# Supervised Nonnegative Matrix Factorization with Dual-Itakura-Saito and Kullback-Leibler Divergences for Music Transcription

Hideaki Kagami and Masahiro Yukawa
Dept. Electronics and Electrical Engineering, Keio University, Japan

***Abstract*—In this paper, we present a convex-analytic approach to supervised nonnegative matrix factorization (SNMF) based on the Dual-Itakura-Saito (Dual-IS) and Kullback-Leibler (KL) divergences for music transcription. The Dual-IS and KL divergences define convex fidelity functions, whereas the IS divergence defines a nonconvex one. The SNMF problem is formulated as minimizing the divergence-based fidelity function penalized by the $\ell_1$ and row-block $\ell_1$ norms subject to the nonnegativity constraint. Simulation results show that (i) the use of the Dual-IS and KL divergences yields better performance than the squared Euclidean distance and that (ii) the use of the Dual-IS divergence prevents from false alarms efficiently.**

## I. Introduction

Nonnegative matrix factorization (NMF) is an attractive approach to separating a nonnegative matrix into a product of two nonnegative matrices [1]–[3]. In NMF, many divergence measures have been presented [4], [5]. For the musical instrument classification, it has been shown experimentally that a use of the Kullback-Leibler (KL) divergence tends to give better performance compared to the squared Euclidean distance or the Itakura-Saito (IS) divergence [6], [7]. Although the unsupervised NMF approaches have no need to prepare dictionaries, those approaches need to estimate the number of sources prior to factorization, and a failure in the estimation causes severe performance deterioration in general. The supervised NMF (SNMF) approaches [8]–[10] are advantageous from this aspect since the source number is not explicitly used subject to the availability of a dictionary matrix. In music applications, for instance, this is a practical assumption because there exist many music databases available to construct a dictionary. In particular, in [9], the SNMF problem is formulated as a sparse optimization problem, where the task is to find an appropriate activation matrix that is row-sparse (as well as sparse componentwise). An iterative method based on convex analysis has been presented therein to solve the sparse optimization problem. So far, the squared Euclidean distance has solely been employed as a measure of data fidelity in this convex-analytic approach.

In this paper, we investigate a use of the KL and Dual-Itakura-Saito (Dual-IS) divergences. Here, both divergences define convex fidelity functions, whereas the IS divergence defines a nonconvex one. Our optimization problem to solve for SNMF involves two sparsity-promoting non-differentiable regularizers (the $\ell_1$ and row-block $\ell_1$ norms) in addition to the fidelity function and the nonnegativity constraint. We apply the alternating direction method of multipliers (ADMM) [11] to this problem after a certain reformulation. Simulation results show that the proposed approach exhibits excellent performance both in the F-measure and the total error. It turns out, in particular, that the small total errors of the Dual-IS divergence come from the prevention of false alarms.

TABLE I
Special cases of the generalized Alpha-Beta divergence

| $\alpha$ | $\beta$ | $d_{\mathrm{AB}}^{(\alpha,\beta)}(y \mid \hat{y})$ | |
|---|---|---|---|
| 1 | 1 | Squared Euclidean distance | $d_{\mathrm{EUC}}(y \mid \hat{y})$ |
| 1 | 0 | KL divergence | $d_{\mathrm{KL}}(y \mid \hat{y})$ |
| 1 | $-1$ | IS divergence | $d_{\mathrm{IS}}(y \mid \hat{y})$ |
| 0 | 1 | Dual KL divergence | $d_{\mathrm{dKL}}(y \mid \hat{y})$ |
| $-1$ | 1 | Dual IS divergence | $d_{\mathrm{dIS}}(y \mid \hat{y})$ |

## II. Generalized Alpha-Beta Divergence with Particular Examples

For a given $y > 0$ and a variable $\hat{y}$, the squared Euclidean distance, the KL divergence, and the IS divergence are defined respectively as

$$d_{\mathrm{EUC}}(y \mid \hat{y}) := \frac{1}{2}(y-\hat{y})^2, \tag{1}$$

$$d_{\mathrm{KL}}(y \mid \hat{y}) := y\log\frac{y}{\hat{y}} - y + \hat{y}, \tag{2}$$

$$d_{\mathrm{IS}}(y \mid \hat{y}) := \frac{y}{\hat{y}} - \log\frac{y}{\hat{y}} - 1. \tag{3}$$

The generalized Alpha-Beta divergence $d_{\mathrm{AB}}^{(\alpha,\beta)}(y|\hat{y}), \alpha, \beta \in \mathbb{R}$ is presented in [4]. It includes the squared Euclidean distance, the KL and IS divergences, and their duals [12]

$$d_{\mathrm{dKL}}(y \mid \hat{y}) := d_{\mathrm{KL}}(\hat{y} \mid y), \tag{4}$$

$$d_{\mathrm{dIS}}(y \mid \hat{y}) := d_{\mathrm{IS}}(\hat{y} \mid y), \tag{5}$$

as its particular cases. Note here that the KL and IS divergences are asymmetric.

