Mixing and Non-Mixing Local Minima of the Entropy Contrast for Blind Source Separation

Frédéric Vrins, *Student Member, IEEE*, Dinh-Tuan Pham, *Member, IEEE*, and Michel Verleysen, *Senior Member, IEEE*

Abstract—In this paper, both non-mixing and mixing local minima of the entropy are analyzed from the viewpoint of blind source separation (BSS); they correspond respectively to acceptable and spurious solutions of the BSS problem. The contribution of this work is twofold. First, a Taylor development is used to show that the *exact* output entropy cost function has a non-mixing minimum when this output is proportional to *any* of the non-Gaussian sources, and not only when the output is proportional to the lowest entropic source. Second, in order to prove that mixing entropy minima exist when the source densities are strongly multimodal, an entropy approximator is proposed. The latter has the major advantage that an error bound can be provided. Even if this approximator (and the associated bound) is used here in the BSS context, it can be applied for estimating the entropy of any random variable with multimodal density.

Index Terms—Blind source separation (BSS), entropy estimation, independent component analysis, mixture distribution, multimodal densities.

I. INTRODUCTION

BLIND source separation (BSS) aims at recovering a vector of independent sources $\boldsymbol{S} = [S_1, \ldots, S_K]^T$ from observed mixtures $\boldsymbol{X} = [X_1, \ldots, X_M]^T$. In this paper, we assume that $K = M$ and $\boldsymbol{X} = \boldsymbol{AS}$, where $\boldsymbol{A}$ is the K-by-K mixing matrix. The sources can be recovered by finding an unmixing matrix $\boldsymbol{B}$ such that $\boldsymbol{W} = \boldsymbol{BA}$ is non-mixing (i.e., with one nonzero entry per row and per column). Such matrices $\boldsymbol{B}$ can be found by minimizing an *ad hoc* cost function (see [1], the books [2]–[4], and references therein).

In practice, the minimum of these criteria is reached by adaptive methods such as gradient descent. Therefore, one has to pay attention to the solutions corresponding to these minima. In most cases, the global minimum is a solution of the BSS problem. By contrast, the possible local minima can either correspond to a desired solution (referred as *non-mixing* minima) or spurious solution (referred as *mixing* minima) of the problem. For example, the optimization algorithm could be trapped in minima that do not correspond to an acceptable solution of the

BSS problem. Therefore, it is of interest to study the possible existence of both non-mixing and mixing local minima.

The paper deals with this issue by extending existing results of related work. The introduction first presents the two main approaches for source separation and details the state-of-the-art related to the local minima of BSS criteria. Then, the objectives and the organization of the paper is presented.

A. Symmetric and Deflation Approaches

To determine matrix $\boldsymbol{B}$, two approaches can be investigated. The first one (called *symmetric*) aims at extracting all sources simultaneously. The second approach (called *deflation*) extracts the sources one by one.

- The common symmetric approach consists in minimizing the Kullback–Leibler divergence between the joint density and the product of the marginal densities of the recovered sources (i.e., their mutual information), which are the components $Y_1, \ldots, Y_K$ of $\boldsymbol{Y} = \boldsymbol{BX}$. This leads to the minimization of (see [5]–[7])

$$C(\boldsymbol{B}) \triangleq \sum_{i=1}^{K} H(Y_k) - \log|\det \boldsymbol{B}| \qquad (1)$$

where $H(Y)$ denotes Shannon's differential entropy Y[5], [6]

$$H(Y) \triangleq -\int p_Y(y) \log(p_Y(y)) dy. \qquad (2)$$

In (2), p_Y denotes the probability density function (pdf) of Y. A variant of this approach applies the unmixing matrix $\boldsymbol{B}$ to a whitened version of the observations. In this case, since the sources are uncorrelated and can be assumed to have the same variance, one can constrain $\boldsymbol{B}$ to be orthogonal [2]. The term $\log \det \boldsymbol{B}$ in criterion (1) disappears and $C(\boldsymbol{B})$ is to be minimized over the group of orthogonal matrices.

- The deflation approach [8] extracts the kth source by computing the kth row $\boldsymbol{b}_k$ of $\boldsymbol{B}$ by minimizing a non-Gaussianity index of $\boldsymbol{b}_k \boldsymbol{X}$ subject to the constraint that $\boldsymbol{b}_k \boldsymbol{X}$ is uncorrelated to $\boldsymbol{b}_i \boldsymbol{X}$ for $i < k$. By taking this index to be the negentropy [9] and assuming (without loss of generality) that the sources have the same variance, the cost function can be written as $H(\boldsymbol{w}_k \boldsymbol{S}) - \log \|\boldsymbol{w}_k\|$ plus a constant, where $\boldsymbol{w}_k = \boldsymbol{b}_k \boldsymbol{A}$ and $\|\boldsymbol{w}_k\|$ denotes the Euclidean norm $\sqrt{\boldsymbol{w}_k \boldsymbol{w}_k^T}$[10], [11]. Since this function is unchanged when $\boldsymbol{w}_k$ is multiplied by a scalar, this leads to minimizing $H(\boldsymbol{w}_k \boldsymbol{S})$ under the $\boldsymbol{w}_i \boldsymbol{w}_k^T = \delta_{i,k}$ constraint for $1 \leq i$, $k \leq K$, where $\delta_{i,k}$ is the Kronecker delta [12].

B. Related Works

Although both symmetric and deflation procedures could be analyzed in this contribution with the same tools, we focus on the entropy $H(Y_k)$, used in the deflation approach.

Several results exist regarding the entropy minima of $Y = wS$ (the subscript "k" has been omitted in the following, since one signal is extracted at a time in the deflation approach). The first kind of results discusses the existence of non-mixing local minima of $H(Y)$ that correspond to the extraction of a single source. The second kind of results discusses the existence of mixing minima that correspond to spurious solutions of the BSS problem: Y is still a mixture of sources despite the fact that $H(Y)$ is a local minimum. These results are summarized below.

- *Non-mixing entropy local minima*
  It has been shown that the global minimum of $H(Y)$ with $Y = wS$ is reached when the output Y is proportional to the source with the lowest entropy [10]. It is proven in [9] that when a fixed-variance output is proportional to one of the sources, then, under some technical conditions, the cumulant-based approximation of entropy $H_J(Y)$ used in FastICA [9] reaches a non-mixing local minimum. Finally, based on the entropy power inequality [13], it is also proven in [14] that, in the two-dimensional case, Shannon's entropy has a local minimum when the output is proportional to a non-Gaussian source.

- *Mixing entropy local minima*
  As for the mutual information, simulations results in [15] suggest that mixing local entropy minima exist in specific cases (i.e., when the source pdfs are strongly multimodal, which sometimes occurs in practice, for sinusoid waveforms among others). These results, based on density estimation using the Parzen kernel method, are confirmed by other simulations using directly entropy estimation, such as Vasicek's one in [16] or based on the approximator analyzed in this paper in [17]. Rigorously speaking, the above results do not constitute an absolute proof since error bounds are not available for the approximation procedure. By contrast, a theoretical proof is given in [18], but for a specific example only (two bimodal sources sharing the same symmetric pdf). The existence of mixing local entropy minima has also been shown in [19] (without detailed proof) in the case of two nonsymmetric sources with strongly multimodal pdfs.

C. Objectives and Organization of the Paper

In this paper, additional results regarding mixing and non-mixing entropy minima are presented. Two main results will be proven.

First, it will be shown in Section II that the exact entropy of an output $H(Y)$ with a fixed variance has local non-mixing minima: the entropy $H(Y)$ has a local minimum when Y is proportional to one of the non-Gaussian sources. This is an extension of the results presented in [18] to the case of $K > 2$ sources. If the output is proportional to the Gaussian source (if it exists), the entropy has a global maximum. Numerical simulations illustrate these results in the $K = 2$ case, for the ease of illustration.

Second, in Section III, an entropy approximator is presented, for which an error bound can be derived. It is suitable for variables having multimodal densities with modes having a low overlap, in the sense that its error bound converges to zero when the mode overlap becomes negligible. This approximator was mentioned in [17] and error bounds have been provided in [19] without proof. In the BSS context, when the sources have such densities, the use of this approximator makes it possible to show that the marginal entropy has local mixing minima. This approach can be applied to a wider class of source densities than the score function-based method derived in [18]. The results presented in this paper further extend those in [19] as they are not restricted to the case of $K = 2$ sources. Finally, we provide a detailed proof of the bound formula for the entropy approximator.

It must be stressed that the aforementioned entropy approximator can be used for other applications that require entropy estimation of multimodal densities.

II. LOCAL NON-MIXING MINIMA OF OUTPUT ENTROPY

In this section, we shall prove that $H(wS)$, under the $\|w\| = 1$ constraint, reaches a local minimum at $w = I_j$, the jth row of the $K \times K$ identity matrix, if S_j is non-Gaussian, or a global maximum otherwise. Note that, as is well known, the global minimum is reached at I_k where $k = \arg\min_k H(S_k)$.

A. Theoretic Development

The starting point is an expansion of the entropy of a random variable Y slightly contaminated with another variable δY up to second order in δY, which has been established in [20]:

$$H(Y + \delta Y) \approx H(Y) + \mathrm{E}[\psi_Y(Y)\delta Y] \\ + \frac{1}{2}\left\{\mathrm{E}\left[\mathrm{var}(\delta Y|Y)\psi_Y'(Y)\right] - \left[\mathrm{E}(\delta Y|Y)\right]'^2\right\}. \quad (3)$$

In this equation, ψ_Y is the score function of Y, defined as[1] $-(\log p_Y)'$, p_Y is the pdf of Y, $'$ denotes the derivative, and $\mathrm{E}(\cdot|Y)$ and $\mathrm{var}(\cdot|Y)$ denote the conditional expectation and conditional variance given Y, respectively.

Assume that w is close from I_j so that its ith component w_i is close to 0 for $i \neq j$. Under the $\|w\| = 1$ constraint, $w_j = \sqrt{1 - \sum_{i \neq j} w_i^2}$ and since $\sqrt{1-x} = 1 - \frac{1}{2}x + o(x)$, one can write

$$w_j = 1 - \frac{1}{2}\sum_{i \neq j} w_i^2 + o\left(\sum_{i \neq j} w_i^2\right).$$

Thus, $wS = S_j + \delta S_j$ with

$$\delta S_j = \sum_{i \neq j} w_i S_i - \frac{1}{2}\left(\sum_{i \neq j} w_i^2\right)S_j + o\left(\sum_{i \neq j} w_i^2\right).$$

[1]In this paper, we use the score function definition presented in [7]. However, several authors define this function with the opposite sign. The reader should have this difference in mind.

