Towards upper bounding the variance of the demixing estimate of a geometry based Independent Component Analysis algorithm - Technical Report

Thomas Pitschel

Braunschweig: Institut für Mathematische Stochastik, 2014

Towards upper bounding the variance of the demixing estimate of a geometry based Independent Component Analysis algorithm - Technical Report

Thomas Pitschel
April 19th, 2014

***Abstract*—For the Independent Component Analysis (ICA) algorithm (using a geometric approach) given in [1], we provide a probabilistic upper bound on the Euclidean error of constituents of the demixing matrix estimate resulting from unmixed input, under the condition that the algorithm converges to the global optimum, and derive an asymptotic bound on the expected squared error for large sample sizes from it. This error bound is found to be of order $\mathcal{O}\left(N^{-2+2\beta}\right)$, where N is the sample size and $\beta \in (0,1)$ is chosen arbitrarily. The result can be used to establish the asymptotic behaviour of the variance of the demixing matrix estimate for arbitrarily linearly mixed independent inputs.**

***Index Terms*—independent component analysis, cylindrical hoof, n-dimensional ball**

I. Introduction

Given N observations $x_i \in \mathbb{R}^d$, $i = 1 \ldots N$, presumed to originate from a discrete-time process x whose values obey $x = As$, where A is a fixed real-valued $d \times d$-matrix, and s is a d-dimensional real-valued source process whose d elements are presumed to be statistically independent, the problem of (linear) independent component analysis is stated as recovering the sources s_i, $i = 1 \ldots N$ and the mixing matrix A such that $x_i \approx As_i$, where the proximity can be measured by an application-suitable norm. Various algorithms have been proposed, specifically in the 1990s (see [2] [3] [4]), for its solution (an older survey e.g. in [5]). It is immediately clear that a solution is determined only up to scaling and permutation of the recovered signal components. Another limitation refers to the (statistically inherent) fact that at most one Gaussian source component can be recovered. Accounting for this indeterminacy, upper bounds on the variances of the demixing estimates of various algorithms, e.g. [6], and the Cramer Rao Lower Bound (CRLB) for the ICA problem in general [6] (with [7]), have been stated.

In [1], a new algorithm for ICA based on a geometric approach has been given. As constituent of the solution, the algorithm determines for a given set of data points optimally bounding hyperplanes for each coordinate direction. In the following, we first give an analysis of the error of the deviation of the normal vector of the optimally bounding hyperplane belonging to a single coordinate direction from its expected value, then expand the result to upper bounding the error of the estimate of the demixing matrix W. The statements are effective for analysis of the actually incurred error *under the condition* that - in loose words - the algorithm indeed converges (from its initial estimate) towards the proximity of the globally optimal solution.

The exposition is organized as follows: The next section revisits the workings of the algorithm and describes the link of the presented statements with it in more detail. In section III, the main results are developed, where emphasis is given to charting the broad route of argument rather than to presenting the strongest results. Finally, a conclusion is stated.

II. Algorithm and relation to presented statements

This section first shortly revisits the algorithm description. Given N samples $X_i \in \mathbb{R}^d$, $i = 1..N$, supposed to originate from sampling a linear mixing of d arbitrarily distributed mutually independent random variables, the algorithm [1] proceeds by alternatingly determining a current best estimate of the demixing matrix and applying it to the latest demixing result to obtain a new demixing result. Within a single iteration, first the margin distributions for each of the coordinates are removed from the current demixing result (yielding a multivariate distribution in $[-1,1]^d$) and then an optimal parallelepiped (consisting of optimally bounding hyperplanes for each coordinate direction) is determined from this empirical distribution.

For independent and non-Gaussian original random variables, it is reasonable to assume that this iterative procedure will – with high probability – eventually yield a demixing result in which all original independent components are present and separated (the proof of which is not given here and left to another exposition), with possibly permuted order and yet undetermined scale. (In order to fix a definite algorithm output, the demixing estimate, i.e. the demixing matrix W, is deemed normalized such that all rows have unit Euclidean norm.)