For the musical instrument classification, it has been reported that a use of the KL divergence tends to give better performance compared to the squared Euclidean distance and the IS divergence [6], [7]. It is well known that the fidelity functions based on the squared Euclidean distance and the KL divergence are proximable (i.e., their proximity operators can be computed easily). See, e.g., [13]. Indeed, the fidelity function based on the Dual-IS divergence is also proximable, as shown in Section III. No special attention has, however, been paid to the Dual-IS divergence so far.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

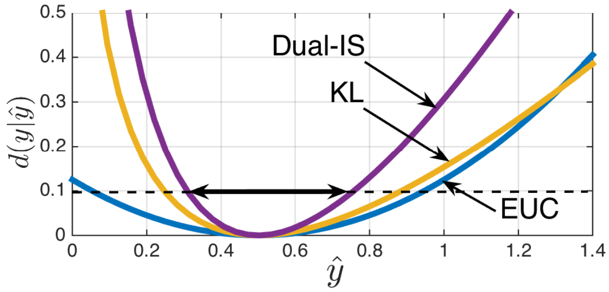
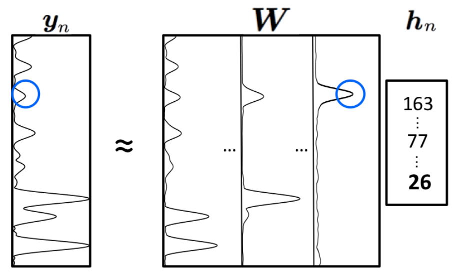
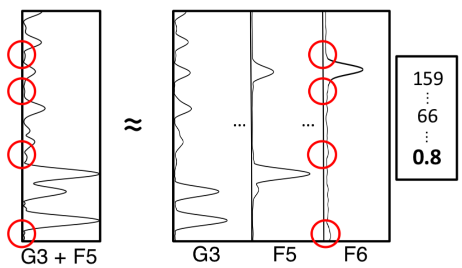



Fig. 1. Illustrations of the squared Euclidean distance and the KL and Dual-IS divergences for $y = 0.5$.




(a) Squared Euclidean distance




(b) Dual-IS (preventing from a false alarm for F6)

Fig. 2. Factorization results based on the Dual-IS divergence and the squared Euclidean distance.

Our focus in the present study is on the KL and Dual-IS divergences.[1] Fig. 1 illustrates the squared Euclidean distance and the KL and Dual-IS divergences as a function of $\hat{y}$ for given $y$. It is seen that the acceptable error range of the Dual-IS divergence is the narrowest among the three curves for $y = 0.5$. (In the figure, the acceptable error range of the Dual-IS divergence is indicated by the bidirectional arrows for the threshold 0.1.) The difference among the three curves becomes larger as $y$ decreases to zero. This implies that the Dual-IS divergence in the SNMF attempts to find from a fixed dictionary a vector that well resembles the coefficients of small amplitudes for each column of the input matrix.

Fig. 2 illustrates how a column $\boldsymbol{y}_n$ of the input matrix $\boldsymbol{Y}$ is factorized. In the case of the squared Euclidean distance, although $\boldsymbol{y}_n$ does not contain the F6 pitch, the coefficient of F6 is large enough to cause a false alarm. This is because the squared Euclidean distance becomes small when the peak is accurately approximated (see the blue circle in Fig.2(a)). In contrast, the Dual-IS divergence correctly suppresses the F6 pitch, because allocating a large coefficient to F6 yields some errors on the small components of $\boldsymbol{y}_n$ (see the red circles in Fig.2(b)) and such errors on the coefficients of small amplitude increase the Dual-IS divergence. This property of the Dual-IS divergence actually leads to considerable reductions of false alarms in music transcriptions.

TABLE II

PROXIMITY OPERATORS OF THE DUAL-IS DIVERGENCE, THE KL DIVERGENCE, AND THE SQUARED EUCLIDEAN DISTANCE.

| $\phi(x)$ | $\operatorname{prox}_{\gamma\phi} x$ |
|---|---|
| $d_{\mathrm{dIS}}(y\,\vert\,x)$ | $p > 0$ s.t. $p^2 + (\gamma y^{-1} - x)p = \gamma$ |
| $d_{\mathrm{KL}}(y\,\vert\,x)$ | $p > 0$ s.t. $p^2 + (\gamma - x)p = \gamma y$ |
| $d_{\mathrm{EUC}}(y\,\vert\,x)$ | $p = \frac{1+\gamma^{-1}}{y + x\gamma^{-1}}$ |

