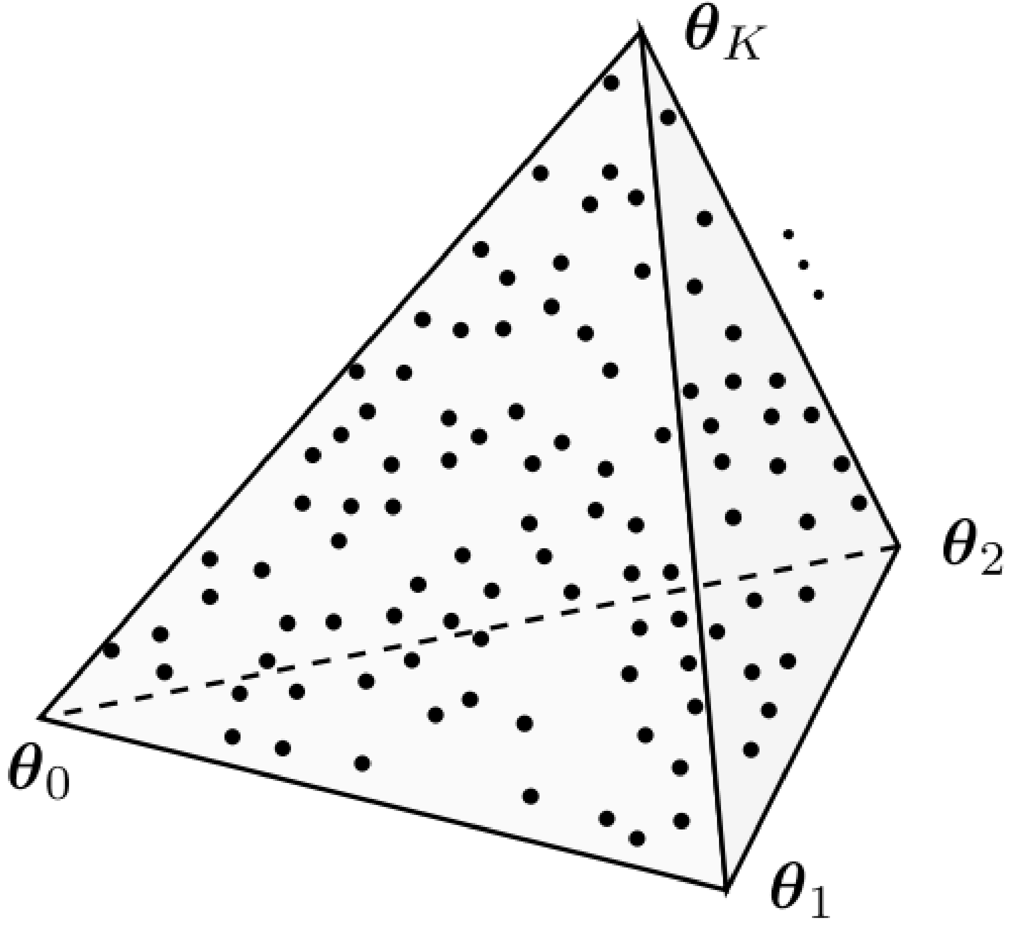
ON STATISTICAL LEARNING OF SIMPLICES UNMIXING PROBLEM REVISITED

Gwangwoo Kim

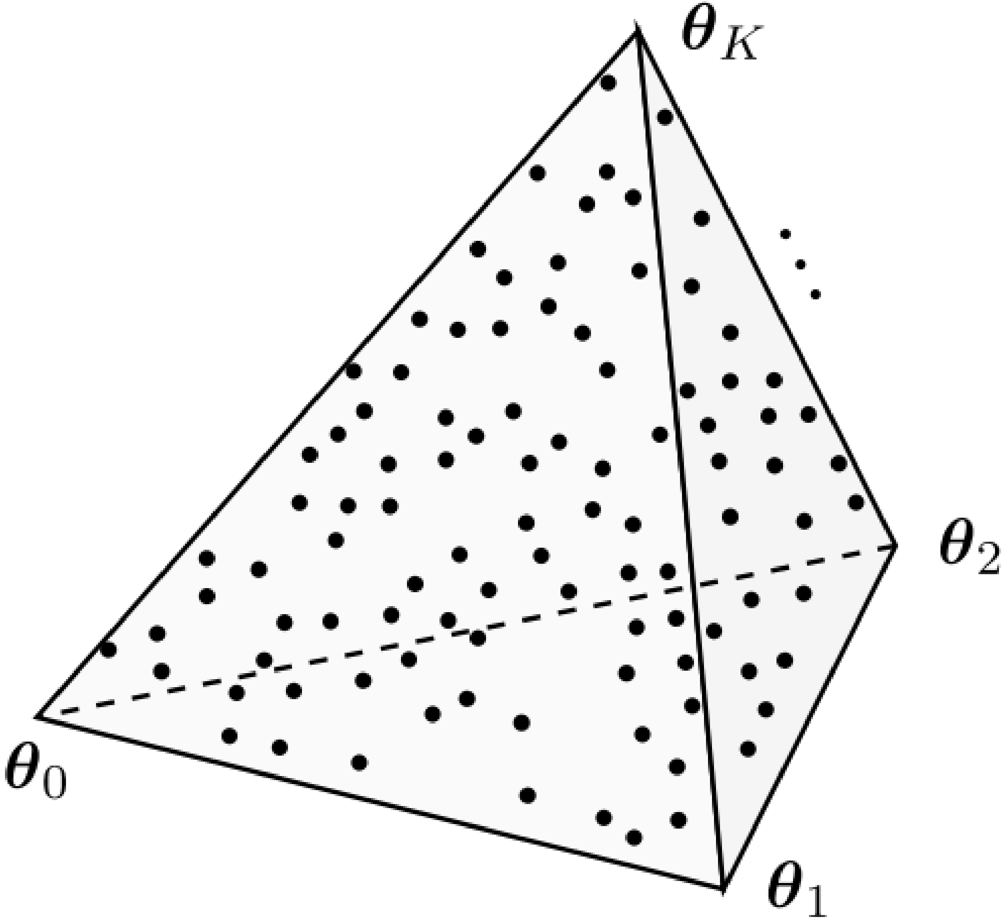
Steptember 26, 2021

Contents

1. Introduction



1. Introduction

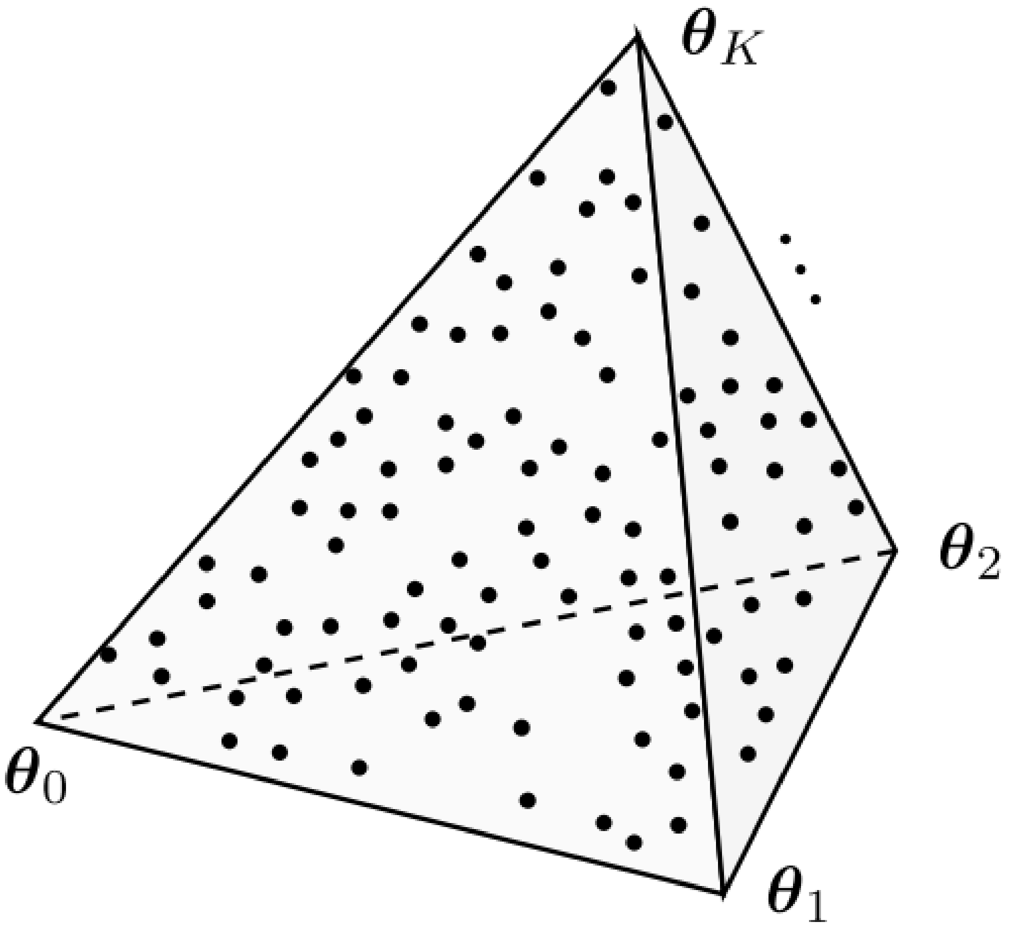


This problem can be formulated by

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

But it is not yet a statistical problem.

1. Introduction



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

Assume p_i are generated from a uniform Dirichlet distribution.

Then our transformed problem is find an estimator for Θ, say $\widehat{\Theta}$, such that the uniform probability measures over the simplices specified by Θ and $\widehat{\Theta}$ have a total variation distance of at most ϵ with probability at least $1 - \zeta$, for any given $\epsilon, \zeta > 0$.

1. Introduction

$$n \geq O(K^{22}/\epsilon^2)$$

Existing bound (2012)

Independent Component Analysis (ICA)

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

Proposed (2021)

heuristic Gradient Descent