Eventually reaching the globally optimal solution (in a so far unspecified number of iterations) means that the accuracy of the algorithm is then determined only by the statistical properties of the data in relation to an algorithm iteration *in the vicinity* of the globally optimal solution. We will call the error of the demixing matrix estimate therefore a *steady state* error, hinting at the condition of completed algorithmic convergence. Further employing the property that the algorithm output obeys an invariance with respect to the choice of the initial demixing estimate, one concludes that it is essentially sufficient to examine only the case where the mixing matrix is the identity. This case will be called *standard setting* in the following.

With similar argument, within this standard setting also the operation of removing the margin distributions from the input is expected to be eventually separable from the main analysis and thus is omitted here.

III. Main results

This section analyzes the steady state error of the demixing estimate in the standard setting.

Throughout this section and remainder, we will denote by w^T the transpose of a vector w. Without further mentioning, d will denote a fixed dimension[1], and we will write C^d for the hypercube $[-1,1]^d$. We will denote by (w,c) the hyperplane given by normal vector w and offset c. For a vector p and a coordinate j ($j \in \{1,..,d\}$), we will also write $p_{\setminus j}$ for $(p_1,\dots,p_{j-1},p_{j+1},\dots,p_d)^T$.

Throughout, we will take the (unindexed) notation $\|\cdot\|$ to mean the Euclidean norm, and will write $\|\cdot\|_\infty$ when meaning the supremum norm. For matrix A, $\operatorname{tr}(A)$ denotes the trace of A. For a set B of points in $\mathbb{R}^d$, we will write $\operatorname{int}(B)$ to mean the interior of B (i.e. the largest open subset of B), based on a metric induced by any of these norms.

[1]Some results will be stated for even d only, in order to avoid cases. It is easy to assemble the results also for d odd.

One of the main results will be best expressed using a modified probabilistic Landau-O notation:

Definition Let $(X_n)_{n\in\mathbb{N}}$ be a family of real-valued random variables defined (respectively) on probability spaces with measures P_n. Let $(a_n)_{n\in\mathbb{N}}$ a real-valued positive sequence. We write

$$X_n = \tilde{\mathcal{O}}_P\left(a_n\right) \text{ as } n \longrightarrow \infty$$

if

$$\forall \epsilon > 0 : \exists M, N_0 : \forall n \geq N_0 : P_n\left(\frac{|X_n|}{a_n} > M\right) < \epsilon.$$

The difference to the standard probabilistic Landau-O notation (see for example [8]) is the inclusion of the sequence index offset N_0, and the two definitions infact coincide if P_n is independent of n.

Lemma III.1. Let $X_i, i = 1..N$, be independently distributed in C^d, let j be a coordinate direction, i.e. $j \in \{1,\dots,d\}$. Let the optimal hyperplane in direction j be given by (p, c_j), i.e. it holds that $p^T X_i \leq c_j$ for all i, $p_j = 1$ and $c_j \geq 0$ is minimal. Denote $\tilde{p}$ for $p_{\setminus j}$. Let $\epsilon > 0$ with $\epsilon \leq \epsilon_{max}$. Then

$$P(\|\tilde{p}\| \geq C_1 \cdot \epsilon) \leq (1-\epsilon)^N =: \vartheta(\epsilon, N)$$

where for even d

$$C_1 = (d/2)! \cdot \left(\tfrac{4}{\pi}\right)^{d/2}$$
$$\epsilon_{max} := 2 \cdot C_1^{-1}.$$

Proof The proof of the lemma proceeds in three steps (following a purely geometrical argumentation): first we link the orientation of the normal vector of the optimal hyperplane to a region within the cube free of sample points, then we find a subregion with known volume, namely a cylindrical hoof properly oriented, and finally use this volume to give a probability bound.

For a given sample $\mathcal{X} := \{X_i, i = 1..N\}$ and a coordinate direction choice j, the optimal hyperplane (p, c_j) is uniquely defined except from cases which aggregated have probability mass zero. In the following we write $p(\mathcal{X})$ and $c_j(\mathcal{X})$ to emphasize the dependence on the sample.