## III. Proximity Operator of Dual-IS Divergence

Fix $y > 0$ arbitrarily in $d_{\mathrm{AB}}^{(\alpha,\beta)}(y \mid x)$ and define the fidelity function $\phi_{\alpha,\beta} : \mathbb{R} \to [0, \infty]$ as

$$\phi_{\alpha,\beta}(x) := \begin{cases} d_{\mathrm{AB}}^{(\alpha,\beta)}(y \mid x) \in [0,\infty) & \text{if } x > 0, \\ +\infty & \text{if } x \le 0. \end{cases} \tag{6}$$

Let $\phi := \phi_{-1,1}$, which is based on the Dual-IS divergence (see Table I). The proximity operator of $\phi$ of index $\gamma > 0$ is defined as follows [13]:

$$\operatorname{prox}_{\gamma\phi} x := \underset{p\in\mathbb{R}}{\operatorname{argmin}} \underbrace{\left(\phi(p) + \frac{1}{2\gamma}(x-p)^2\right)}_{=:F_1(p)}. \tag{7}$$

Here, the proximity operator of $\phi$ is well defined because $\phi$ is proper, lower semi-continuous, and convex.[2] Since the function $F_1(p)$ is strictly convex and also differentiable over $(0, \infty)$, $\operatorname{prox}_{\gamma\phi} x$ can be characterized by $\frac{\partial}{\partial p} F_1(p) = 0$ for $p > 0$, from which it follows that

$$\operatorname{prox}_{\gamma\phi} x = \left\{ p > 0 \mid p^2 + (\gamma y^{-1} - x)p = \gamma \right\}. \tag{8}$$

Since $F_2(p) := p^2 + (\gamma y^{-1} - x)p - \gamma$ is convex and $F_2(0) = -\gamma < 0$, the quadratic equation $F_2(p) = 0$ has a unique positive solution. Table II summarizes the proximity operators of the fidelity functions based on the Dual-IS divergence, the KL divergence, and the squared Euclidean distance.

## IV. Proposed Method

### A. Problem formulation

Let $\mathbb{R}_+$ be the set of nonnegative real numbers. We consider the SNMF problem: given a data matrix $\boldsymbol{Y} \in \mathbb{R}_+^{M\times N}$ to be factorized and a redundant dictionary matrix $\boldsymbol{W} \in \mathbb{R}_+^{M\times L}$, find $\boldsymbol{H} \in \mathcal{C} := \mathbb{R}_+^{L\times N}$ such that $\boldsymbol{Y} \approx \boldsymbol{WH}$. Here, $\boldsymbol{W}$ is assumed to have full column-rank. We formulate the SNMF

[1] The proximity operator of the fidelity function based on the Dual-KL divergence is known to be expressed by using the Lambert W-function [13], [14]. We do not consider the Dual-KL divergence in the present study.

[2] Let $\mathcal{X}$ be a real Hilbert space. A function $f : \mathcal{X} \to (-\infty, \infty]$ is called *proper*, if $\operatorname{dom} f := \{x \in \mathcal{X} \mid f(x) < \infty\} \neq \emptyset$. If the level set $\operatorname{lev}_{\le a} f := \{x \in \mathcal{X} \mid f(x) \le a\}$ is closed for any $a \in \mathbb{R}$, then $f$ is called *lower semi-continuous*. If $f(\eta x + (1-\eta)y) \le \eta f(x) + (1-\eta) f(y)$ for any $x, y \in \mathcal{X}$ and $\eta \in (0, 1)$, then $f$ is called *convex*.

problem as the following convex optimization problem in $\mathcal{H} := \mathbb{R}^{L\times N}$:

$$(\mathrm{P}_0) \quad \min_{\boldsymbol{H}\in\mathcal{H}} \quad \underbrace{\lambda_1 \sum_{l=1}^{L} \left\| \boldsymbol{h}_l^{\mathsf{T}} \right\|_2}_{=:g_1(\boldsymbol{H})} + \underbrace{\lambda_2 \sum_{l=1}^{L}\sum_{n=1}^{N} \mid h_{l,n} \mid}_{=:g_2(\boldsymbol{H})} + \underbrace{i_{\mathcal{C}}(\boldsymbol{H})}_{=:g_3(\boldsymbol{H})} + \underbrace{D_{\mathrm{AB}}^{(\alpha,\beta)}(\boldsymbol{Y} \mid \boldsymbol{WH})}_{=:g_4(\boldsymbol{WH})},$$

where $\boldsymbol{h}_l^{\mathsf{T}}$ denotes the $l$ th row vector of the matrix $\boldsymbol{H}$, $g_1(\boldsymbol{H})$ is the row-block $\ell_1$ norm with the $\ell_2$ norm $\|\cdot\|_2$, $g_2(\boldsymbol{H})$ is the $\ell_1$ norm with the $(l,n)$ entry $h_{l,n}$ of $\boldsymbol{H}$,

$$g_3(\boldsymbol{H}) := i_{\mathcal{C}}(\boldsymbol{H}) := \begin{cases} 0 & \text{if } \ \boldsymbol{H}\in\mathcal{C}, \\ +\infty & \text{otherwise,} \end{cases} \tag{9}$$