Therefore, applying (3) and dropping higher order terms, one gets that $H(\boldsymbol{wS})$ equals

$$
\begin{aligned}
H(S_j) &+ \left(\sum_{i \neq j} w_i\right) \mathrm{E}[\psi_{S_j}(S_j)S_i] \\
&- \frac{1}{2}\left(\sum_{i \neq j} w_i^2\right) \mathrm{E}[\psi_{S_j}(S_j)S_j] \\
&+ \frac{1}{2}\left\{\mathrm{E}\left[\mathrm{var}\left(\sum_{i \neq j} w_i^2 S_i \Big| S_j\right) \psi'_{S_j}(S_j)\right]\right. \\
&\left.- \left[\sum_{i \neq j} w_i \mathrm{E}(S_i|S_j)\right]'^2\right\} + o\left(\sum_{i \neq j} w_i^2\right).
\end{aligned}
$$

Since the sources are mutually independent, any nonlinear mapping of them is uncorrelated so that $\mathrm{E}[\psi_{S_j}(S_j)S_i] = 0$, for $i \neq j$. Furthermore $\mathrm{E}(S_i|S_j) = \mathrm{E}(S_i) = 0$ for $i \neq j$, $\mathrm{E}[\psi_{S_j}(S_j)S_j] = 1$ (by integration by parts), and

$$
\mathrm{var}\left(\sum_{i \neq j} w_i S_i \Big| S_j\right) = \mathrm{var}\left(\sum_{i \neq j} w_i S_i\right) = \left(\sum_{i \neq j} w_i^2\right)\sigma_S^2
$$

where σ_S^2 denotes the common variance of the sources. Therefore

$$
H(\boldsymbol{wS}) = H(S_j) + \frac{1}{2}\left(\sum_{i \neq j} w_i^2\right)\left\{\sigma_S^2 \mathrm{E}\left[\psi'_{S_j}(S_j)\right] - 1\right\}
$$
$$
+ o\left(\sum_{i \neq j} w_i^2\right). \quad (4)
$$

Note that again by integration by parts, $\mathrm{E}[\psi'_{S_j}(S_j)]$ can be rewritten as $\mathrm{E}[\psi^2_{S_j}(S_j)]$, which is precisely Fisher's information [5]. In addition, by Schwarz's inequality [5], one has

$$
|\mathrm{E}\{[S_j - \mathrm{E}(S_j)]\psi_{S_j}(S_j)\}| \leq \sqrt{\sigma_S^2 \mathrm{E}\left[\psi^2_{S_j} S_j\right]}
$$

with equality if and only if ψ_{S_j} is a linear function. But since as mentioned earlier $\mathrm{E}[\psi_{S_j}(S_j)] = 0$ and $\mathrm{E}[S_j\psi_{S_j}(S_j)] = 1$, the left-hand side of the above inequality equals 1. Thus, $\sigma_S^2 \mathrm{E}[\psi^2_{S_j}(S_j)] > 1$ unless ψ_{S_j} is linear (which means that S_j is Gaussian) in which case $\sigma_S^2 \mathrm{E}\left[\psi^2_{S_j}(S_j)\right] = 1$. One concludes from (4) that $H(\boldsymbol{wS}) > H(S_j)$ for all $\boldsymbol{w}$ sufficiently close to $\boldsymbol{I}_j$ if S_j is non-Gaussian. Thus, $H(\boldsymbol{wS})$ reaches local non-mixing minima at $\boldsymbol{w} = \pm\boldsymbol{I}_j$ (since $H(-\boldsymbol{wS}) = H(\boldsymbol{wS})$), as long as S_j is non-Gaussian. If S_j is Gaussian, then $H(S_j)$ is a global maximum since Gaussian random variables have the highest entropy for a given variance. Equality (4) is of no use in this case, since the second term in this equality vanishes.

B. Numerical Simulations

In this subsection, three simple examples are analyzed in the $K = 2$ case. In this case, the unit-norm vector $\boldsymbol{w}$ can be rewritten

as $[\sin\theta, \cos\theta]$ and $H(\boldsymbol{wS})$ is considered as a function of θ. The entropy is computed through (2), in which the pdf were estimated from a finite sample set (1000 samples), using Parzen density estimation [21], [22] with Gaussian kernels of standard deviation $\sigma_K = 0.5\hat{\sigma}_X * S^{-1/5}$ (S denotes the number of samples and $\hat{\sigma}_X$ is the empirical standard deviation, enforced to be equal to one here) and Riemannian summation instead of exact integration.

Example 1: Assume that S_1 and S_2 have uniform densities. According to the above results, local minima exist for $\theta \in \{p\pi/2 | p \in \mathbb{Z}\}$. In this example, no mixing minimum can be observed (Fig. 1(a)).

Example 2: Suppose now that S_1 and S_2 have uniform and Gaussian distributions, respectively. Local minima are found for $\theta \in \{(2p+1)\pi/2\}$, $p \in \mathbb{Z}$, and local maxima for $\theta \in \{p\pi\}$ Fig. 1(b). Again, no spurious minimum can be observed in this example.

Example 3: Consider two source symmetric pdfs p_{s_1} and p_{s_2} that are constituted by i) two nonoverlapping uniform modes and ii) two Gaussian modes with negligible overlap, respectively. One can observe that non-mixing solutions occur for $\theta \in \{p\pi/2\}$ Fig. 1(c).

In addition to an illustration of the above theoretical result, the last example shows the existence of spurious (mixing) local minima for $\theta \notin \{p\pi/2\}$. However, the figure does not constitute a proof of the existence of local minima of $H(\boldsymbol{wS})$; the minima visible on the figure could indeed be a consequence of the entropy estimator (more precisely, of the pdf estimation). In Section III, we derive an entropy estimator and an associated error bound. This approximator is efficient for estimating the entropy of variables having multimodal densities, in the sense that the error bound tends to zero when the mode overlaps decrease. Next, thanks to this approximator, it will be theoretically proven that mixing local minima exist for strongly multimodal source densities.

III. ENTROPY APPROXIMATOR

In this section, we introduce the entropy approximator first derived in [17]. The detailed proofs of the upper and lower bounds of the entropy based on this approximator, already mentioned in [19] without proof, are given. Illustrative examples are further provided. The entropy bounds will be used in Section IV to prove that for a specific class of source distributions, the entropy function $H(\boldsymbol{wS})$ can have a local minimum that does not correspond to a row of the identity matrix. The presented approach yields more general results than those in [18], since it is no longer constrained that the sources share a common symmetric pdf.

This approach relies on an entropy approximation of a multimodal pdf of the form

$$
p(y) = \sum_{n=1}^{N} \pi_n K_n(y) \quad (5)
$$

where $N > 1$, $\pi_1, \ldots, \pi_N$ are (strictly positive) probabilities summing to 1, and $K_1, \ldots, K_N$ are unimodal pdfs. We focus
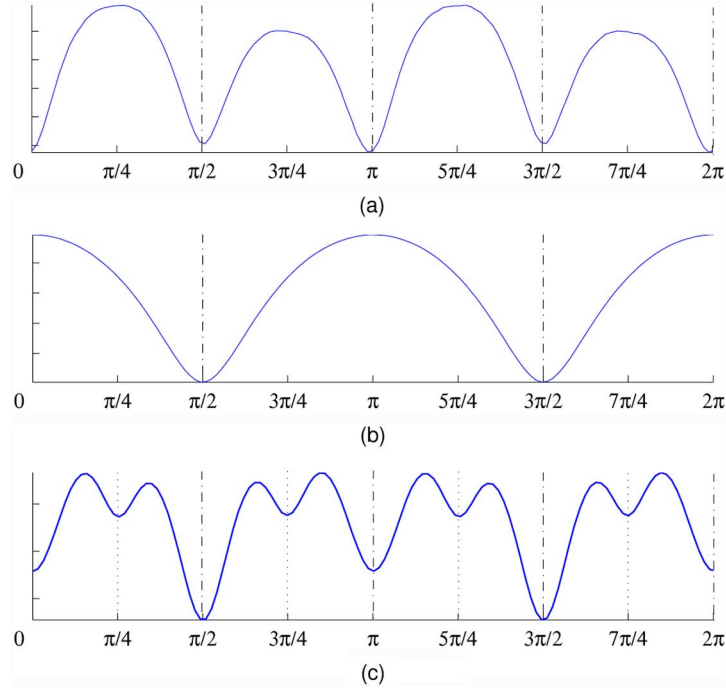
Fig. 1. Evolution of $H(\boldsymbol{w}\boldsymbol{S})$ versus θ. (a) Example 1: Two uniform sources. (b) Example 2: Uniform (S_1) and Gaussian (S_2) sources. (c) Example 3: Two bimodal sources. The non-mixing minima are indicated by dash-dotted vertical lines, the mixing ones by dotted lines.

on the case where the supports of the K_n can be nearly covered by disjoint subsets Ω_n $(n = 1, \ldots, N)$ so that p is strongly multimodal (with N modes). In this case, a good approximation to the entropy of a random variable of density p can be obtained; this entropy will be denoted (with abuse of notation) by $H(p)$ instead of $H(Y)$ where Y is a random variable with pdf p. Such approximation will be first derived informally (for ease of comprehension) and then a formal development giving the error bounds of the approximator is provided.

A. Informal Derivation of Entropy Approximator

If the random variable has a pdf of the form (5), then its entropy equals

$$H(p) = -\int_{-\infty}^{\infty} \sum_{n=1}^{N} \pi_n K_n(y) \log \left[\sum_{n=1}^{N} \pi_n K_n(y) \right] dy. \quad (6)$$

Suppose that there exists disjoint sets $\Omega_1, \ldots, \Omega_N$ that *nearly* cover the supports of the K_n densities; even if the K_n have a finite support, the Ω_n may differ from the true support of the K_i since these supports may be not disjoint. Then, assuming that $\pi_n K_n(y) \geq 0$ is small or zero for all $y \notin \Omega_n$ and noting that $0 \log 0 = 0$ by convention (more rigorously: $\lim_{x \to 0^+} x \log x = 0$), one gets

$$H(p) \approx -\sum_{m=1}^{N} \int_{\Omega_m} \sum_{n=1}^{N} \pi_n K_n(y) \log \left[\sum_{n=1}^{N} \pi_n K_n(y) \right] dy$$

$$\approx -\sum_{m=1}^{N} \pi_m \int_{\Omega_m} K_m(y) \log[\pi_m K_m(y)] dy.$$

If we note $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_n]$ and $h(\boldsymbol{\pi}) \triangleq -\sum_{n=1}^{N} \pi_n \log \pi_n$ the entropy of a discrete random variable taking N distinct values with probabilities $\pi_1, \ldots, \pi_N$, then $H(p) \approx \mathcal{H}(p)$ where

$$\mathcal{H}(p) \triangleq \sum_{n=1}^{N} \pi_n H(K_n) + h(\boldsymbol{\pi}). \quad (7)$$

B. Upper and Lower Bounds of the Entropy of a Multimodal Distribution

The entropy approximator $\mathcal{H}(p)$ in the preceding subsection is actually an upper bound for the entropy. This claim is proved in the following; in addition, a lower bound of the entropy will be further provided. These bounds permit to analyze how accurate is the approximation $H(p) \approx \mathcal{H}(p)$; they are explicitly computed when all K_n are Gaussian kernels.

1) General Results: The following lemma provides upper and lower bounds for the entropy.

Lemma 1: Let p be given by (5), then

$$H(p) \leq \mathcal{H}(p) \quad (8)$$

where $\mathcal{H}(p)$ is given by (7).

In addition, assume that

$$\sup K_n = \sup_{y \in \mathbb{R}} K_n(y) < \infty \qquad (1 \leq n \leq N)$$

and let $\Omega_1, \ldots, \Omega_N$ be disjoint subsets which approximately cover the supports of $K_1, \ldots, K_N$, in the sense that

$$\begin{cases} \epsilon_n \triangleq \int_{\mathbb{R}\backslash\Omega_n} K_n(y) dy \\ \epsilon_n' \triangleq \int_{\mathbb{R}\backslash\Omega_n} K_n(y) \log \frac{\sup K_n}{K_n(y)} dy \end{cases}$$

are small. Then, we have

$$H(p) \geq \mathcal{H}(p) - \sum_{n=1}^{N} \pi_n \epsilon_n'$$

$$- \sum_{n=1}^{N} \pi_n \left[\log \left(\frac{\max_{1 \leq m \leq N} \sup K_m}{\pi_n \sup K_n} \right) + 1 \right] \epsilon_n. \quad (9)$$

The proof of this lemma is given in the Appendix.