2. Preliminaries

$$\boldsymbol{\Phi} \triangleq \{\boldsymbol{p} \in \mathbb{R}^{K+1} | \textstyle\sum_k p_k = 1,\, p_k \geq 0\}$$

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

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

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

2. Preliminaries

Assume $X_1, ..., X_n \in \mathbb{R}^K$ to be n i.i.d. samples which are generated uniformly from $\mathcal{S}_T \in \mathbb{S}_K$, that is, $X_1, ..., X_n \sim \mathbb{P}_{\mathcal{S}_T}$.

The problem is to find an approximation of $\mathcal{S}_T$, denoted by $\mathcal{S}^*$, from the dataset $D = \{X_1, ..., X_n\}$ 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. Method and results

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

<Notes>

- If S does not contain some points of D, the likelihood would be -∞.
- The estimator requires the smallest simplex in terms of volume.
- As mentioned, the transformed problem is to find the MLE.
- **The likelihood is not continuous (hence, not differentiable).**
- Moreover, finding MLE is NP-hard.

3. Method and results

THEOREM 3.1 (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 $\mathcal{D}_{\mathrm{TV}}(\mathbb{P}_{\mathcal{S}_T}, \mathbb{P}_{\mathcal{S}^*_{\mathrm{ML}}}) \leq \epsilon$.

<Remark> 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 the true simplex.

3. Method and results

By the computational hardness of MLE, we should replace the objective function with a continuously relaxed surrogate. Let's reformulate our problem.

$$\mathcal{S}^*_{\mathrm{ML}} = \underset{\mathcal{S}\in\mathbb{S}_K}{\operatorname{argmin}} \quad \mathrm{Vol}(\mathcal{S})$$

$$\text{subject to} \quad d_{\mathcal{S}}(\boldsymbol{X}_i) = 0, \forall i.$$

3. Method and results

DEFINITION 3.1 (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 $\hat{R}_{\text{CRR}} : \mathbb{S}_K \to \mathbb{R}$ is defined as

$$\hat{R}_{\text{CRR}}(\mathcal{S}; \boldsymbol{D}, \gamma, \ell) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell(d_{\mathcal{S}}(\boldsymbol{X}_i)) + \gamma \operatorname{Vol}(\mathcal{S}). \tag{3.3}$$