With $d \times d$-matrix $(S_{1,d})_{i,k} = \delta_{i,k} + h \cdot \delta_{i,d}\delta_{k,1}$ and an $h > 0$, denote by $G_{\mathrm{H},\infty}(h)$ the set of points $\{S_{1,d}(h)x \,|\, x \in \mathbb{R}^d, \|x_{\setminus d}\|_\infty \leq 1, x_d \leq 0\}$, and by $G_{\mathrm{H},2}(h)$ the set $\{S_{1,d}(h)x \,|\, x \in \mathbb{R}^d, \|x_{\setminus d}\|_2 \leq 1, x_d \leq 0\}$, i.e. $G_{\mathrm{H},\infty}(h)$ and $G_{\mathrm{H},2}(h)$ are the cubical resp. cylindrical hoof of height h, sloped along direction x_1, and extending infinitely in negative d-th coordinate direction. We will call these sets *standard hoofs*. It is shown in the appendix (Proposition

A.1) that, given the sample $\mathcal{X}$ and corresponding optimal hyperplane (p, c_j), there exists a matrix $A = VA'$ such that V is permutation matrix, A' is orthogonal, $A'e_d = e_d$ and $p^T AG_{\mathrm{H},\infty}(\|\tilde{p}\|) \leq 0$, i.e. a standard hoof properly sloped can be rotated so as to completely be located in one half-space of the hyperplane described by $(p, 0)$. By suitable arbitration we can uniquely choose this matrix for the given sample $\mathcal{X}$. We will refer to this choice as $A(\mathcal{X})$. Defining $G'_2(j, A, h) := e_j - AG_{\mathrm{H},2}(h)$, which corresponds to the rotated mirrored and translated version of a standard cylindrical hoof, it is then easy to understand that $p^T G'_2(j, A(\mathcal{X}), \|\tilde{p}(\mathcal{X})\|) \geq 1$. In the following, we will–for abbreviation–further set $G''_2(j, A, h) := \mathrm{int}(G'_2(j, A, h)) \cap C^d$.

With an (initially) arbitrary $\gamma > 0$, with $\gamma \leq 2$, we now have

$$\begin{aligned}
&P(\|\tilde{p}(\mathcal{X})\| \geq \gamma) \\
&\quad \leq P(\|\tilde{p}(\mathcal{X})\| \geq \gamma, \\
&\qquad \forall i : X_i \notin \{x \,|\, p(\mathcal{X})^T x > c_j(\mathcal{X})\} \cap C^d) \\
&\quad \leq P(\|\tilde{p}(\mathcal{X})\| \geq \gamma, \\
&\qquad \forall i : X_i \notin G''_2(j, A(\mathcal{X}), \|\tilde{p}(\mathcal{X})\|)) \\
&\quad \leq P(\forall i : X_i \notin G''_2(j, A(\mathcal{X}), \gamma))
\end{aligned}$$

Partitioning the event according to equal values for the rotation matrix $A(\mathcal{X})$, this is further expressed as

$$\begin{aligned}
&P(\forall i : X_i \notin G''_2(j, A(\mathcal{X}), \gamma)) \\
&\quad = \int P(\forall i : X_i \notin G''_2(j, A(\mathcal{X}), \gamma) \,|\, A(\mathcal{X}) = A') \\
&\qquad \mathrm{d}P_{A(\cdot)}(A') \\
&\quad = \int P(\forall i : X_i \notin G''_2(j, A', \gamma)) \,\mathrm{d}P_{A(\cdot)}(A')
\end{aligned}$$