Let us consider now the case where the densities K_n in (5) all have the same form

$$K_n(y) = (1/\sigma_n) K[(y - \mu_n)/\sigma_n] \quad (10)$$

where K is a bounded density of finite entropy. Hence, $H(K_n) = H(K) + \log \sigma_n$ and the upper bound (7) becomes

$$H(p) \leq \mathcal{H}(p) = H(K) + \sum_{n=1}^{N} \pi_n \log \sigma_n + h(\pi). \quad (11)$$

Also, the lower bound of the entropy given by (9) reduces to

$$H(p) \geq \mathcal{H}(p) - \sum_{n=1}^{N} \pi_n \left[\epsilon_n' + \left(\log \pi_n^{-1} + 1 \right) \epsilon_n \right]. \quad (12)$$

Let us arrange the μ_n by increasing order and take σ_n small with respect to

$$d_n \triangleq \min(\mu_n - \mu_{n-1}, \mu_{n+1} - \mu_n) \quad (13)$$

where $\mu_0 = -\infty$ and $\mu_{N+1} = \infty$ by convention. Under this assumption, the density (5) is strongly multimodal and Ω_n in the above lemma can be taken to be intervals centered at μ_n of length d_n

$$\Omega_n \triangleq (\mu_n - d_n/2, \mu_n + d_n/2). \quad (14)$$

Then simple calculations give

$$\begin{cases} \epsilon_n = 1 - \int_{-d_n/(2\sigma_n)}^{d_n/(2\sigma_n)} K(x) dx \\ \epsilon_n' = H(K) - H_{d_n/\sigma_n}(K) + \epsilon_n \log(\sup K) \end{cases}$$

where

$$H_\alpha(K) \triangleq - \int_{-\alpha/2}^{\alpha/2} K(x) \log K(x) dx.$$

It is clear that ϵ_n and ϵ_n' both tend to 0 as $d_n/\sigma_n \to \infty$. Thus, one gets the following corollary.

Corollary 1: Let p be given by (5) with K_n of the form (10) and $\sup_x K(x) < \infty$. Then $H(p)$ is bounded above by $\mathcal{H}(p)$ and converges to this bound as $\min_n(d_n/\sigma_n) \to \infty$, d_n being defined in (13).

2) Explicit Calculation in the Gaussian Case: Let us focus on the $K(x) = \Phi(x)$ case where $\Phi(x)$ denotes the standard Gaussian density: $\Phi(x) = (1/\sqrt{2\pi}) e^{-x^2/2}$.

The upper and lower bounds of $H(p)$ are given by (11) and (12) with $H(\Phi)$ instead of $H(K)$; ϵ_n and ϵ_n' can now be obtained explicitly

$$\begin{cases} \epsilon_n = \text{Erfc} \left(\frac{d_n}{2\sqrt{2}\sigma_n} \right) \\ \epsilon_n' = H(\Phi) - H_{d_n/\sigma_n}(\Phi) - \epsilon_n \log \sqrt{2\pi} \end{cases}$$

where Erfc is the complementary error function defined as $\text{Erfc}(x) = (2/\sqrt{\pi}) \int_x^\infty \exp(-z^2) dz$. By double integration by parts and noting that $\int \text{Erf}(x) dx = x \text{Erf}(x) + \exp(-x^2)/\sqrt{\pi}$ with $\text{Erf}(x) = 1 - \text{Erfc}(x)$, some algebraic manipulations give

$$H_{d_n/\sigma_n}(\Phi) = \frac{1}{2} \text{Erf} \left(\frac{d_n}{2\sqrt{2}\sigma_n} \right) \log(2\pi e) - \frac{d_n}{2\sqrt{2\pi}\sigma_n} e^{-d_n^2/(8\sigma_n^2)}.$$

One can see that $H_{d_n/\sigma_n}(\Phi) \to H(\Phi) = \log \sqrt{2\pi e}$ as $d_n/\sigma_n \to \infty$, as it should be. Finally

$$\begin{cases} \epsilon_n = \text{Erfc} \left(\frac{d_n}{2\sqrt{2}\sigma_n} \right) \\ \epsilon_n' = \frac{1}{2} \text{Erfc} \left(\frac{d_n}{2\sqrt{2}\sigma_n} \right) + \frac{d_n}{2\sqrt{2\pi}\sigma_n} e^{-[d_n/(2\sqrt{2}\sigma_n)]^2}. \end{cases}$$

Example 4: To illustrate Corollary 1, Fig. 2 plots the entropy of a trimodal variable Y with density p as in (5) with K_n given by (10), $\sigma_n = \sigma$ (for the ease of illustration), $K = \Phi$, $\mu = [0, 5, 10]$, and $\pi = [1/4, 1/2, 1/4]$. Such variable can be represented as $Y = U + \sigma Z$ where U is a discrete random variable taking values in $\{0, 5, 10\}$ with probabilities $1/4, 1/2, 1/4$, and Z is a standard Gaussian variable independent from U. The upper and lower bounds of the entropy are computed as in Lemma 1 with the above expressions for ϵ_n, ϵ_n', and plotted on the same figure. One can see that the lower the σ, the better the approximation of $H(Y)$ by its upper and lower bounds. On the contrary, when σ increases, the difference between the entropy and its bounds tend to increase, which seems natural. These differences however can be seen to tend toward a constant for $\sigma \to \infty$. This can be explained as follows. When σ is large, p is no longer multimodal and tends to the Gaussian density of variance σ^2. Thus $H(Y)$ grows with σ as $\log \sigma$. On the other hand, the upper bound of $\mathcal{H}(p)$ of $H(Y)$ also grows as $\log \sigma$. The same is true for the lower bound of $H(Y)$ which equals

$$\mathcal{H}(p) - \sum_{n=1}^{3} \pi_n \left[\epsilon_n' + \epsilon_n \left(\log \pi_n^{-1} + 1 \right) \right]$$

where the last term tends to $h(\pi) + \frac{3}{2}$ as $\sigma \to \infty$ since for fixed $d_n, \epsilon_n \to 1$, and $\epsilon_n' \to 1/2$ as $\sigma \to \infty$.

C. Entropy Bounds and Decision Theory

The entropy estimator given in (7) has actually close connections with decision problems, and a tighter upper bound for $H(p)$ can be found in this framework. Assume that we have an N-class classification problem consisting in finding the class label C of an observation y_n, knowing the densities and the priors of the classes. In such kinds of classification problems, one is often interested in quantifying the Bayes probability of
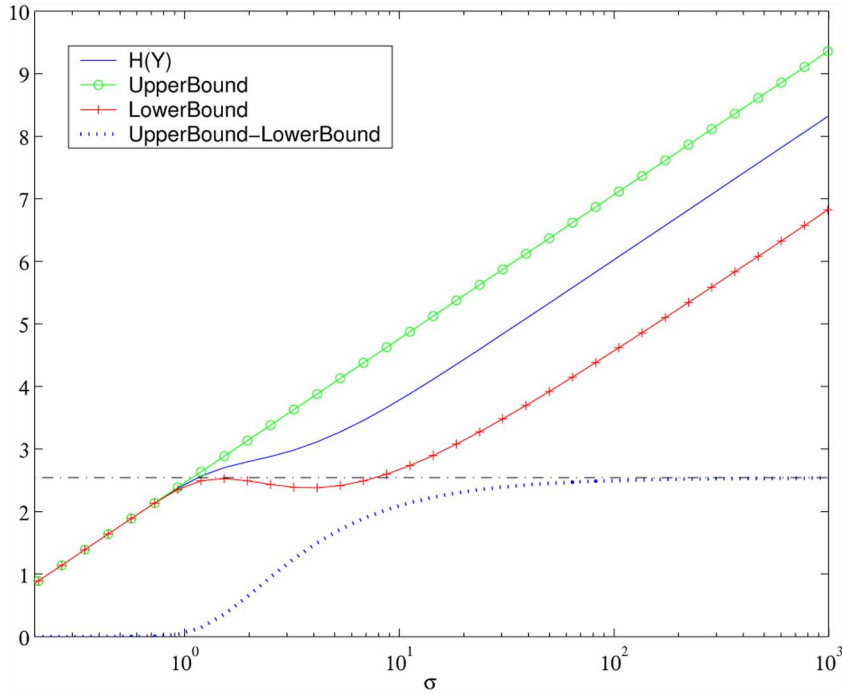
Fig. 2. Illustration of Example 4: Evolution of $H(Y)$ and its bounds versus σ, where $Y = U + \sigma Z$, U is a discrete random variable taking values in $\{0, 5, 10\}$ with probabilities $\boldsymbol{\pi} = [1/4, 1/2, 1/4]$ and Z is a standard Gaussian variable independent from U. The lower bound converges to the upper bound as $\sigma \to 0$ and the difference between upper and lower bounds tends to $3/2 + h(\boldsymbol{\pi})$ as $\sigma \to \infty$ (note that the horizontal axis scale is logarithmic).

error $P(e)$. In our context, each of the pdf mode K_n represents the density of a given class c_n, i.e., the conditional density of Y given $C = c_n$ is K_n. Furthermore, π_n is the *a priori* probability of c_n: $P(C = c_n) = \pi_n$, and p is the density of Y, which can thus be seen as a "mixture density." Defining $h(C) = -\sum_{n=1}^{N} P(C = c_n) \log P(C = c_n)$, it can be shown [23], [24] that

$$P(e) \leq \frac{1}{2} h(C|Y) = \frac{1}{2}[H(Y|C) + h(C) - H(Y)]$$

$$= \frac{1}{2}\left[\sum_{n=1}^{N} \pi_n H(K_n) + h(\pi) - H(Y)\right] \qquad (15)$$

where $H(Y|C) \triangleq \mathrm{E}_C[H(Y|C = c_i)]$, which shows that half the difference between the $\mathcal{H}(p)$ and $H(p)$ is precisely an upper bound of Bayes' probability of error $P(e) \triangleq \mathrm{E}_Y[1 - \max_i p(c_i|y)]$. The error vanishes when the modes have no overlap (the classes are separable, i.e., disjoint).

Clearly, $\mathcal{H}(p) - 2P(e)$ is a tighter upper bound of $H(p)$ than $\mathcal{H}(p)$ as $P(e) \geq 0$. On the other hand, it can be proved that $\mathcal{H}(p) - 2\sqrt{(N-1)P(e)}$ is a lower bound for $H(p)$ [24]. However, the lower bound in Lemma 1 is tighter when σ is small enough. Both bounds in this lemma are easier to deal with in more general theoretical developments, are more related to the multimodality of $p(y)$, and suffice for our purposes. Therefore, in the following theoretical developments, the last pair of bounds shall be used.

IV. MIXING LOCAL MINIMA IN MULTIMODAL BSS

Based on the results derived in Section III-B, it will be shown that mixing local minima of the entropy exist in the context

of the blind separation of multimodal sources with Gaussian modes if the mode standard deviations σ_n are *small enough*.