is the indicator function to enforce the matrix $\boldsymbol{H}$ to be non-negative, and $g_4(\boldsymbol{WH}){:=}D_{\mathrm{AB}}^{(\alpha,\beta)}(\boldsymbol{Y}|\boldsymbol{WH}) := \sum_{m=1}^{M}\sum_{n=1}^{N} d_{\mathrm{AB}}^{(\alpha,\beta)}\ (y_{m,n}|[\boldsymbol{WH}]_{m,n})$. Here, $y_{m,n}$ (or $[\boldsymbol{WH}]_{m,n}$) is the $(m,n)$ entry of $\boldsymbol{Y}$ (or $\boldsymbol{WH}$). The function $g_4$ is proximable in the cases of $\alpha=-1,\beta=1$ (the Dual-IS divergence), $\alpha=1,\beta=0$ (the KL divergence), and $\alpha=1,\beta=1$ (the squared Euclidean distance). In all cases, the function is convex (see Section III and the reference [4]).

We shall reformulate $(\mathrm{P}_0)$ into a tractable convex optimization problem in the large space $\tilde{\mathcal{H}} := \mathbb{R}^{4L\times N}$. For any matrix $\tilde{\boldsymbol{A}}\in\tilde{\mathcal{H}}$ and for any $(3L+M)\times N$ matrix $\boldsymbol{B}$, we denote by $\tilde{\boldsymbol{A}}^{(1)},\tilde{\boldsymbol{A}}^{(2)},\tilde{\boldsymbol{A}}^{(3)},\tilde{\boldsymbol{A}}^{(4)}\in\mathcal{H}$, $\boldsymbol{B}^{(1)},\boldsymbol{B}^{(2)},\boldsymbol{B}^{(3)}\in\mathcal{H}$, and $\boldsymbol{B}^{(4)}\in\mathbb{R}^{M\times N}$ their partitioned submatrices such that

$$\tilde{\boldsymbol{A}}{=}\begin{bmatrix}\tilde{\boldsymbol{A}}^{(1)}\\ \tilde{\boldsymbol{A}}^{(2)}\\ \tilde{\boldsymbol{A}}^{(3)}\\ \tilde{\boldsymbol{A}}^{(4)}\end{bmatrix}\in\tilde{\mathcal{H}},\quad \boldsymbol{B}{=}\begin{bmatrix}\boldsymbol{B}^{(1)}\\ \boldsymbol{B}^{(2)}\\ \boldsymbol{B}^{(3)}\\ \boldsymbol{B}^{(4)}\end{bmatrix}\in\mathbb{R}^{(3L+M)\times N}. \tag{10}$$

The problem $(P_0)$ can be reformulated equivalently as follows:

$$(\mathrm{P}_1) \quad \min_{\tilde{\boldsymbol{H}}\in\tilde{\mathcal{H}}} \sum_{q=1}^{3} g_q(\tilde{\boldsymbol{H}}^{(q)}) + g_4(\boldsymbol{W}\tilde{\boldsymbol{H}}^{(4)}) \quad \text{s.t.} \quad \tilde{\boldsymbol{H}}^{(1)} = \tilde{\boldsymbol{H}}^{(2)} = \tilde{\boldsymbol{H}}^{(3)} = \tilde{\boldsymbol{H}}^{(4)}. \tag{11}$$

Here, the linear constraints in (11) can be expressed as

$$\tilde{\boldsymbol{H}}\in\mathcal{M} := \left\{ \left[\, \boldsymbol{H}^{\mathsf{T}}\ \boldsymbol{H}^{\mathsf{T}}\ \boldsymbol{H}^{\mathsf{T}}\ \boldsymbol{H}^{\mathsf{T}} \,\right]^{\mathsf{T}} \in\tilde{\mathcal{H}} \mid \boldsymbol{H}\in\mathcal{H} \right\}. \tag{12}$$

The problem $(\mathrm{P}_1)$ can therefore be reformulated equivalently as follows:

$$(\mathrm{P}_2) \quad \min_{\tilde{\boldsymbol{H}}\in\tilde{\mathcal{H}}} \quad i_{\mathcal{M}}(\tilde{\boldsymbol{H}}) + g(\boldsymbol{G}\tilde{\boldsymbol{H}}), \tag{13}$$

where $i_{\mathcal{M}}(\cdot)$ is the indicator function defined as in (9),

$$\boldsymbol{G} := \mathrm{diag}(\boldsymbol{I},\boldsymbol{I},\boldsymbol{I},\boldsymbol{W}) \in\mathbb{R}_{+}^{(3L+M)\times 4L} \tag{14}$$

TABLE III
ADMM TO SOLVE $(\mathrm{P}_2)$

**Requirement:** $\gamma>0$
**Initialization:** $\boldsymbol{Q}_0,\boldsymbol{R}_0$ such that $\boldsymbol{Q}_0-\boldsymbol{R}_0\in\mathbb{R}_{+}^{(3L+M)\times N}$