Using the fact that the probability for the sample points being located outside of G''_2 does not depend on the rotation matrix A', and using a result about the volume of a d-dimensional cylindrical hoof (appendix B), we finally yield

$$\begin{aligned}
&P(\forall i : X_i \notin G''_2(j, A(\mathcal{X}), \gamma)) \\
&\quad = P(\forall i : X_i \notin G''_2(j, I, \gamma)) \cdot \int \mathrm{d}P_{A(\cdot)}(A') \\
&\quad = P(\forall i : X_i \notin G''_2(j, I, \gamma)) \\
&\quad = (1 - \epsilon)^N
\end{aligned}$$

where the last equation holds by defining γ such that (for d even)

$$\epsilon = \gamma \cdot \frac{\pi^{d/2}}{(d/2)! \cdot 2^d}.$$

Solving for γ completes the proof. □

An immediate conclusion from the Lemma is a result about the rate of convergence to zero of the normal vector norm as N tends to infinity, expressed with modified probabilistic Landau-O notation. We have:

Corollary III.2. Let $X_i, i = 1..N$, coordinate direction j, (p, c_j) and C_1 be given as in Lemma III.1. Let $1 > \beta > 0$. Then

$$\|\tilde{p}\| = \tilde{\mathcal{O}}_P\left(N^{-1+\beta}\right).$$

Proof To ensure the probability bound in Lemma III.1 converges to zero, it is clear ϵ must decrease not too fast (in N). We choose $\epsilon := N^{-1+\beta}$. Then it is easy to see that $(1-\epsilon)^N \leq \exp(-N^\beta)$ for all $N \in \mathbb{N}$. Thus

$$P(\|\tilde{p}\| \geq C_1 \cdot N^{-1+\beta}) \leq \exp(-N^\beta)$$

for all N such that $C_1 \cdot N^{-1+\beta} \leq 2$, from which the result follows. □

While it would be natural to further pursue a deterministic asymptotic result for the expectation of $\|\tilde{p}\|$, a problem arises in bounding the norm in the seldom event. (First note that the algorithm perfectly allows for yielding a p such that $\|\tilde{p}\|$ exceeds any given bound. Second, to bound the vector p artificially is not possible since the true demixing matrix is not known in the non-standard setting.) We *can* give such a result however for $\|\tilde{w}\|$, where w is the L_2-normalized hyperplane normal vector:

Theorem III.3. Let $X_i, i = 1..N$, coordinate direction j, (p, c_j) be given as in Lemma III.1. Let $1 > \beta > 0$. Let $w := p/\|p\|$, and denote again $\tilde{w} := w_{\setminus j}$. Then for $a \in \mathbb{N}$

$$\mathrm{E}(\|\tilde{w}\|^a) = \mathcal{O}\left(N^{a(-1+\beta)}\right).$$

Proof (of Theorem III.3): Since $p = w/w_j$ and $|w_j| \leq 1$, it clearly follows that $\|\tilde{w}\| \geq C$ for some C implies $\|\tilde{p}\| = \|\tilde{w}\|/|w_j| \geq C$. Thus

$$P(\|\tilde{w}\| \geq C_1 \cdot \epsilon) \leq P(\|\tilde{p}\| \geq C_1 \cdot \epsilon).$$

Since $\|\tilde{w}\|$ is also bounded by one, we obtain for sufficiently large N, using Lemma III.1,

$$\mathrm{E}(\|\tilde{w}\|^a) \leq 1 \cdot \exp(-N^\beta) + C_1^a \cdot N^{a(-1+\beta)} \cdot 1,$$

from which again the result follows. □

We now proceed towards a result about the variance of the normalized demixing matrix estimate. For this note that the expectation of this matrix is not exactly, but still asymptotically (in N), the identity matrix.

Theorem III.4. Let $X_i, i = 1..N$, be given as in Lemma III.1, and for the coordinate directions $j = 1..d$, let $(p^{[j]}, c_j)$ be the corresponding optimal hyperplanes, i.e. for all j, $(p^{[j]})^T X_i \leq c_j$ for all i, $p_j^{[j]} = 1$, $c_j \geq 0$ minimal. Let $w^{[j]} = p^{[j]}/\|p^{[j]}\|$, and denote by W the matrix formed by assembling the $(w^{[j]})^T$ in rows. Then, for $1 > \beta > 0$,

$$1 - \mathrm{E}(w_j^{[j]}) = \mathcal{O}\left(N^{-2+2\beta}\right) \quad \forall j$$
$$\mathrm{E}(w_i^{[j]}) = 0 \quad \forall i \neq j$$
$$\mathrm{E}(\|W - \mathrm{E}(W)\|^2) = \mathcal{O}\left(N^{-2+2\beta}\right),$$

as $N \longrightarrow \infty$.