We are interested in the (mixing) local minima of $H(\boldsymbol{wS})$ on the unit sphere $\mathcal{S} \triangleq \{\boldsymbol{w} : \|\boldsymbol{w}\| = 1\}$ of $\mathbb{R}^K$. We shall assume that the sources have a pdf of the form (5), with K_n being Gaussian with identical variance σ^2 (but with distinct means). Thus, as in Example 4, we may represent S_k as $U_k + \sigma Z_k$ where U_k is a discrete random variable and Z_k is a standard Gaussian variable independent from U_k. Further, $(U_1, Z_1), \ldots, (U_K, Z_K)$ are assumed to be independent so that the sources are independent as required. From this representation, $\boldsymbol{wS} = \boldsymbol{wU} + \sigma Z$ where $\boldsymbol{U}$ is the column vector with components U_k and Z is again a standard Gaussian variable (since any linear combination of independent Gaussian variables is a Gaussian variable and $\sum_{k=1}^{K} w_k Z_k$ has zero mean and unit variance). Since $\boldsymbol{wU}$ is clearly a discrete random variable, $\boldsymbol{wS}$ also has a multimodal distribution of the form (5) with K_n again the Gaussian density with variance σ^2. Note that the number of modes is the number of distinct values $\boldsymbol{wU}$ can have and the mode centers (the means of the K_n) are these values; they depend on $\boldsymbol{w}$. However, as long as σ is small enough with respect to the distances d_n defined in (13) the approximation (7) of the entropy is justified. Thus, we are led to the approximation $H(\boldsymbol{wS}) \approx h(\boldsymbol{wU}) + \log \sigma + H(\Phi)$, where $h(\boldsymbol{wU})$ denotes abusively the entropy of the discrete random variable $\boldsymbol{wU}$ (the entropy of a discrete random variable U with probability vector $\boldsymbol{\pi}$ is noted either $h(U)$ or $h(\boldsymbol{\pi})$).

The above approximation suggests that there is a relationship between the local minimum points of $H(\boldsymbol{wS})$ and those of $h(\boldsymbol{wU})$. Therefore, we shall first focus on the local minimum points of the entropy of $\boldsymbol{wU}$ before analyzing those of $H(\boldsymbol{wS})$.
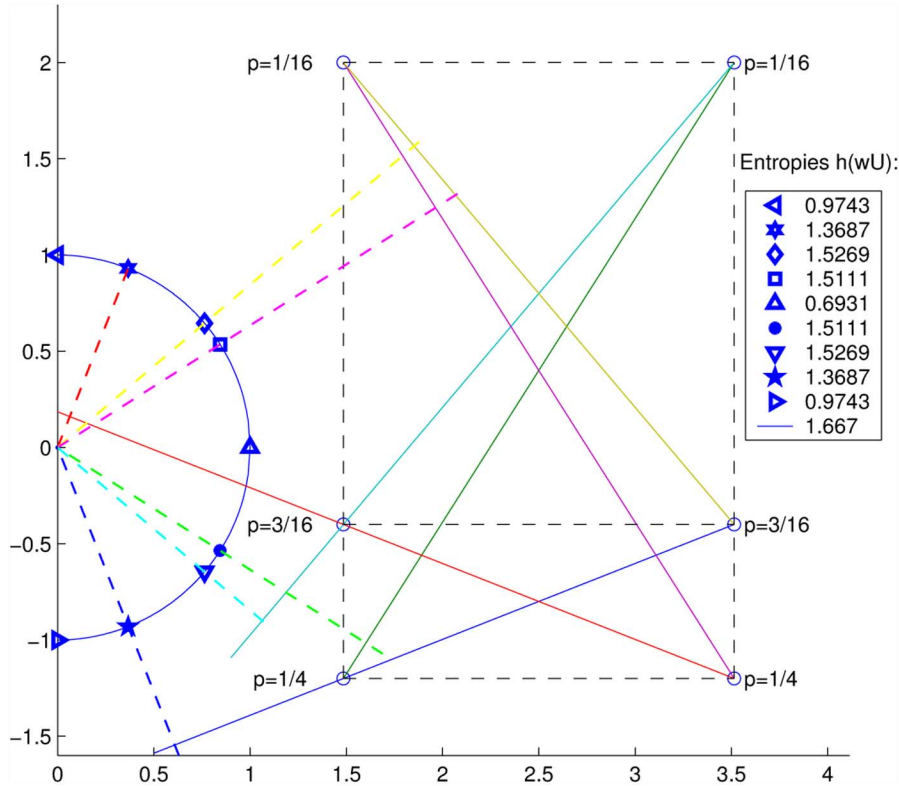
Fig. 3.   Example 5: illustration of Lemma 2. The discrete random variables U_1 and U_2 take values in $\{-\sqrt{1.03}+2.5, \sqrt{1.03}+2.5\}$, and $\{-1.2, -0.4, 2\}$ with probabilities $[0.5, 0.5]$ and $[1/2, 3/8, 1/8]$, respectively. The entropies at the points located by the corresponding markers shown on the half-circle are given in the legend.

A. Local Minimum Points of $h(\boldsymbol{w}\boldsymbol{U})$

The function $h(\boldsymbol{w}\boldsymbol{U})$ does not depend on the values that $\boldsymbol{w}\boldsymbol{U}$ can take but only on the associated probabilities; these probabilities remain constant as $\boldsymbol{w}$ changes unless the number of distinct values that $\boldsymbol{w}\boldsymbol{U}$ can take varies. Such number would decrease when an equality $\boldsymbol{w}\boldsymbol{u} = \boldsymbol{w}\boldsymbol{u}'$ is attained for some distinct column vectors $\boldsymbol{u}$ and $\boldsymbol{u}'$ in the set of possible values of $\boldsymbol{U}$. A deeper analysis yields the following result, which is helpful to find the local minimum point of $h(\boldsymbol{w}\boldsymbol{U})$.

Lemma 2: Let $\boldsymbol{U}$ be a discrete random vector in $\mathbb{R}^K$ and $\mathcal{U}$ be the set of distinct values it can take. Assume that there exists $r \geq 1$ disjoint subsets $\mathcal{U}_1, \ldots, \mathcal{U}_r$ of $\mathcal{U}$ each containing at least two elements, such that the linear subspace V spanned by the vectors $\boldsymbol{u} - \boldsymbol{u}_1, \boldsymbol{u} \in \mathcal{U}_1 \setminus \{\boldsymbol{u}_1\}, \ldots, \boldsymbol{u} - \boldsymbol{u}_r, \boldsymbol{u} \in \mathcal{U}_r \setminus \{\boldsymbol{u}_r\}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ being arbitrary elements of $\mathcal{U}_1, \ldots, \mathcal{U}_r$, is of dimension $K - 1$. (Note that V does not depend on the choice of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$, since $\boldsymbol{u} - \boldsymbol{u}_j' = (\boldsymbol{u} - \boldsymbol{u}_j) - (\boldsymbol{u}_j' - \boldsymbol{u}_j)$ for any other $\boldsymbol{u}_j' \in \mathcal{U}_j$.) Then for $\boldsymbol{w}^* \in \mathcal{S}$ and orthogonal to V, there exists a neighborhood $\mathcal{W}$ of $\boldsymbol{w}^*$ in $\mathcal{S}$ and $\alpha > 0$ such that $h(\boldsymbol{w}\boldsymbol{U}) \geq h(\boldsymbol{w}^*\boldsymbol{U}) + \alpha$ for all $\boldsymbol{w} \in \mathcal{W} \setminus \{\boldsymbol{w}^*\}$. In the case $K = 2$, one has a stronger result that $h(\boldsymbol{w}\boldsymbol{U}) = h(\boldsymbol{U}) > h(\boldsymbol{w}^*\boldsymbol{U})$ for all $\boldsymbol{w} \in \mathcal{W} \setminus \{\boldsymbol{w}^*\}$.

The proof is given in Appendix II.

Example 5: An illustration of Lemma 2 in the $K = 2$ case (again for clarity) is provided in Fig. 3. We note $\boldsymbol{U} = [U_1, U_2]^T$ where the discrete variables U_1 and U_2 take the values $-\sqrt{1.03} + 2.5, \sqrt{1.03} + 2.5$ with probabilities and $0.5, 0.5$

and the values $-1.2, -4, 2$ with probabilities $1/2, 3/8, 1/8$, respectively. They are chosen to have the same variance, as we need that the $S_k = U_k + \sigma Z_k, k = 1, 2$, have the same variance. But their mean can be arbitrary since $H(\boldsymbol{w}\boldsymbol{S})$ does not depend on them. In this $K = 2$ example, each line that links two distinct points $\boldsymbol{u}, \boldsymbol{u}' \in \mathcal{U}$ span a one-dimensional linear subspace, which constitutes a possible subspace V, as stated in Lemma 2. There are thus many possibilities for V, each corresponding to a specific vector $\boldsymbol{w}^*$.

Two simple possibilities for V are the subspaces with direction given by $[0, 1]^T$ and $[1, 0]^T$. In the first case, the subsets $\mathcal{U}_i$ are built by grouping the points of $\mathcal{U}$ laying on a same vertical dashed line. There are two such subsets $(r = 2)$ consisting of $\boldsymbol{u} \in \mathcal{U}$ with first component equal to $-\sqrt{1.03} + 2.5$ and $\sqrt{1.03} + 2.5$, respectively. In the second case, the subsets $\mathcal{U}_i$ are built by grouping the points of $\mathcal{U}$ laying on a same horizontal dashed line. There are three such subsets $(r = 3)$ consisting of $\boldsymbol{u} \in \mathcal{U}$ with second component equal to $-1.2, -0.4$, and 2, respectively.

There also exist other subspaces V, corresponding to "diagonal lines" (i.e., to solid lines in Fig. 3). This last kind of one-dimensional linear subspace V corresponds to directions given by two-dimensional vectors $\boldsymbol{w}^*$ with two nonzero elements.

On the plot, the points on the half circle correspond to the vectors $\boldsymbol{w}^*$ of the lemma; each $\boldsymbol{w}^*$ is orthogonal to a line joining a pair of distinct points in $\mathcal{U}$, $\mathcal{U}$ being the set of all possible values of $[U_1, U_2]^T$. The points of $\mathcal{U}$ are displayed in the plot together with their probabilities. The entropies $h(\boldsymbol{w}\boldsymbol{U})$ are also
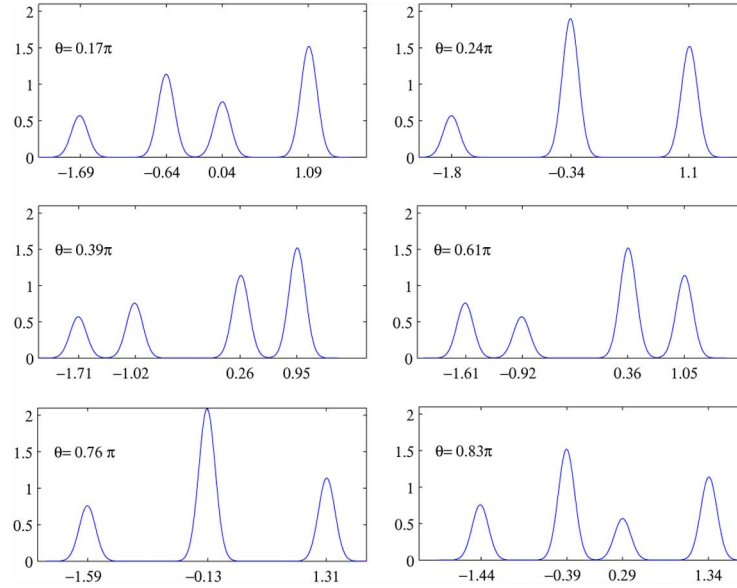
Fig. 4. Example 6: pdf of $\boldsymbol{wS}$ for various angles θ.

given in the plot; one can see that they are lower for $\boldsymbol{w} = \boldsymbol{w}^*$ than for other points $\boldsymbol{w}$.