For $k=0,1,2...$

(a) $\tilde{\boldsymbol{H}}_k = \underset{\tilde{\boldsymbol{Z}}\in\tilde{\mathcal{H}}}{\mathrm{argmin}} \left[ i_{\mathcal{M}}(\tilde{\boldsymbol{Z}}) + \frac{1}{2\gamma}\left\| \boldsymbol{G}\tilde{\boldsymbol{Z}} - (\boldsymbol{Q}_k - \boldsymbol{R}_k) \right\|_{\mathrm{F}}^2 \right]$
(b) $\boldsymbol{Q}_{k+1} = \mathrm{prox}_{\gamma g}(\boldsymbol{G}\tilde{\boldsymbol{H}}_k + \boldsymbol{R}_k)$
(c) $\boldsymbol{R}_{k+1} = \boldsymbol{R}_k + \boldsymbol{G}\tilde{\boldsymbol{H}}_k - \boldsymbol{Q}_{k+1}$

$\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

is the block diagonal matrix whose block-diagonal entries are given by $\boldsymbol{I},\boldsymbol{I},\boldsymbol{I}$, and $\boldsymbol{W}$, where $\boldsymbol{I}$ is the $L\times L$ identity matrix, and

$$g(\boldsymbol{G}\tilde{\boldsymbol{H}}) := \sum_{q=1}^{3} g_q(\tilde{\boldsymbol{H}}^{(q)}) + g_4(\boldsymbol{W}\tilde{\boldsymbol{H}}^{(4)}). \tag{15}$$

As both functions $i_{\mathcal{M}}$ and $g$ are proximable, $(\mathrm{P}_2)$ can be solved by ADMM, which is presented in the following subsection.

### *B. ADMM for Problem ($P_2$)*

The ADMM algorithm to solve $(\mathrm{P}_2)$ is presented in Table III. Note that $\tilde{\boldsymbol{H}}_k$ in the table is well-defined since the minimizer is unique due to the strict convexity of the quadratic term, which can be verified by the nonsingularity of $\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W}$. Steps (a) and (b) are elaborated below.

(a) Due to the presence of $i_{\mathcal{M}}(\tilde{\boldsymbol{Z}})$, it is guaranteed that $\tilde{\boldsymbol{Z}}^{(1)} = \tilde{\boldsymbol{Z}}^{(2)} = \tilde{\boldsymbol{Z}}^{(3)} = \tilde{\boldsymbol{Z}}^{(4)} \in\mathcal{H}$. Hence, it can be verified that $\tilde{\boldsymbol{H}}_k^{\mathsf{T}} = \left[ \tilde{\boldsymbol{H}}_*^{\mathsf{T}}\ \tilde{\boldsymbol{H}}_*^{\mathsf{T}}\ \tilde{\boldsymbol{H}}_*^{\mathsf{T}}\ \tilde{\boldsymbol{H}}_*^{\mathsf{T}} \right]$ with

$$\tilde{\boldsymbol{H}}_* := (3\boldsymbol{I} + \boldsymbol{W}^{\mathsf{T}}\boldsymbol{W})^{-1} \times \left[ \sum_{q=1}^{3} (\boldsymbol{Q}_k - \boldsymbol{R}_k)^{(q)} + \boldsymbol{W}^{\mathsf{T}}(\boldsymbol{Q}_k - \boldsymbol{R}_k)^{(4)} \right]. \tag{16}$$

(b) Let $\boldsymbol{S} := \boldsymbol{G}\tilde{\boldsymbol{H}}_k + \boldsymbol{R}_k \in\mathbb{R}^{(3L+M)\times N}$. Each submatrix of $\boldsymbol{Q}_{k+1}$ is then given as follows (see (10)):

$$\boldsymbol{Q}_{k+1}^{(1)} = \mathrm{prox}_{\gamma g_1}(\boldsymbol{S}^{(1)}) = \sum_{l=1}^{L} \max\left\{ 1 - \frac{\lambda_1\gamma}{\left\| \boldsymbol{s}_l^{(1)} \right\|_2}, 0 \right\} \boldsymbol{e}_l {\boldsymbol{s}_l^{(1)}}^{\mathsf{T}},$$

$$\boldsymbol{Q}_{k+1}^{(2)} = \mathrm{prox}_{\gamma g_2}(\boldsymbol{S}^{(2)}) = \sum_{l=1}^{L}\sum_{n=1}^{N} \mathrm{sgn}(s_{l,n}^{(2)}) \max\left\{ |s_{l,n}^{(2)}| - \lambda_2\gamma, 0 \right\} \boldsymbol{E}_{l,n},$$