Also, let us define

$$\mathcal{S}^* = \mathcal{S}^*(\boldsymbol{D}, \gamma, \ell) \triangleq \underset{\mathcal{S} \in \mathbb{S}_K}{\operatorname{argmin}}\, \hat{R}_{\text{CRR}}(\mathcal{S}; \boldsymbol{D}, \gamma, \ell), \tag{3.4}$$

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

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

3. Method and results

THEOREM 3.2 (Sample complexity for general ℓ). *Assume a K-simplex $\mathcal{S}_T$ with Lebesgue measure $V_T \triangleq \mathrm{Vol}(\mathcal{S}_T)$, which is $(\underline{\lambda}, \bar{\lambda})$-isoperimetric for some $\underline{\lambda}, \bar{\lambda} \geq 0$. Also, assume $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ to be n i.i.d. samples drawn from $\mathbb{P}_{\mathcal{S}_T}$. Assume for $\zeta, \epsilon > 0$, the following condition holds for n:*

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

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 (3.4).

The proof follows from Vapnik–Chervonenkis (VC) theory of statistical learning.

3. Method and results

COROLLARY 3.1 (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 (3.4), denoted by $\mathcal{S}^*$, satisfies the inequality $\mathcal{D}_{\mathrm{TV}}(\mathbb{P}_{\mathcal{S}_T}, \mathbb{P}_{\mathcal{S}^*}) \leq \epsilon$.

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

3. Method and results

COROLLARY 3.1 (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 (3.4), denoted by $\mathcal{S}^*$, satisfies the inequality $\mathcal{D}_{\mathrm{TV}}(\mathbb{P}_{\mathcal{S}_T}, \mathbb{P}_{\mathcal{S}^*}) \leq \epsilon$.

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

3. Method and results

THEOREM 3.3 (Gradient of the planar distance).

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

where $\boldsymbol{G}_i \triangleq \nabla_{\boldsymbol{\Theta}} \ell(d_{\mathcal{S}(\boldsymbol{\Theta})}(\boldsymbol{X}_i))$

Algorithm 1 Learning of simplices via gradient descent

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

4. Experimental results

- We define error of two simplices by

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

- The initialization of the algorithm is not important, but such a convex hull lead to a substantially faster convergence.

- when γ is chosen to be relatively low, which means the objective function of the Algorithm becomes more similar to that of MLE.