The preceding lemma only provides a mean to find a local minimum point of the function $h(\boldsymbol{wU})$, but does not prove the existence of such a point, since the existence of V was only *assumed* in the lemma. Nevertheless, in the case where the components of $\boldsymbol{U}$ are independent and can take at least two distinct values, subset $\mathcal{U}_i$ ensuring the existence of V can be built as follows. Let j be any index in $\{1, \ldots, K\}$ and $\lambda_{j,1}, \ldots, \lambda_{j,r_j}$ be the possible value of U_j, the jth component of $\boldsymbol{U}$. One can take $\mathcal{U}_i, 1 \le i \le r_j$ to be the set of $\boldsymbol{u} \in \mathcal{U}$ such that its jth components equal to $\lambda_{j,i}$. Then it is clear that the corresponding subspace V consists of all vectors orthogonal to the jth row of the identity matrix (hence V is of dimension $K - 1$) and that the associated vector $\boldsymbol{w}^*$ is simply this row or it opposite. By Lemma 2, this point $\boldsymbol{w}^*$ would be a local minimum point of $h(\boldsymbol{wU})$. But, as explained above, it is a non-mixing point while we are interested in the mixing point, i.e., not proportional to a row of the identity matrix. However, the above construction can be extended by looking for a set of K vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K$ in $\mathcal{U}$, such that the vectors $\boldsymbol{u}_i - \boldsymbol{u}_j, 1 \le i < j \le K$ span any linear subspace of dimension $K - 1$ of $\mathbb{R}^K$. If such a set can be found, then V is simply this linear subspace by taking $\mathcal{U}_1 = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K\}$ and $r = 1$. In addition, if $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K$ do not all have the same jth component, for some j, then the corresponding $\boldsymbol{w}^*$ is a mixing local minimum point. In view of the fact that there are at least 2^K points in $\mathcal{U}$ to choose from for the $\boldsymbol{u}_i$ and that the last construction procedure meant not to find all local minimum points of $h(\boldsymbol{wU})$, chances are that there exists both non-mixing and mixing local minimum points of $h(\boldsymbol{wU})$. In the $K = 2$ case this is really the case: it suffices to take two distinct points $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ in $\mathcal{U}$, then by the precediing lemma, the vector $\boldsymbol{w}^*$ orthogonal to $\boldsymbol{u}_1 - \boldsymbol{u}_2$ is a local minimum point of $h(\boldsymbol{wU})$. If one chooses $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ such that both components of $\boldsymbol{u}_1 - \boldsymbol{u}_2$ are nonzero, the associated orthogonal vector $\boldsymbol{w}^*$ is not proportional to any row of the identity matrix; it is a mixing local

minimum point of $h(\boldsymbol{wU})$. Note that in the particular $K = 2$ case, the aforementioned method identifies all local minimum points of $h(\boldsymbol{wU})$. Indeed, for any $\boldsymbol{w} \in \mathcal{S}$, either there exists a pair of distinct vectors $\boldsymbol{u}_1, \boldsymbol{u}_2$ in $\mathcal{U}$ such that $\boldsymbol{w}(\boldsymbol{u}_1 - \boldsymbol{u}_2) = 0$ or there exists no such pair. In the first case, $\boldsymbol{w}$ is a local minimum point and in the second case, one has $h(\boldsymbol{wU}) = h(\boldsymbol{U})$. Since there is only a finite number of the differences $\boldsymbol{u}_1 - \boldsymbol{u}_2$, for distinct $\boldsymbol{u}_1, \boldsymbol{u}_2$ in $\mathcal{U}$, there can be only a finite number of local minimum points of $h(\boldsymbol{wU})$, and for all other points $h(\boldsymbol{wU})$ take the maximum value $h(\boldsymbol{U})$.

B. Local Minimum Points of $H(\boldsymbol{wS})$

This subsection shows that the local minima points of $H(\boldsymbol{wS})$ can be related to those of $h(\boldsymbol{wU})$.

Lemma 3: Define $S_i, i = 1, \ldots, K$, as $S_i = U_i + \sigma Z_i$ described at the beginning of Section IV-B2 and $\boldsymbol{w}^*$ be a vector satisfying the assumption of Lemma 2 ($\boldsymbol{U}$ being the vector with component U_i). Then for σ sufficiently small, $H(\boldsymbol{wS})$ admits a local minimum point converging to $\boldsymbol{w}^*$ as $\sigma \to 0$.

The proof of this lemma is relegated to the Appendix.

Example 6: Thanks to the entropy approximator, we shall illustrate the existence of the local minima of $H(\boldsymbol{wS})$ in the following $K = 2$ example, so that vectors $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}\| = 1$ can be written as $[\sin \theta, \cos \theta]$. We take $S_1 = U_{\pi/2} + \sigma Z_1$ and $S_2 = U_0 + \sigma Z_2$, where $U_0, U_{\pi/2}$ are independent discrete random variables taking the values $-2\sqrt{3}/3, \sqrt{3}/2$ with probabilities $1/3, 2/3$, and $-\sqrt{2}, \sqrt{2}/2$ with probabilities $3/7, 4/7$, respectively, and Z_1, Z_2 are standard Gaussian variables. The parameter σ is set to 0.1. Thus, $Y_\theta = \boldsymbol{wS}$ can be represented as $U_\theta + \sigma Z$ where $U_\theta = \sin \theta U_{\pi/2} + \cos \theta U_0$ and Z is a standard Gaussian variable independent from U_θ. Fig. 4 plots the pdf of Y_θ for various angles θ. It can be seen that the modality (i.e., the number of modes) changes with θ. Fig. 5 shows the entropy of Y_θ together with its upper and lower bounds, for
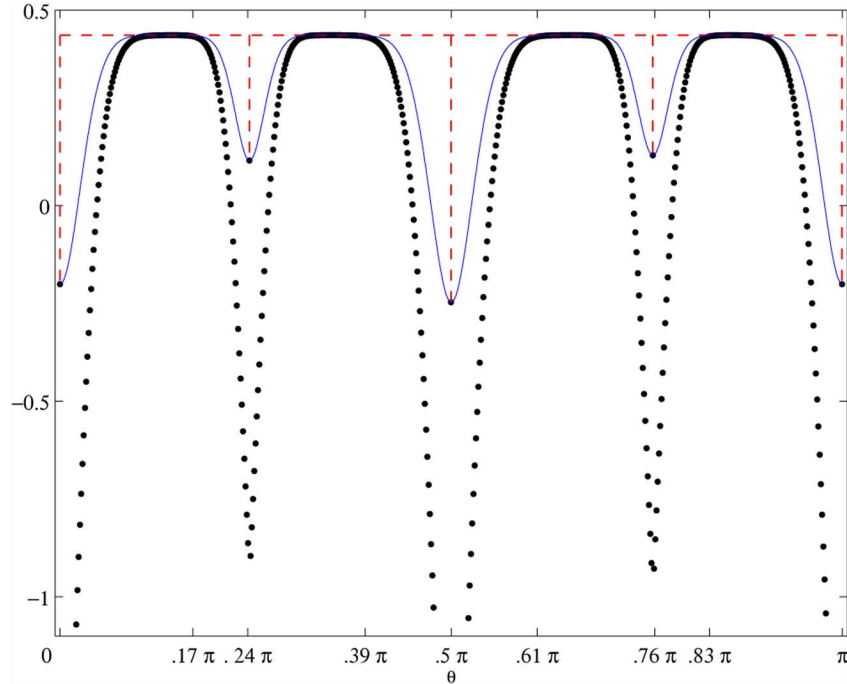
Fig. 5.   Example 6: Upper bound (dashed line), lower bound (dots), and entropy estimation of Y_θ using finite Riemannian sum (solid). It can be seen that the upper and lower bounds of the entropy converge to each other when the density becomes strongly multimodal (see the corresponding plots in Fig. 4).

$\theta \in [0, \pi]$. In addition to non-mixing local minima at $\theta \in \{p\pi/2 | p \in \mathbb{Z}\}$, mixing local minima exist when $\boldsymbol{w}(\boldsymbol{u}_1 - \boldsymbol{u}_2) = 0$, where $\boldsymbol{u}_1 = [-2\sqrt{3}/3, \sqrt{2}/2]^T$, $\boldsymbol{u}_2 = [\sqrt{3}/2, -\sqrt{2}]^T$, i.e., when $|\tan(\theta)| = 0.9526$, or $\theta \in \{(0.2423 + p)\pi, (0.7577 + p)\pi | p \in \mathbb{Z}\}$. One can observe that the upper bound is a constant function except for a finite number of angles for which we observe negative peaks (see Lemma 2). For these angles, the pdf is strongly multimodal, and the upper and lower bounds are very close, though not clearly visible on the figure. This results from a discontinuity of the lower bound at these angles, due to the superimposition of several modes at these angles.

V. COMPLEMENTARY OBSERVATIONS

This section provides two observations that can be drawn regarding the impact of the *mode variance* σ^2 on the existence of local minima and the symmetry of the entropy with respect to θ.

A. Impact of "Mode Variance" σ^2

In the example of Fig. 6, the discrete variables U_1 and U_2 in the expression of S_1 and S_2 are taken as in Example 5. One can observe that the mixing minima of the entropy tends to disappear when the mode variance increases. This is a direct consequence of the fact that the mode overlaps increase. When σ increases, the source densities become more and more Gaussian and the $H(\boldsymbol{w}\boldsymbol{S})$ versus θ curve tends to be more and more flat, approaching the constant function $\log \sqrt{2\pi e} + \log \sigma$. The upper and lower bounds have only been plotted for the $\sigma = 0.05$, for visibility purposes. Again, at angles corresponding to the upper bound negative peaks, the error bound is very tight, as explained in Example 6.

B. Note on Symmetry of $H(\boldsymbol{w}_\theta \boldsymbol{S})$

In the above graphs plotting the entropy (and its bounds) versus θ, some symmetry can be observed. First, if we note $\boldsymbol{w}_\theta \triangleq [\sin \theta \cos \theta]$, observe that $H(\boldsymbol{w}_\theta \boldsymbol{S}) = H(\boldsymbol{w}_{\theta+\pi} \boldsymbol{S})$ whatever are the source pdfs; this is a direct consequence of the fact the entropy is not sign sensitive. Second, if one of the source densities is symmetric, i.e., if there exists $\mu \in \mathbb{R}$ such that $p_{S_j}(\mu - s) = p_{S_j}(\mu + s)$ for all $s \in \mathbb{R}$, then $H(\boldsymbol{w}_\theta \boldsymbol{S}) = H(\boldsymbol{w}_{-\theta} \boldsymbol{S})$. Third, if the two sources share the same pdf, then $H(\boldsymbol{w}_\theta \boldsymbol{S}) = H(\boldsymbol{w}_{\pi/2-\theta} \boldsymbol{S})$. Finally, if the two sources can be expressed as in Lemma 3, then the vectors $\boldsymbol{w}^*$ for which $h(\boldsymbol{w}^* \boldsymbol{U}) < h(\boldsymbol{U})$ (as obtained in Lemma 2) are symmetric in the sense that their angles are pairwise opposite. This means that for σ small enough, if a local minimum of $H(\boldsymbol{w}_\theta \boldsymbol{S})$ appears at θ^*, then another local minimum point will exist near $-\theta^*$ (and thus near $p\pi - \theta, \forall p \in \mathbb{Z}$). The above symmetry property can be seen from Fig. 3 and can be proved formally as follows. From Lemma 2, $\boldsymbol{w}^*$ must be orthogonal to $\boldsymbol{u}_1 - \boldsymbol{u}_2$ for some pair of distinct vectors in the set of all possible values of $\boldsymbol{U}$. Define $\boldsymbol{u}_i^\dagger$ ($i = 1, 2$) to be the vector with first coordinate the same as that of $\boldsymbol{u}_{3-i}$ and second coordinate the same as that of $\boldsymbol{u}_i$. Then it can be seen that the vector orthogonal to $\boldsymbol{u}_1^\dagger - \boldsymbol{u}_2^\dagger$ has an angle opposite to the angle of $\boldsymbol{w}^*$, yielding the desired result.