Proof Clearly $0 < w_j^{[j]} \leq 1$. Observing $p^{[j]} = w^{[j]}/w_j^{[j]}$ we have $w_j^{[j]} = 1/\|p^{[j]}\|$. Omitting superscript $[j]$, this is $w_j = 1/\|p\| = 1/\sqrt{1 + \|\tilde{p}\|^2}$. Using Lemma III.1 with $\epsilon = N^{-1+\beta}$, we conclude, with $B(N) := 1/\sqrt{1 + C_1^2 N^{-2+2\beta}}$,

$$P(w_j \leq B(N)) \leq \exp(-N^\beta).$$

In the seldom event, w_j is bounded below; thus

$$\begin{aligned}\mathrm{E}(w_j) &\geq 0 \cdot P(w_j \leq B(N)) + B(N) \cdot P(w_j > B(N)) \\ &= B(N) \cdot (1 - \exp(-N^\beta))\end{aligned}$$

and therefore, for all N such that $C_1 \cdot N^{-1+\beta} \leq 2$,

$$\begin{aligned}1 - \mathrm{E}(w_j) &\leq 1 - B(N) + B(N)\exp(-N^\beta) \\ &\leq C_2 \cdot N^{-2+2\beta} + \exp(-N^\beta), \quad (*)\end{aligned}$$

with suitable C_2, from which the first result follows. We have $\mathrm{E}(w_i^{[j]}) = 0$ for $i \neq j$ for symmetry reason. Finally,

$$\begin{aligned}&\mathrm{E}(\|W - \mathrm{E}(W)\|^2) \\ &= \mathrm{E}(\mathrm{tr}\left((W - \mathrm{E}(W))^T (W - \mathrm{E}(W))\right)) \\ &= \mathrm{E}(\sum_j \|\tilde{w}^{[j]}\|^2) + \mathrm{E}(\sum_j (w_j^{[j]} - \mathrm{E}(w_j^{[j]}))^2)\end{aligned}$$

The first summand is of order $\mathcal{O}\left(N^{-2+2\beta}\right)$ by applying the argument given in Theorem III.3 simultaneously to the vectors $\tilde{w}^{[j]}$. As already hinted by the convergence of the expectation of $w_j^{[j]}$, the second summand is of order $\mathcal{O}\left(N^{-4+4\beta}\right)$ (elaborated in appendix C), completing the proof. □

IV. Conclusion

An analysis of the steady state error of the ICA algorithm [1] (using a geometric approach) has been given - deriving both an elementary probabilistic and an (in expectation) asymptotic result - for the case where mutually independent variables are provided as input to the algorithm. An outline has been sketched how the statements relate to that error for the case of arbitrary inputs, and thus eventually to the variance of the estimator. The result presents a theoretical limit that the algorithm attains if the convergence to the globally optimal solution can be assured for a particular implementation of the algorithm. The proof of this global convergence remains open to examination.

An obvious point for further improvement is sharpening the statement for use with finite sample sizes. Concretely, the constant C_1 in Lemma III.1 could be considerably improved, in terms of the order in d (in lieu of the fact that the volume of a unit d-ball decreases exponentially in d), by rather bounding the volume of what would be called $G''_\infty(j, A, \gamma)$, i.e. the volume of the halfspace intersected with C^d actually being free of sample points.

Another remaining question regards the asymptotics of the expectation of $\|\tilde{p}\|$. Establising a corresponding result could be helpful for analysing modified versions of the algorithm.

References

[1] Thomas Pitschel. Independent component analysis using a geometric approach. Digitale Bibliothek Braunschweig, 2013.

[2] A. J. Bell and T. J. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, pages 757–763. MIT Press, 1996.