4. Experimental results

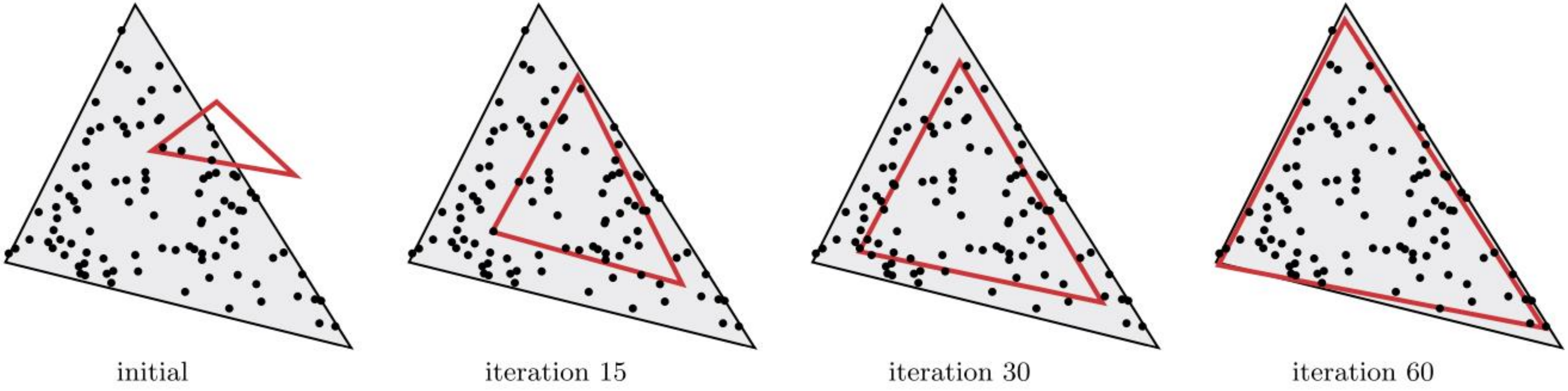


FIG. 3. *Snapshots from running Algorithm 1 on a set of $n = 100$ noiseless samples drawn uniformly from a two-dimensional simplex. The original triangle is drawn in black and the outputs of the proposed method for four different iteration steps are shown in red.*