VI. CONCLUSION

In this paper, new results regarding both non-mixing and mixing entropy local minima have been derived in the context of the blind separation of K sources. First, it is shown that a local entropy minimum exists when the output is proportional to one of the non-Gaussian source. Second, it is shown that
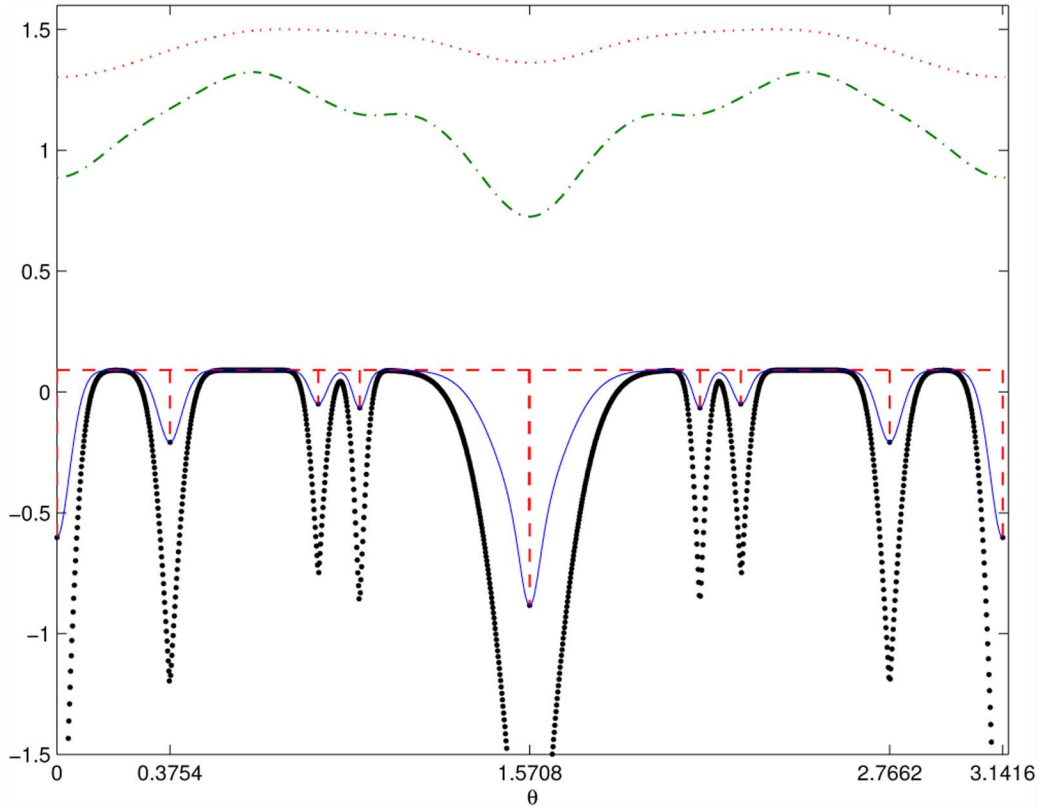
Fig. 6. Entropy of $\boldsymbol{wS}$ (estimated using finite Riemannian sum) versus θ for $S_1 = U_1 + \sigma Z_1$, $S_2 = U_2 + \sigma Z_2$, where U_1 and U_2 are taken from Example 5 (and Fig. 3) and the four random variables are all independent. The parameter σ is set to 0.5 (solid), 0.25 (dashed-dotted), and 0.5 (dotted). The upper and lower bounds have been added for the $\sigma = 0.05$ case only, for visibility purposes. It can be seen that the upper and lower bounds of the entropy converge to each other when the density becomes strongly multimodal.

mixing entropy minima may exist when the source densities are strongly multimodal (i.e., multimodal with sufficiently small overlap); therefore, spurious BSS solutions can be obtained when minimizing this entropic criterion. Some attention must be paid to the obtained solutions when they are found by adaptive gradient minimization.

To prove the existence of mixing entropy minima, a theoretical framework using an entropy approximator and its associated error bounds has been provided. Even if this approximator is considered here in the context of BSS, its use can be extended to other applications involving entropy estimation.

APPENDIX
PROOFS OF LEMMAS

Proof of Lemma 1: We have from (6) that $H(Y) = \sum_{n=1}^{N} \pi_n H_n$ where

$$H_n \triangleq - \int K_n(y) \log \left[\sum_{m=1}^{N} \pi_m K_m(y) \right] dy. \quad (16)$$

Since all $K_m \geq 0$, the last right-hand side is bounded above by $- \int K_n(y) \log[\pi_n K_n(y)] dy = H(K_n) - \log \pi_n$, yielding the inequality (8).

A more elegant derivation of this inequality can be obtained from the entropy properties. Indeed, the density given in (5) can be interpreted as the marginal density of an augmented model (Y, U) where U is a discrete variable with N values $u_1, \ldots, u_n$ with probabilities $\pi_1, \ldots, \pi_n$ and Y has a conditional density given $U = u_n$ equal to K_n. The joint entropy $H(Y, U)$ of (the

"continuous-discrete" pair of random variables) Y, U equals $H(Y|U) + h(U)$ where $h(U) = h(\boldsymbol{\pi})$ is the discrete entropy of U and $H(Y|U) = \sum_{n=1}^{N} \pi_n H(K_n)$ is the conditional entropy of Y given U. But $H(Y, U) = h(U|Y) + H(Y)$ (where $h(U|Y)$ is the conditional entropy of U given Y) and thus $\mathcal{H}(p) - H(p)$ equals $h(U|Y)$ which is always nonnegative because U is a discrete variable.

Yet another way to prove the above inequality is to exploit its connection to the decision problem discussed in Section III-C. Indeed, (15) yields immediately $\mathcal{H}(p) - H(p) \geq P(e) \geq 0$.

To prove the second result, noting that $\log(1 + x) \leq x$, the term $\log \left[\sum_{m=1}^{N} \pi_m K_m(y) \right]$ can be bounded above by

$$\begin{cases} \log[\pi_n K_n(y)] + \sum_{1 \leq m \leq N, m \neq n} \frac{\pi_m K_m(y)}{\pi_n K_n(y)}, & \text{if } y \in \Omega_n \\ \log(\max_{1 \leq m \leq N} \sup K_m), & \text{otherwise.} \end{cases} \quad (17)$$

Therefore, with

$$H_n \triangleq - \int K_n(y) \log \left[\sum_{m=1}^{N} \pi_m K_m(y) \right] dy \quad (18)$$

one gets

$$\begin{aligned} H_n \geq &- \int_{\Omega_n} K_n(y) \log[\pi_n K_n(y)] dy \\ &- \sum_{1 \leq m \leq N, m \neq n} \frac{\pi_m}{\pi_n} \int_{\Omega_n} K_m(y) dy \\ &- \log \left(\max_{1 \leq m \leq N} \sup K_m \right) \epsilon_n. \end{aligned}$$

But since $\Omega_1, \ldots, \Omega_N$ are disjoint

$$\sum_{n=1}^{N} \pi_n \sum_{1 \le m \le N, m \ne n} \frac{\pi_m}{\pi_n} \int_{\Omega_n} K_m(y) dy$$

$$= \sum_{m=1}^{N} \pi_m \int_{\cup_{1 \le n \ne m \le N} \Omega_n} K_m(y) dy$$

and $\cup_{1 \le n \ne m \le N} \Omega_n \subseteq \mathbb{R} \setminus \Omega_m$. Therefore, the right-hand side of the above equality is bounded above by $\sum_{m=1}^{N} \pi_m \epsilon_m$. It follows that $H(p) = \sum_{n=1}^{N} \pi_n H_n$ is bounded below by

$$h(\boldsymbol{\pi}) + \sum_{n=1}^{N} \pi_n H(K_n) + \sum_{n=1}^{N} \pi_n \log(\pi_n \sup K_n) \epsilon_n - \sum_{n=1}^{N} \pi_n \epsilon'_n$$

$$- \sum_{m=1}^{N} \pi_m \epsilon_m - \sum_{n=1}^{N} \pi_n \log \left(\max_{1 \le m \le N} \sup K_m \right) \epsilon_n.$$

After some manipulations, the above expression reduces to the lower bound for $\sum_{n=1}^{N} \pi_n H_n$ given in the lemma $\qquad \square$

Proof of Lemma 2: By construction, for each $j = 1, \ldots, r$, $\boldsymbol{w}^* \boldsymbol{u}$ take the same values for $\boldsymbol{u} \in \mathcal{U}_j$. On the other hand, by grouping the vectors $\boldsymbol{u} \in \mathcal{U}$ which produce the same value of $\boldsymbol{w}^* \boldsymbol{u}$ into subsets of $\mathcal{U}$, one gets a partition of $\mathcal{U}$ into $r^* + 1$ subsets $\mathcal{U}_0^*, \ldots, \mathcal{U}_{r^*}^*$, such that each $\mathcal{U}_j^*, 1 \le j \le r^*$ contains at least two elements and $\boldsymbol{w}^* \boldsymbol{u}$ takes the same values for $\boldsymbol{u} \in \mathcal{U}_j^*$ and the values associated with different $\mathcal{U}_j^*$ and the $\boldsymbol{w}^* \boldsymbol{u}, \boldsymbol{u} \in \mathcal{U}_0^*$, are all distinct. Obviously $r^* \ge 1$ and each of the $\mathcal{U}_1, \ldots, \mathcal{U}_r$ must be contained in one of the $\mathcal{U}_1^*, \ldots, \mathcal{U}_{r^*}^*$. Therefore, the space V must be contained in the space spanned by the vectors $\boldsymbol{u} - \boldsymbol{u}_j, \boldsymbol{u} \in \mathcal{U}_j^* \setminus \{\boldsymbol{u}_j\}, j = 1, \ldots, r^*, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r^*}$ being arbitrary elements of $\mathcal{U}_1^*, \ldots, \mathcal{U}_{r^*}^*$. But the last space is orthogonal to $\boldsymbol{w}^*$ by construction and thus cannot have dimension greater than $K - 1$, hence it must coincide with V.