[4] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[5] Aapo Hyvärinen. Survey on independent component analysis. http://cis.legacy.ics.tkk.fi/aapo/papers/NCS99web/.

[6] Petr Tichavský, Zbyněk Koldovský, and Erkki Oja. Performance analysis of the FastICA algorithm and cramérrao bounds for linear independent component analysis. *IEEE Trans. on Signal Processing*, 54(4), 2006.

[7] Petr Tichavský, Zbyněk Koldovský, and Erkki Oja. Corrections to 'performance analysis of the FastICA algorithm and cramér-rao bounds for linear independent component analysis' tsp 04/06. *IEEE Trans. on Signal Processing*, 56(4):1715–1716, 2008.

[8] Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete multivariate analysis - Theory and Practice*. Springer Verlag, 2007.

[9] Eric W. Weisstein. Cylindrical wedge. Retrieved from MathWorld (Wolfram Web Resource) 2014-03-15. http://mathworld.wolfram.com/CylindricalWedge.html, 1948.

[10] Equation 5.19.4, NIST Digital Library of Mathematical Functions, Release 1.0.7 of 2014-03-21. http://dlmf.nist.gov/5.19.E4, 2014.

Appendix

A. Existence of rotation $A = VA'$

By allowing for application of the permutation matrix V, it is sufficient to establish a relation between a hyperplane normal vector and a point set representing a cubical hoof only for the case that the chosen coordinate direction j equals d. We will write A in this sub-section when referring to A'. For a vector v, we will denote $\tilde{v} := v_{\setminus d}$.

For $h > 0$, define $G_{H,\infty}(h) := \{S_{1,d}(h)x \mid x \in \mathbb{R}^d, x_d \leq 0, \|\tilde{x}\|_\infty \leq 1\}$, where $S_{1,d}(h)$ is a matrix $\in \mathbb{R}^{d\times d}$ which shifts the d-th component by $h \cdot x_1$, i.e. $(S_{1,d})_{i,k} = \delta_{i,k} + h \cdot \delta_{i,d}\delta_{k,1}$. Thus $G_{H,\infty}(h)$ describes a standard (cubical) hoof sloped along direction x_1 and infinitely extending towards negative d-th coordinate direction. For a given normal vector w of a hyperplane $(w, 0)$ with $w_d = 1$, we can give a relation to the standard hoof with slope $\|\tilde{w}\|$:

Proposition A.1. Let $w \in \mathbb{R}^d$ with $w_d = 1$. Set $h := \|\tilde{w}\|$. Then exists a matrix $A \in \mathbb{R}^{d\times d}$ such that $Ae_d = e_d$, A is orthogonal matrix and $w^T A G_{H,\infty}(h) \leq 0$, i.e. there exists a rotation matrix leaving d-th direction invariant and such that the rotated hoof is fully on one side of the hyperplane $(w, 0)$.

Proof Matrix A must be of form $A_{d,j} = \delta_{j,d}$ and therefore because of orthogonality $A_{i,d} = \delta_{i,d}$. The matrix must therefore have block form containing a $(d-1) \times (d-1)$ matrix $\tilde{A}$. Noting that the crucial face of the hoof to fulfill the remaining condition is the one with $x_d = 0$ (in the definition of $G_{H,\infty}$), it is then easy to check that choosing $\tilde{A}_{\cdot,1} = -\tilde{w}/\|\tilde{w}\|$ and all other columns arbitrary but orthogonal to the previous ones yields the desired matrix. (Choose $x = (x_1, 0, \dots)$.) If $\|\tilde{w}\| = 0$ then $\tilde{A} = I$ will yield a suitable choice for A. ☐

During usage of the proposition, we will make use of the uniqueness of the matrix A. As an examplary construction rule, successively choose the columns $i = 2, \dots, (d-1)$ of $\tilde{A}$ such that they are orthogonal to w and to all previously chosen ones, they are normal, and such that the i-th component is maximal. (Almost surely, this will define $\tilde{A}$ uniquely.)