4. Experimental results

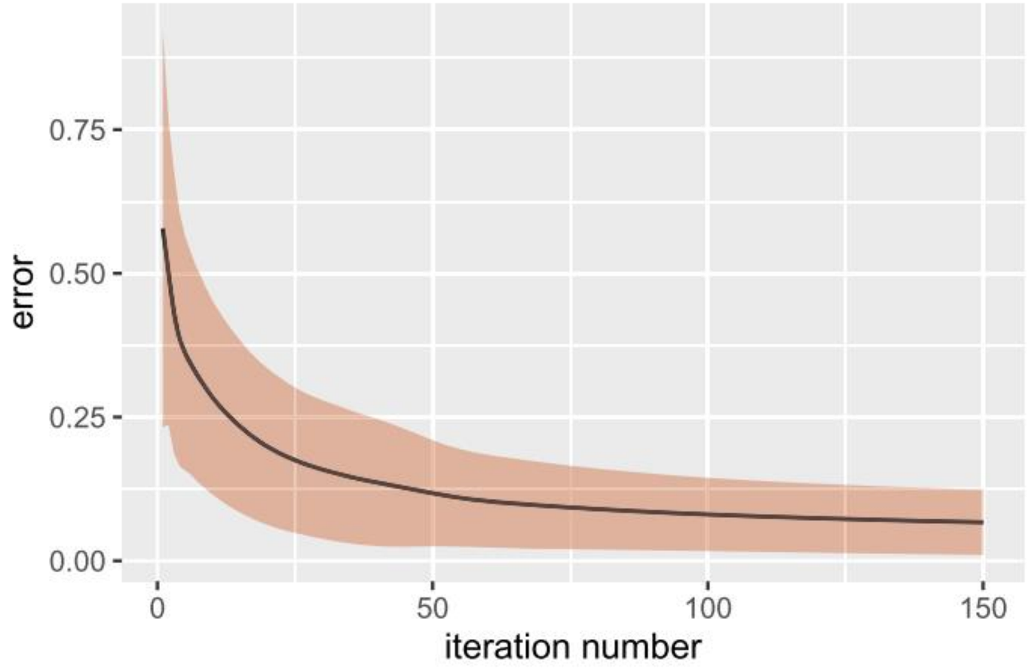
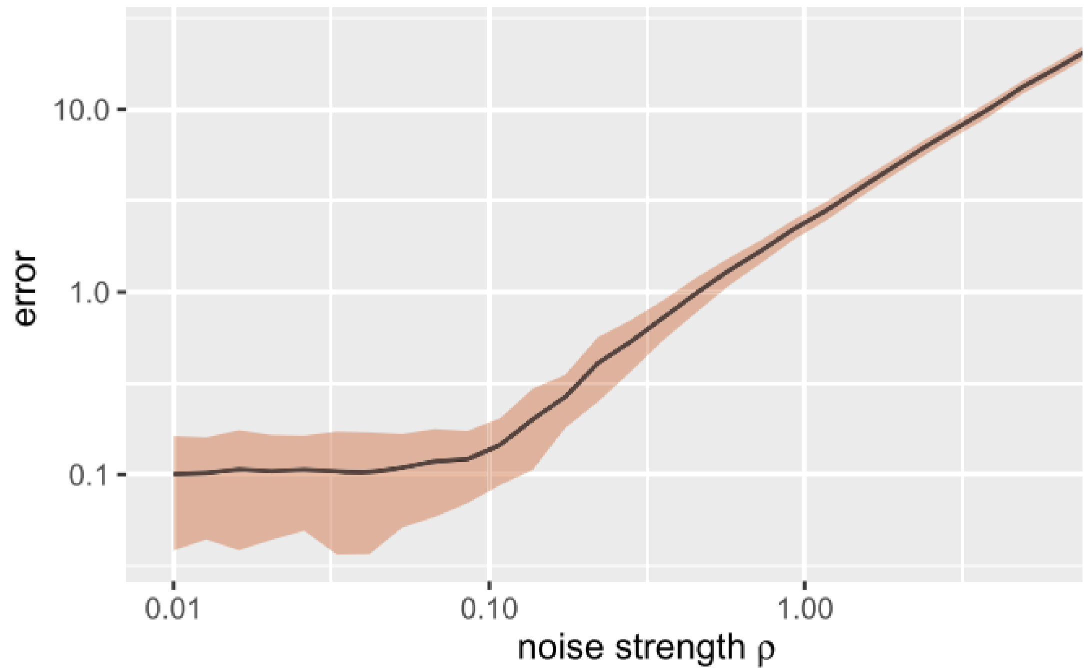
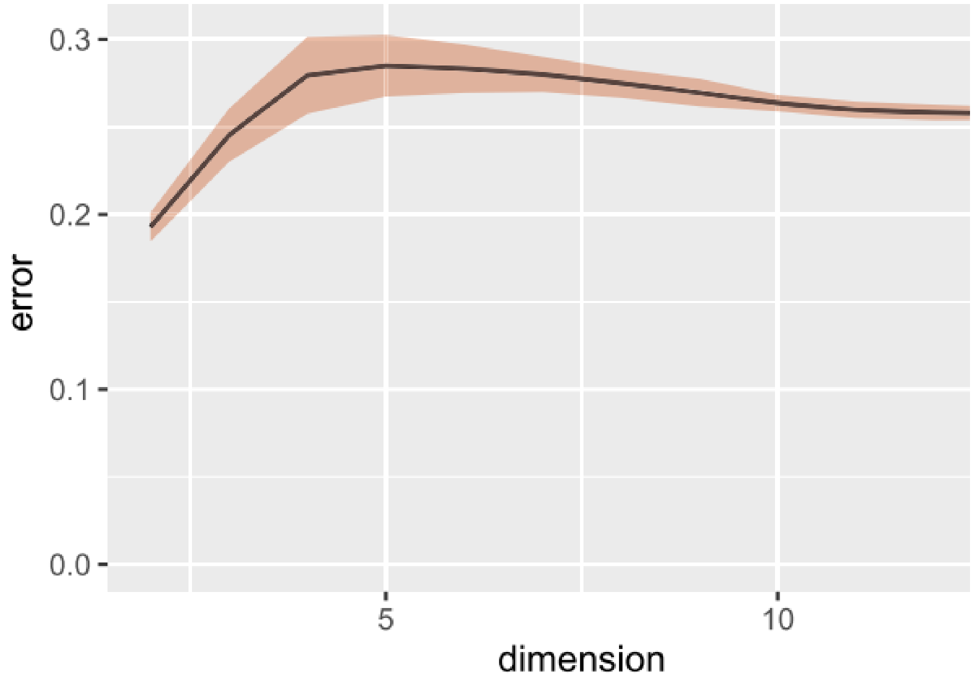


FIG. 4. *Depiction of error in (4.1) as a function of iteration number for Algorithm 1. The experiment has been performed on $n = 100$ data points uniformly sampled from a two-dimensional simplex. Parameters γ and optimization step α have been adjusted to optimize the performance. According to the curve, sample mean and the standard deviation of error decay as the number of iterations is increased.*

4. Experimental results



(a)



(b)