Putting $P(\boldsymbol{u})$ for $P(\boldsymbol{U} = \boldsymbol{u})$ for short and $P(\mathcal{U}_j^*) = \sum_{\boldsymbol{u} \in \mathcal{U}_j^*} P(\boldsymbol{u})$, one has

$$h(\boldsymbol{w}^* \boldsymbol{U}) = - \sum_{\boldsymbol{u} \in \mathcal{U}_0^*} P(\boldsymbol{u}) \log P(\boldsymbol{u}) - \sum_{j=1}^{r^*} P(\mathcal{U}_j^*) \log P(\mathcal{U}_j^*).$$

For a given pair $\boldsymbol{u}, \boldsymbol{u}'$ of distinct vectors in $\mathcal{U}$, if $\boldsymbol{w}^*(\boldsymbol{u} - \boldsymbol{u}') \ne 0$ then it remains so when $\boldsymbol{w}^*$ is changed to $\boldsymbol{w}$ provided that the change is sufficiently small. But if $\boldsymbol{w}^*(\boldsymbol{u} - \boldsymbol{u}') = 0$ then this equality may break no matter how small the change. In fact, if $\boldsymbol{w}$ is not proportional to $\boldsymbol{w}^*$, it is not orthogonal to V, hence, $\boldsymbol{w}(\boldsymbol{u} - \boldsymbol{u}') \ne 0$ for at least one pair $\boldsymbol{u}, \boldsymbol{u}'$ of distinct points in some $\mathcal{U}_j^*$, meaning that $\boldsymbol{w}\boldsymbol{u}$ takes at least two distinct values in $\mathcal{U}_j^*$. Thus, there exists a neighborhood of $\mathcal{W}$ of $\boldsymbol{w}^*$ in $\mathcal{S}$ such that for all $\boldsymbol{w} \in \mathcal{W} \setminus \{\boldsymbol{w}^*\}$, each subset $\mathcal{U}_j^*$ be partitioned into subsets $\mathcal{U}_{j,k}(\boldsymbol{w}), k = 1, \ldots, n_j(\boldsymbol{w})$ ($n_j(\boldsymbol{w})$ can be 1) such that $\boldsymbol{w}\boldsymbol{u}$ takes the same value on $\mathcal{U}_{j,k}(\boldsymbol{w})$, and the values of $\boldsymbol{w}\boldsymbol{u}$ on the subsets $\mathcal{U}_{j,k}(\boldsymbol{w})$ and on each point of $\mathcal{U}_0^*$ are distinct. Further, there exists at least one index i for which $n_i(\boldsymbol{w}) > 1$. For such an index

$$P(\mathcal{U}_i^*) \log P(\mathcal{U}_i^*) = \sum_{k=1}^{n_i(\boldsymbol{w})} P[\mathcal{U}_{i,k}(\boldsymbol{w})] \log P[\mathcal{U}_{i,k}(\boldsymbol{w})]$$

$$+ \sum_{k=1}^{n_i(\boldsymbol{w})} P[\mathcal{U}_{i,k}(\boldsymbol{w})] \frac{\log P(\mathcal{U}_i^*)}{P[\mathcal{U}_{i,k}\boldsymbol{w}]}.$$

The last term can be seen to be a strictly positive number, as $P(\mathcal{U}_i^*) > P[\mathcal{U}_{i,k}(\boldsymbol{w})]$ for $1 \le k \le n_i(\boldsymbol{w})$. Note that this term does not depend directly on $\boldsymbol{w}$ but only indirectly via the set $\mathcal{U}_{j,k}(\boldsymbol{w}), k = 1, \ldots, n_j(\boldsymbol{w}), j = 1, \ldots, r^*$, and there is only a finite number of possible such sets. Therefore, $h(\boldsymbol{w}\boldsymbol{U}) \ge h(\boldsymbol{w}^*\boldsymbol{U}) + \alpha$ for some $\alpha > 0$ for all $\boldsymbol{w} \in \mathcal{W}$.

In the case $K = 2$, the space V reduces to a line and thus the differences $\boldsymbol{u} - \boldsymbol{u}'$ for distinct $\boldsymbol{u}, \boldsymbol{u}'$ in $\mathcal{U}_j^*$, for all j, are proportional to this line. Thus, if $\boldsymbol{w}$ is not proportional to $\boldsymbol{w}^*$, and hence not orthogonal to this line, $\boldsymbol{w}\boldsymbol{u}$ take distinct values on each of the sets $\mathcal{U}_1^* \ldots, \mathcal{U}_{r^*}^*$, and if $\boldsymbol{w}$ is close enough to $\boldsymbol{w}^*$, these values are also distinct for different sets and distinct from the values of $\boldsymbol{w}\boldsymbol{u}$ on $\mathcal{U}_0^*$, which are distinct themselves. Thus, for such $\boldsymbol{w}, h(\boldsymbol{w}\boldsymbol{U}) = h(\boldsymbol{U})$. $\qquad \square$

Proof of Lemma 3: The proof of this lemma is quite involved in the $K > 2$ case, therefore, we will first give the proof for the $K = 2$ case which is much simpler, and then proceed by extending it to $K > 2$. As already shown in the beginning of Section IV, $\boldsymbol{w}\boldsymbol{S} = \boldsymbol{w}\boldsymbol{U} + \sigma Z$ where Z is a standard Gaussian distribution. Thus, the density of $\boldsymbol{w}\boldsymbol{S}$ is of the form (5) with $K_n(y) = \Phi[(y - \mu_n)/\sigma]/\sigma, \mu_1, \ldots, \mu_N$ being the possible values of $h(\boldsymbol{w}\boldsymbol{U})$, and Φ being the standard Gaussian density. For $\boldsymbol{w} = \boldsymbol{w}^*$, one has by Lemma 1

$$H(\boldsymbol{w}^* \boldsymbol{S}) \le h(\boldsymbol{w}^* \boldsymbol{U}) + H(\Phi) + \log \sigma.$$

On the other hand, we have seen in the proof of Lemma 2 that for $\boldsymbol{w}$ in some neighborhood $\mathcal{W}$ of $\boldsymbol{w}^*$ and distinct from $\boldsymbol{w}$, the $\boldsymbol{w}\boldsymbol{u}, \boldsymbol{u} \in \mathcal{U}$ ($\mathcal{U}$ denoting the set of possible values of $\boldsymbol{U}$) are all distinct (in the $K = 2$ case). Thus, the maps $\boldsymbol{u} \mapsto \boldsymbol{w}\boldsymbol{u}$ map different points $\boldsymbol{u} \in \mathcal{U}$ to different μ_n. However, when $\boldsymbol{w}$ approaches $\boldsymbol{w}^*$, some of the μ_n tend to coincide and thus some of the d_n defined in (13) approach zero. To avoid this, we restrict $\boldsymbol{w}$ to $\mathcal{W} \setminus \mathcal{W}'$ where $\mathcal{W}'$ is any open neighborhood of $\boldsymbol{w}^*$ strictly included in $\mathcal{W}$. Then $\min_n d_n \ge d$ for all $\boldsymbol{w} \in \mathcal{W} \setminus \mathcal{W}'$ for some $d > 0$ (which depends on $\mathcal{W}'$). Thus by Corollary 1, $H(\boldsymbol{w}\boldsymbol{S})$ can be made arbitrarily close to $h(\boldsymbol{w}\boldsymbol{U}) + H(\Phi) + \log \sigma$ for all $\boldsymbol{w} \in \mathcal{W} \setminus \mathcal{W}'$ by taking σ small enough. But $h(\boldsymbol{w}\boldsymbol{U}) = h(\boldsymbol{U}) > h(\boldsymbol{w}^*\boldsymbol{U})$, therefore, $H(\boldsymbol{w}\boldsymbol{S}) > H(\boldsymbol{w}^*\boldsymbol{S})$ for all $\boldsymbol{w} \in \mathcal{W} \setminus \mathcal{W}'$, for σ small enough.

One can always choose $\mathcal{W}$ to be a close set in $\mathcal{S}$; hence it is compact. Since the function $\boldsymbol{w} \in \mathcal{W} \mapsto H(\boldsymbol{w}\boldsymbol{S})$ is continuous, it must admit a minimum, which by the above result must be in $\mathcal{W}'$ and thus is not on the boundary of $\mathcal{W}$. This shows that this minimum is a local minimum. Finally, as one can choose $\mathcal{W}'$ arbitrarily small, the above result shows that the above local minimum converges to $\boldsymbol{w}^*$ as $\sigma \to 0$.

Consider now the case $K > 2$. The difficulty is that it is no longer true that for $\boldsymbol{w}$ in some neighborhood $\mathcal{W}$ of $\boldsymbol{w}^*$ and distinct from $\boldsymbol{w}^*$, the $\boldsymbol{w}\boldsymbol{u}, \boldsymbol{u} \in \mathcal{U}$ are all distinct. Indeed, by construction of $\boldsymbol{w}^*$, there exists $K - 1$ pairs $(\boldsymbol{u}_j, \boldsymbol{u}'_j), 1 \le j < K$, of distinct vectors in $\mathcal{U}$ such that the differences $\boldsymbol{u}_j - \boldsymbol{u}'_j$ are linearly independent and $\boldsymbol{w}^*(\boldsymbol{u}_j - \boldsymbol{u}'_j) = 0, 1 \le j < K$. For $\boldsymbol{w}$ not proportional to $\boldsymbol{w}^*$, at least one (but not necessary all) of the above equalities will break. Therefore, all the $\boldsymbol{w}\boldsymbol{u}, \boldsymbol{u} \in \mathcal{U}$ may be not distinct, even if $\boldsymbol{w}$ is restricted to $\mathcal{W} \setminus \mathcal{W}'$. But the set of $\boldsymbol{w}$ for which this property is not true anymore is the union of a finite number of linear subspaces of dimension $K - 1$ of $\mathbb{R}^K$ and

thus is not dense in $\mathbb{R}^K$. Therefore, for most of the $\boldsymbol{w} \in \mathcal{W} \setminus \mathcal{W}'$, the $\boldsymbol{wu}, \boldsymbol{u} \in \mathcal{U}$ are all distinct.

The pdf of $\boldsymbol{wS}$ can be written as

$$p(y) = \sum_{\boldsymbol{u} \in \mathcal{U}} P(\boldsymbol{U} = \boldsymbol{u}) \frac{1}{\sigma} \Phi\left(\frac{y - \boldsymbol{wu}}{\sigma}\right); \qquad (19)$$

but some of the $\boldsymbol{wu}, \boldsymbol{u} \in \mathcal{U}$ can be arbitrarily close to each other. In this case, it is of interest to group the corresponding terms in (19) together. Thus, we rewrite $p(y)$ as

$$p(y) = \sum_{n=1}^{N} \sum_{\boldsymbol{u} \in \mathcal{V}_n} P(\boldsymbol{u}) \left[\sum_{\boldsymbol{u} \in \mathcal{V}_n} \frac{P(\boldsymbol{u})}{\sum_{\boldsymbol{u} \in \mathcal{V}_n} P(\boldsymbol{u})} \frac{1}{\sigma} \Phi\left(\frac{y - \boldsymbol{wu}}{\sigma}\right) \right]$$

where $\mathcal{V}_1, \ldots, \mathcal{V}_N$ is a partition of $\mathcal{U}$. This pdf is still of the form (5) with

$$\pi_n = \sum_{\boldsymbol{u} \in \mathcal{V}_n} P(\boldsymbol{u}), \quad K_n(y) = \sum_{\boldsymbol{u} \in \mathcal{V}_n} \frac{P(\boldsymbol{u})}{\pi_n} \frac{1}{\sigma} \Phi\left(\frac{y - \boldsymbol{wu}}{\sigma}\right).$$

The partition $\mathcal{V}_1, \ldots, \mathcal{V}_N$ can and should be chosen so that

$$d(\boldsymbol{w}) \triangleq \min_{1 \leq n \neq m \leq N} \min_{\boldsymbol{u} \in \mathcal{V}_n, \boldsymbol{u}' \in \mathcal{V}_m} |\boldsymbol{wu} - \boldsymbol{wu}'|$$

is bounded below by some given positive number. To this end, note that, as shown in the proof of Lemma 2, $\boldsymbol{w}^*$ is associated with a partition $\mathcal{U}_0^*, \ldots, \mathcal{U}_r^*$ of $\mathcal{U}$ such that $\boldsymbol{w}^*\boldsymbol{u}$ take the same value for all $\boldsymbol{u} \in \mathcal{U}_j^*$ ($1 \leq j \leq r^*$), and the values associated with different $\mathcal{U}_j^*$ and the $\boldsymbol{w}^*\boldsymbol{u}, \boldsymbol{u} \in \mathcal{U}_0^*$, are all distinct. Thus, $\inf_{\boldsymbol{w} \in \mathcal{W}} |\boldsymbol{wu} - \boldsymbol{wu}'| \geq \delta$ for some $\delta > 0$ for all $\boldsymbol{u} \neq \boldsymbol{u}'$ and $\boldsymbol{u}, \boldsymbol{u}'$ do not belong to a same $\mathcal{U}_j^*$, $j = 1, \ldots, r^*$.