$$\boldsymbol{Q}_{k+1}^{(3)} = \mathrm{prox}_{\gamma g_3}(\boldsymbol{S}^{(3)}) = P_{\mathcal{C}}(\boldsymbol{S}^{(3)}) = \begin{bmatrix} \max\left\{ s_{1,1}^{(3)}, 0 \right\} & \cdots & \max\left\{ s_{1,N}^{(3)}, 0 \right\} \\ \vdots & \ddots & \vdots \\ \max\left\{ s_{L,1}^{(3)}, 0 \right\} & \cdots & \max\left\{ s_{L,N}^{(3)}, 0 \right\} \end{bmatrix},$$

$$
\begin{aligned}
\boldsymbol{Q}_{k+1}^{(4)} &= \operatorname{prox}_{\gamma g_4}(\boldsymbol{S}^{(4)}) \\
&= \begin{cases}
\displaystyle \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\Bigg[-\frac{\gamma}{y_{m,n}}+s_{m,n}^{(4)} \\
\displaystyle \quad +\sqrt{\left(\frac{\gamma}{y_{m,n}}-s_{m,n}^{(4)}\right)^2+4\gamma}\,\Bigg]\boldsymbol{E}_{m,n}, & \text{(Dual-IS)} \\
\displaystyle \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\Bigg[-\gamma+s_{m,n}^{(4)} \\
\displaystyle \quad +\sqrt{\left(\gamma-s_{m,n}^{(4)}\right)^2+4\gamma y_{m,n}}\,\Bigg]\boldsymbol{E}_{m,n}. & \text{(KL)}
\end{cases}
\end{aligned}
$$

Here, $\boldsymbol{S}^{(1)} = [\boldsymbol{s}_1^{(1)}\ \boldsymbol{s}_2^{(1)}\ \dots\ \boldsymbol{s}_L^{(1)}]^\top$, $\{\boldsymbol{e}_l\}_{l=1}^{L}$ denotes the standard basis of $\mathbb{R}^L$, $s_{l,n}^{(q)}$ denotes the $(l,n)$ entry of $\boldsymbol{S}^{(q)}$,

$$
\operatorname{sgn}(s_{l,n}^{(2)}) := \begin{cases} s_{l,n}^{(2)}/\left|s_{l,n}^{(2)}\right| & \text{if } s_{l,n}^{(2)} \neq 0, \\ 0 & \text{if } s_{l,n}^{(2)} = 0, \end{cases} \tag{17}
$$

is the signum function, $\boldsymbol{E}_{l,n}$ (or $\boldsymbol{E}_{m,n}$) is the $L \times N$ (or $M \times N$) matrix that has one at the $(l,n)$ position (or the $(m,n)$ position) and zeros elsewhere.

A remarkable advantage of the convex analytic approach is the guarantee of the global convergence. Indeed, a qualification condition[3] (to guarantee the convergence) [16]

$$
\operatorname{int}(\operatorname{dom} g) \cap \boldsymbol{G}(\operatorname{dom} i_{\mathcal{M}}) \neq \emptyset \tag{18}
$$

is satisfied as shown below. Here,

$$
\operatorname{int}(\operatorname{dom} g) = \mathcal{H} \times \mathcal{H} \times \tilde{\mathcal{C}} \times \mathbb{R}^{M\times N} \tag{19}
$$

is the interior of dom $g$ (the domain of $g$) with $\tilde{\mathcal{C}} := \operatorname{int}(\mathcal{C})$, and

$$
\begin{aligned}
\boldsymbol{G}(\operatorname{dom} i_{\mathcal{M}}) &= \boldsymbol{G}(\mathcal{M}) := \{\boldsymbol{G}\tilde{\boldsymbol{H}} \mid \tilde{\boldsymbol{H}} \in \mathcal{M}\} \\
&= \left\{\left[\,\boldsymbol{H}^\top\ \boldsymbol{H}^\top\ \boldsymbol{H}^\top\ (\boldsymbol{W}\boldsymbol{H})^\top\,\right]^\top \middle| \boldsymbol{H} \in \mathcal{H}\right\}. 
\end{aligned} \tag{20}
$$

Hence, it follows that

$$
\begin{aligned}
(18) &\Leftrightarrow \tilde{\mathcal{C}} \times \tilde{\mathcal{C}} \times \tilde{\mathcal{C}} \times \boldsymbol{W}(\tilde{\mathcal{C}}) \neq \emptyset \\
&\Leftrightarrow \tilde{\mathcal{C}} \neq \emptyset.
\end{aligned} \tag{21}
$$

The set $\tilde{\mathcal{C}}$ of positive-valued matrices is clearly nonempty, which verifies (18).

## V. Simulation Results

### A. Simulation Conditions

We show the efficacy of the proposed approach for music transcription. We compose four different patterns (A–D) by using several tones from RWC music database [17]. The input matrix $\boldsymbol{Y}$ for each pattern is the magnitude spectrogram computed by the short-time Fourier transform (STFT) using a Hamming window of length 23 ms with 50% overlap. The columns of the basis matrix $\boldsymbol{W}$ are composed of piano sounds of 88 pitches, and are computed by STFT in the same way as for the input matrix.