B. Volume of a d-dimensional cylindrical hoof

This sub-section details the volume of a d-dimensional cylindrical hoof, i.e. a region formed by intersection of a cylinder (extending infinitely in one direction) with a half-space described by a hyperplane $(w, 0)$. The following will assume the cylinder has e_d as rotational axis and the plane being solely sloped with respect to coordinate x_1, i.e. $w^T = (w_1, 0, \dots, 0)$. The following mimicks the calculation given in [9].

With the notation $\tilde{x} := x_{\setminus d}$ and $\tilde{\tilde{x}} = \tilde{x}_{\setminus 1}$, the volume V of $G := \{x \mid \|\tilde{\tilde{x}}\| \leq R, w^T x \leq 0, w_1 = h/R, x_d \geq 0\}$, where R is the radius and h is the height of the hoof, is given by

$$\begin{aligned} V &= \int_{x_1\in[0,R]} \int_{\|\tilde{\tilde{x}}\|\leq\sqrt{R^2-x_1^2}} \frac{h}{R} \cdot x_1 \,\mathrm{d}\tilde{\tilde{x}} \,\mathrm{d}x_1 \\ &= \int_{x_1\in[0,R]} \frac{h}{R} \cdot x_1 \cdot V^B_{d-2}(1) \cdot \sqrt{R^2 - x_1^2}^{\,d-2} \,\mathrm{d}x_1 \\ &= -\frac{1}{2}\frac{h}{R} \cdot V^B_{d-2}(1) \cdot \frac{2}{d}\sqrt{R^2 - x_1^2}^{\,d}\Big|_{[0,R]} \\ &= \frac{h \cdot V^B_{d-2}(1)}{R \cdot d} \cdot R^d = \frac{h \cdot V^B_{d-2}(1)}{d} \cdot R^{d-1}, \end{aligned}$$

where $V^B_d(1)$ is the volume of the d-dimensional unit ball, given as [10]

$$\begin{aligned} V^B_d(1) &= \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} \\ &= \begin{cases} \frac{\pi^{d/2}}{(d/2)!} & d \text{ even} \\ \frac{2\left(\frac{d-1}{2}\right)!\cdot(4\pi)^{(d-1)/2}}{d!} & d \text{ odd} \end{cases}, \end{aligned}$$

with Γ being Euler's gamma function.

C. Bounding $(w_j - E(w_j))^2$

Writing w_j for $w_j^{[j]}$, we have

$$\begin{aligned} (w_j - \mathrm{E}(w_j))^2 \leq (w_j - 1)^2 + 2\,|w_j - 1| \cdot |1 - \mathrm{E}(w_j)| \\ + (1 - \mathrm{E}(w_j))^2. \end{aligned}$$

Using the argument given in proving Theorem III.4 (cf. $(*)$), for all N with $C_1 \cdot N^{-1+\beta} \leq 2$ and some sufficiently large finite $\tilde{C}$, we have

$$|1 - \mathrm{E}(w_j)| \leq \tilde{C} \cdot N^{-2+2\beta}.$$

Therefore, for this $\tilde{C}$ and such N,

$$\begin{aligned} P((w_j - \mathrm{E}(w_j))^2 \geq (C_2 \cdot N^{-2+2\beta})^2 \\ + 2C_2 \cdot N^{-2+2\beta} \cdot \tilde{C} \cdot N^{-2+2\beta} \\ + (\tilde{C} \cdot N^{-2+2\beta})^2) \\ \leq \exp(-N^\beta). \end{aligned}$$

Applying the argument simultaneously for all j, we obtain

$$\begin{aligned} P(\sum_j (w_j - \mathrm{E}(w_j))^2 \geq d \cdot (C_2 + \tilde{C})^2 \cdot N^{4(-1+\beta)}) \\ \leq d \cdot \exp(-N^\beta). \end{aligned}$$

Since the $(w_j - \mathrm{E}(w_j))^2$ are bounded (by one), it follows (as $N \longrightarrow \infty$)

$$\mathrm{E}(\sum_j (w_j - \mathrm{E}(w_j))^2) = \mathcal{O}\left(N^{4(-1+\beta)}\right).$$