We take $N = r^* + \#(\mathcal{U}_0^*)$, where $\#(\mathcal{U}_0^*)$ denotes the number of elements of $\mathcal{U}_0^*$, $\mathcal{V}_j = \mathcal{U}_j^*, j = 1, \ldots, r^*$ and the remaining $\mathcal{V}_j$ to be disjoint sets containing only a single element of $\mathcal{U}_0^*$. Then the partition

$$\{\mathcal{V}_1, \ldots, \mathcal{V}_N\} = \{\{\boldsymbol{u}\}, \boldsymbol{u} \in \mathcal{U}_0^*, \mathcal{U}_1^*, \ldots, \mathcal{U}_{r^*}^*\}$$

satisfies $d(\boldsymbol{w}) \geq \delta, \forall \boldsymbol{w} \in \mathcal{W}$. The above partition is not fine enough in order to apply Lemma 1 and to obtain the desired lower bound of $H(p)$. The application of this lemma with π_n, K_n ($n = 1, \ldots, N$) defined as above, would yield a lower bound involving $h(\boldsymbol{\pi}) = -\sum_{n=1}^{N} \pi_n \log \pi_n$. By construction, $h(\boldsymbol{\pi}) = h(\boldsymbol{w}^*\boldsymbol{U})$ while we would need a strict inequality. By using a finer partition, one would get a higher value of $h(\boldsymbol{\pi})$. We thus refine the partition $\{\mathcal{V}_1, \ldots, \mathcal{V}_N\}$ by splitting one of the sets $\mathcal{V}_j = \mathcal{U}_j^*, j = 1, \ldots, r^*$ into two subsets. The splitting rule is as follows: for each $\mathcal{U}_j^*$ arrange the $\boldsymbol{wu}, \boldsymbol{u} \in \mathcal{U}_j^*$ in ascending order and look for the maximum gap between two consecutive values. The set $\mathcal{U}_j^*$ that produces the largest gap will be split and the splitting is done at the gap. For $\boldsymbol{w} \in \mathcal{W} \setminus \mathcal{W}'$, this maximum gap can be bounded below by a positive number δ' (noting that there is only a finite number of elements in each $\mathcal{U}_j^*$); hence for the refined partition, $d(\boldsymbol{w}) \geq \min(\delta, \delta')$. Of course, the partition constructed this way depends on $\boldsymbol{w}$, but there can be only a finite number of possible partitions. Hence, one can find a finite number of subsets $\mathcal{W}_1, \ldots, \mathcal{W}_q$ which cover $\mathcal{W} \setminus \mathcal{W}'$, each of which is associated with a partition of $\mathcal{U}$ such that the corresponding $d(\boldsymbol{w})$ is bounded below by $\min(\delta, \delta')$ for all $\boldsymbol{w}$ in this subset. In the following, we shall restrict $\boldsymbol{w}$ to one such subset, $\mathcal{W}_p$ say, and we denote by $\mathcal{V}_1, \ldots, \mathcal{V}_N$ the associated partition.[2]

We now apply Lemma 1 with the sets Ω_n defined by

$$\Omega_n \triangleq \left\{ y : \min_{\boldsymbol{u} \in \mathcal{V}_n} |y - \boldsymbol{wu}| < d(\boldsymbol{w})/2 \right\}.$$

Then we have, writing d in place of $d(\boldsymbol{w})$ for short

$$\epsilon_n \leq 1 - \int_{-d/(2\sigma)}^{d/(2\sigma)} \Phi(x) dx = \text{Erfc}\left(\frac{d}{2\sqrt{2}\sigma}\right)$$

$$\epsilon'_n = \sum_{\boldsymbol{u} \in \mathcal{V}_n} \frac{P(\boldsymbol{u})}{\pi_n} \int_{\mathbb{R} \setminus \Omega_n} \frac{1}{\sigma} \Phi\left(\frac{y - \boldsymbol{wu}}{\sigma}\right) \log \frac{\sup K_n}{K_n(y)} dy.$$

In each term in the sum in that last right-hand side, one applies the bound

$$\frac{\sup K_n}{K_n(y)} \leq \frac{\sigma \sup K_n}{[P(\boldsymbol{u})/\pi_n] \Phi[(y - \boldsymbol{wu})/\sigma]}$$

which yields

$$\epsilon'_n \leq \sum_{\boldsymbol{u} \in \mathcal{V}_n} \frac{P(\boldsymbol{u})}{\pi_n} \int_{|x| > d/(2\sigma)} \Phi(x) \log \frac{\sigma \sup K_n}{[P(\boldsymbol{u})/\pi_n] \Phi(x)} dx$$

$$= \left[\log \sup(\sigma K_n) - \sum_{\boldsymbol{u} \in \mathcal{V}_n} \frac{P(\boldsymbol{u})}{\pi_n} \log \frac{P(\boldsymbol{u})}{\pi_n} \right]$$

$$\times \text{Erfc}\left(\frac{d}{2\sqrt{2}\sigma}\right) + H(\Phi) - H_{d/\sigma}(\Phi).$$

Therefore, putting $h_n \triangleq -\sum_{\boldsymbol{u} \in \mathcal{V}_n} [P(\boldsymbol{u})/\pi_n] \log[P(\boldsymbol{u})/\pi_n]$ and noting that $\sup(\sigma K_n) \leq \sup \Phi = (2\pi)^{-1/2}$, one gets

$$\sum_{n=1}^{N} \pi_n \epsilon'_n + \sum_{n=1}^{N} \pi_n \left[\log\left(\frac{\max_{1 \leq m \leq N} \sup K_m}{\pi_n \sup K_n}\right) + 1 \right] \epsilon_n$$

$$\leq \left[1 - \frac{\log(2\pi)}{2} + \sum_{n=1}^{N} \pi_n h_n \right] \text{Erfc}\left(\frac{d}{2\sqrt{2}\sigma}\right)$$

$$+ H(\Phi) - H_{d/\sigma}(\Phi).$$

Since $d = d(\boldsymbol{w}) \geq \min(\delta, \delta'), \forall \boldsymbol{w} \in \mathcal{W}_p$, the last inequality shows that for any $\eta > 0$

$$H(p) \geq \sum_{n=1}^{N} \pi_n H(K_n) + h(\boldsymbol{\pi}) - \eta, \qquad \forall \boldsymbol{w} \in \mathcal{W}_p,$$

for σ small enough. On the other hand, since $\log x \leq x - 1$

$$\int \frac{1}{\sigma} \Phi\left(\frac{y - \boldsymbol{wu}}{\sigma}\right) \log \frac{K_n(y)}{\Phi[(y - \boldsymbol{wu})/\sigma]/\sigma} dy \leq 0.$$

Multiplying both members of the above inequality by $P(\boldsymbol{u})/\pi_n$ and summing up with respect to $\boldsymbol{u} \in \mathcal{V}_n$, one gets $H(\Phi) + \log \sigma - H(K_n) \leq 0$. Therefore

$$H(p) \geq H(\Phi) + \log \sigma + h(\boldsymbol{\pi}) - \eta.$$

But by construction $h(\boldsymbol{\pi}) > h(\boldsymbol{w}^*\boldsymbol{U})$ (see the proof of Lemma 2); therefore, taking $\eta < h(\boldsymbol{\pi}) - h(\boldsymbol{w}^*\boldsymbol{U})$, one sees that for σ small enough $H(\boldsymbol{wS}) = H(p) > H(\boldsymbol{w}^*\boldsymbol{S})$ for all $\boldsymbol{w} \in \mathcal{W}_p$. Since this is true for all $p = 1, \ldots, q$, we conclude as before that $H(\boldsymbol{wS})$ admits a local minimum in $\mathcal{W}'$. $\qquad \square$

---

[2]Note that the partition obtained *after the split* obviously counts one more element than the corresponding partition *before the split*. However, the same symbol N is used for both partitions to simplify the notation.

REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Willey, 2001.

[3] S. Haykin, Ed., *Unsupervised Adaptive Filtering: Blind Source Separation*. New York: Willey, 2000, vol. 1.

[4] A. Cichocki and S.-I Amari, *Adaptive Blind Signal and Image Processing*. London, U.K.: Willey, 2002.

[5] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: Wiley, 1991.

[6] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1991.

[7] D.-T. Pham, "Mutual information approach to blind separation of stationary sources," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1935–1946, Jul. 2002.

[8] N. Delfosse and P. Loubaton, "Adaptibe blind separation of sources: A deflation approach," *Signal Process.*, vol. 45, pp. 59–83, 1995.

[9] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, Jul. 1999.

[10] S. Cruces, A. Cichocki, and S. Amari, "From blind signal extraction to blind instantaneous signal separation: Criteria, algorithms and stability," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 859–873, Jul. 2004.

[11] D.-T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. Signal Process.*, vol. 44, no. 11, pp. 2768–2779, Nov. 1996.

[12] R. M. Gray and L. D. Davisson, *An Introduction to Statistical Signal Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[13] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1501–1518, Nov. 1991.

[14] F. Vrins and M. Verleysen, "On the entropy minimization of a linear mixture of variables for source separation," *Signal Processing*, vol. 85, no. 5.

[15] ——, "Information theoretic vs cumulant-based contrasts for multimodal source separation," *IEEE Signal Process. Lett.*, vol. 12, no. 3, pp. 190–193, Mar. 2005.

[16] E. G. Learned-Miller and J. W. Fisher III, "ICA using spacings estimates of entropy," *J, Machine Learning Res.*, vol. 4, pp. 1271–1295, 2003.

[17] F. Vrins, J. Lee, and M. Verleysen, "Can we always trust entropy minima in the ica context?," in *Proc. Europ. Signal Processing Conf. (EUSIPCO'05)*, Antalya, Turkey, 2005, pp. cr1107.1–cr1107.14.

[18] D.-T. Pham and F. Vrins, "Local minima of information-theoretic criteria in blind source separation," *IEEE Signal Process. Lett.*, vol. 12, no. 11, pp. 788–791, Nov. 2005.

[19] D.-T. Pham, F. Vrins, and M. Verleysen, "Spurious entropy minima for multimodal source separation," in *Proc. Int. Symp. Signal Processing and Applications (ISSPA'05)*, Sidney, Australia, 2005, pp. 37–40.

[20] D.-T. Pham, "Entropy of a variable slightly contaminated with another," *IEEE Signal Process. Lett.*, vol. 12, no. 7, pp. 536–539, Jul. 2005.

[21] B. W. Silverman, *Density Estimation*. London, U.K.: Chapman &Hall/CRC, 1986.

[22] D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley, 1992.

[23] M. HellmanJ. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 368–372, Jul. 1970.

[24] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.