FIG. 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.*

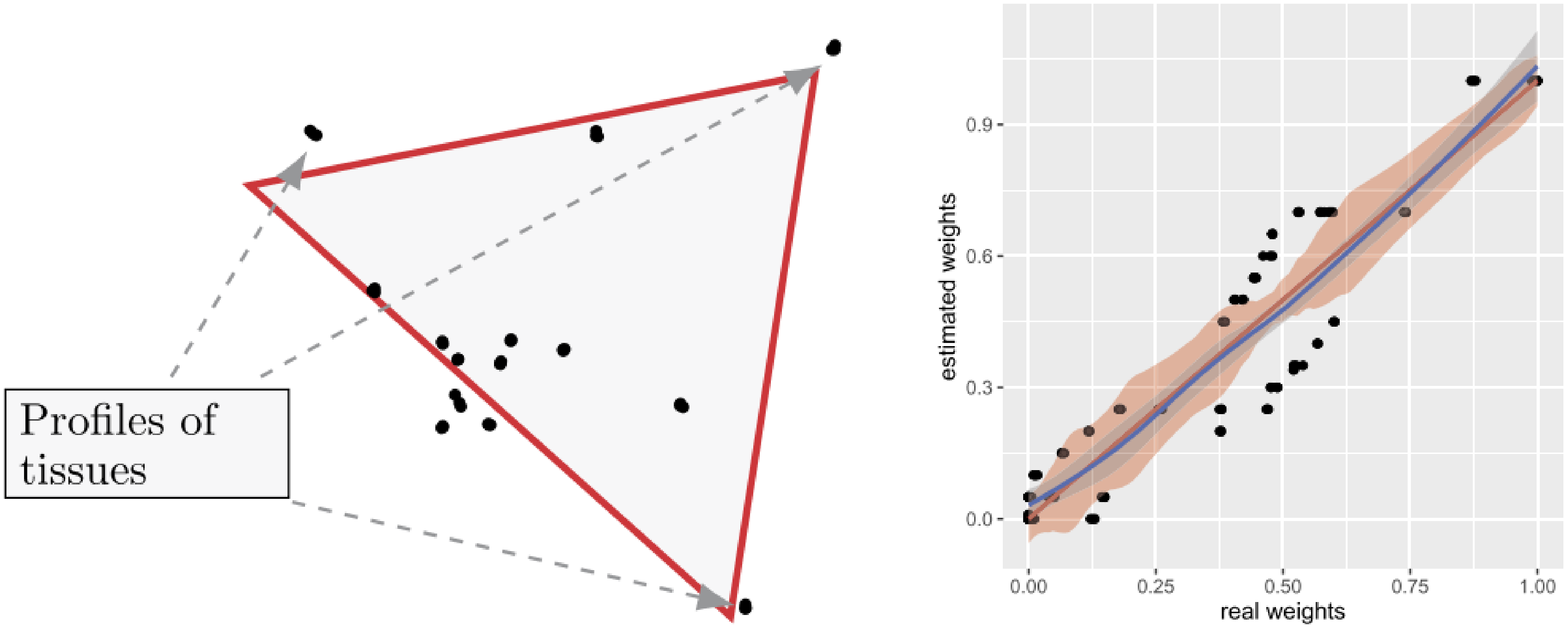
4. Experimental results

TABLE 1

Comparison of the proposed method with MVSA [22], *SISAL* [7], *VCA* [26] *and UNMIX* [33]. *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*

	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	–

4. Experimental results



(a) (b)

FIG. 6. *Cell-type identification from micro-array data given in* [32]. 6*a*: *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.* 6*b*: *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.*

5. References

1. Anderson, J., Goyal, N., & Rademacher, L. (2013, June). Efficient learning of simplices. In *Conference on Learning Theory* (pp. 1020-1045). PMLR.

2. Najafi, A., Ilchi, S., Saberi, A. H., Motahari, S. A., Khalaj, B. H., & Rabiee, H. R. (2021). On statistical learning of simplices: Unmixing problem revisited. *The Annals of Statistics*, *49*(3), 1626-1655.