[3]The qualification condition (18) can be weakened by using the concept of relative interiors [13], [15].

TABLE IV
The number of sources, duration, and parameter settings.

| | #sources | duration | | Proposed-Dual-IS | Proposed-KL | GFBS-EUC |
|---|---|---|---|---|---|---|
| pattern A | 4 | 13 sec. | $\lambda_1$ | 0.93 | 0.93 | 150 |
| | | | $\lambda_2$ | 0.43 | 0.95 | 25 |
| pattern B | 4 | 15 sec. | $\lambda_1$ | 0.97 | 1.0 | 150 |
| | | | $\lambda_2$ | 0.49 | 0.38 | 25 |
| pattern C | 4 | 23 sec. | $\lambda_1$ | 0.73 | 0.76 | 150 |
| | | | $\lambda_2$ | 0.97 | 0.77 | 10 |
| pattern D | 3 | 23 sec. | $\lambda_1$ | 0.75 | 0.98 | 150 |
| | | | $\lambda_2$ | 0.92 | 0.98 | 20 |

TABLE V
The evaluation results in the F-measure, the total error, and the false alarms.

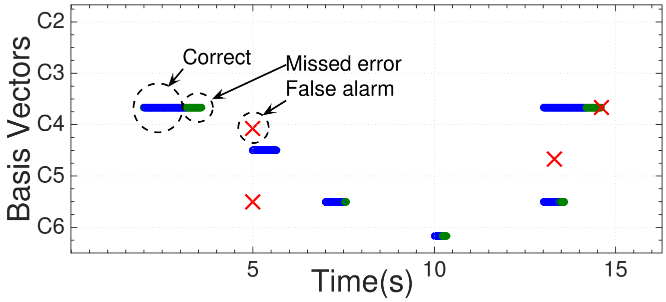
| | | Proposed-Dual-IS | Proposed-KL | GFBS-EUC | BND-KL |
|---|---|---|---|---|---|
| pattern A | $\mathcal{F}$ | **95.5** | 94.2 | 88.5 | 91.5 |
| | $\mathcal{E}_{\text{tot}}$ | **8.35** | 10.8 | 21.0 | 16.0 |
| | $\mathcal{E}_{\text{fals}}$ | **1.60** | 5.68 | 2.49 | 9.59 |
| pattern B | $\mathcal{F}$ | 88.3 | **92.3** | 85.5 | 91.4 |
| | $\mathcal{E}_{\text{tot}}$ | 21.2 | **15.0** | 26.5 | 17 |
| | $\mathcal{E}_{\text{fals}}$ | **0.88** | 6.84 | 4.64 | 9.49 |
| pattern C | $\mathcal{F}$ | **93.7** | 92.0 | 85.1 | 90.7 |
| | $\mathcal{E}_{\text{tot}}$ | **12.2** | 15.2 | 27.0 | 17.5 |
| | $\mathcal{E}_{\text{fals}}$ | **2.72** | 3.13 | 4.44 | 3.45 |
| pattern D | $\mathcal{F}$ | **94.6** | 94.5 | 87.2 | 91.6 |
| | $\mathcal{E}_{\text{tot}}$ | **10.4** | 10.5 | 23.4 | 16.1 |
| | $\mathcal{E}_{\text{fals}}$ | **1.61** | 2.6 | 3.1 | 4.71 |

We compare the proposed method with (i) the generalized forward-backward splitting method to solve the problem ($\text{P}_0$) with the squared Euclidean distance in [9] (GFBS-EUC), and (ii) the multiplicative algorithm to solve the SNMF with the KL divergence (BND-KL) [18]. For each algorithm, the output matrix is binarized with the threshold 5% of the maximum value of the output matrix.

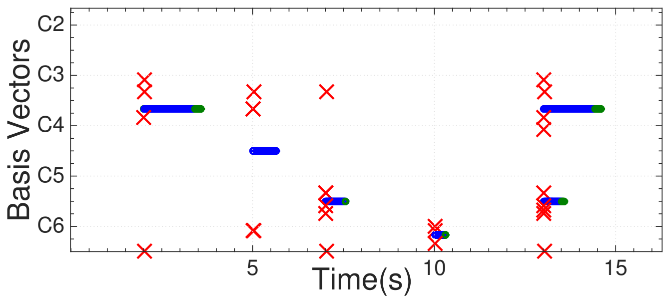
### B. Results and Discussion

All algorithms are run for 300 iterations for $\gamma := 10$. The parameters of each algorithm are chosen to attain the best performance in total errors. Table IV shows the number of sources existing in each pattern and the parameters of each algorithm. Table V summarizes the results in the standard evaluation metrics (see [19]). It can be seen that, the Dual-IS divergence gives the best performance in the total errors for patterns A, C, and D, while the KL divergence does for pattern B. It should be remarked that the Dual-IS divergence achieves the smallest scores in false alarms for all patterns at the expense of some increases of missed errors (i.e., $\mathcal{E}_{\text{tot}} - \mathcal{E}_{\text{fals}}$).

