right] K\mathrm{Vol}\left(\mathcal{S}_{\mathrm{reg}}\right)^{\frac{K-1}{K}} \\
&\leq \left[e\sqrt{(K/e^2)^{-1/K}}\right]\mathrm{Vol}\left(\mathcal{S}_{\mathrm{reg}}\right)^{\frac{K-1}{K}} \\
&\leq e\mathrm{Vol}\left(\mathcal{S}_{\mathrm{reg}}\right)^{\frac{K-1}{K}}, \qquad (91)
\end{aligned}$$

Again, it means $\bar{\lambda}$ can be chosen to be as small as e. This completes the proof. □