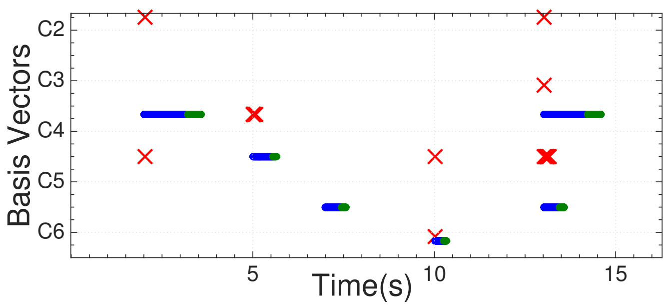
To show that the use of the Dual-IS divergence leads to the prevention of the false alarms, we plot in Fig. 3 the resulting $\boldsymbol{H}$ for pattern B, highlighting the C2 – C6 pitches. In the figures, the false alarms are marked by "x" in red color, the missed errors are plotted in green color; the blue lines indicate that the true pitches are correctly detected. One can see that the proposed algorithm contains only a few false alarms. This verifies the small errors of the Dual-IS divergence in false alarms. If one tries to reduce the false alarms in Proposed-KL or GFBS-EUC by tuning the threshold, the total errors will be increased considerably. This implies that the use of the Dual-IS divergence is a reasonable way to prevent from the false alarms.



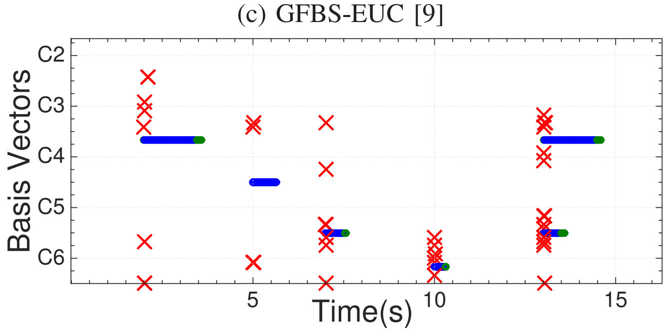

(a) Proposed-Dual-IS




(b) Proposed-KL




(c) GFBS-EUC [9]




(d) BND-KL [18]

Fig. 3. Simulation results for pattern B.

## VI. Conclusion

We have studied the use of the Dual-IS and KL divergences in SNMF. The proximity operator of the fidelity function based on the Dual-IS divergence has been derived. The problem has been formulated as a minimization problem of the divergence-based fidelity function penalized by the three terms: the $\ell_1$ norm, the row-block $\ell_1$ norms, and the indicator function to enforce the nonnegativity. ADMM has been applied to the problem reformulated in the large Euclidean space. The simulation results have demonstrated that (i) the use of the Dual-IS and KL divergences yields better performance than the squared Euclidean distance, and that (ii) the use of the Dual-IS divergence prevents from the false alarms.

## Acknowledgment

This work was supported by the Support Center for Advanced Telecommunications Technology Research (SCAT) and JSPS Grants-in-Aid (15K06081, 15H02757).

## References

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, pp. 556–562. 2001.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. WASPAA*, Oct. 2003, pp. 177–180.

[3] D. L. Sun and C. Fevotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," in *Proc. IEEE ICASSP*, May 2014, pp. 6201–6205.

[4] A. Cichocki, S. Cruces, and S. Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134, 2011.

[5] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations : Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 1st edition, 2009.

[6] D. Fitzgerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Proc. Irish Signals Syst. Conf.*, 2009.

[7] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. 13, no. 5, pp. 1113–1118, 2014.

[8] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. LVA/ICA 2010*, 2010, pp. 140–148.

[9] Y. Morikawa and M. Yukawa, "A sparse optimization approach to supervised nmf based on convex analytic method," in *Proc. IEEE ICASSP*, May 2013, pp. 6078–6082.

[10] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *Proc. 11th International Society for Music Information Retrieval Conference (ISMIR)*, Aug. 2010, pp. 489–494.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[12] I. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with bregman divergences," in *Proc. Advances in Neural Information Processing Systems*, pp. 283–290. MIT Press, 2006.

[13] P. L. Combettes and J. C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*, pp. 185–212. Springer, 2011.

[14] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[15] H. H. Bauschke and P. L. Combettes, *Convex Analysis And Monotone Operator Theory in Hilbert Spaces*, Springer, New York: NY, 1st edition, 2011.

[16] N. Komodakis and J. C. Pesquet, "Playing with duality: An overview of recent primal?dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, Nov 2015.

[17] M. Goto, "Development of the RWC music database," in *Proc. ICA*, 2004, pp. 553–556.

[18] A. Dessein, A. Cont, and G. Lemaitr, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, pp. 341–371. Springer, 2012.

[19] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *Proc. ISMIR*, 2009, pp. 315–320.