Proximity operators of discrete information divergences

Mireille El Gheche, Giovanni Chierchia, Jean-Christophe Pesquet

Proximity Operators of Discrete Information Divergences

Mireille El Gheche, Giovanni Chierchia, *Member, IEEE*, and Jean-Christophe Pesquet, *Fellow, IEEE*

***Abstract*—While φ-divergences have been extensively studied in convex analysis, their use in optimization problems often remains challenging. In this regard, one of the main shortcomings of existing methods is that the minimization of φ-divergences is usually performed with respect to one of their arguments, possibly within alternating optimization techniques. In this paper, we overcome this limitation by deriving new closed-form expressions for the proximity operator of such two-variable functions. This makes it possible to employ standard proximal methods for efficiently solving a wide range of convex optimization problems involving φ-divergences. In addition, we show that these proximity operators are useful to compute the epigraphical projection of several functions. The proposed proximal tools are numerically validated in the context of optimal query execution within database management systems, where the problem of selectivity estimation plays a central role. Experiments are carried out on small to large scale scenarios.**

***Index Terms*—Convex Optimization, Divergences, Proximity Operator, Proximal Algorithms, Epigraphical Projection.**

I. Introduction

DIVERGENCE measures play a crucial role in evaluating the dissimilarity between two information sources. The idea of quantifying how much information is shared between two probability distributions can be traced back to the work by Pearson [3] and Hellinger [4]. Later, Shannon [5] introduced a powerful mathematical framework that links the notion of information with communications and related areas, laying the foundations for information theory. In this context, a key measure of information is the Kullback-Leibler divergence [6], which can be regarded as an instance of the wider class of φ-divergences [7]–[9], including also Jeffreys, Hellinger, Chi-square, Rényi, and I_α divergences [10].

The Kullback-Leibler (KL) divergence has been known to play a prominent role in the computation of channel capacity and rate-distortion functions [11], [12]. These problems can be addressed either with alternating minimization approaches [13], [14] or geometric programming [15]. The KL divergence was also used as a metric for maximizing a log-likelihood in a proximal method generalizing the EM algorithm [16], but here one of its two variables is fixed. The generalized KL divergence (also called I-divergence) is widely used in inverse problems for recovering a signal of interest from an observation degraded by Poisson noise. In such a case, the generalized KL divergence is employed as a data fidelity term, and the resulting optimization approach can be solved through an alternating projection scheme [17]. The problem was formulated in a similar manner by Richardson [18], Lucy [19], and others [20]–[27]. However, in the latter works, one of the two variables of the I-divergence is fixed.

The classical symmetrization of KL divergence, known as Jeffreys (Jef) divergence [28], was recently used in the k-means algorithm as a replacement of the squared difference [29], [30], yielding an analytical expression of the centroids in terms of the Lambert W function. Moreover, tight bounds for this divergence were recently derived in terms of the total variation distance [31], similarly to KL divergence [32].

The Hellinger (Hel) divergence was originally introduced in [33] and later rediscovered under different names [34]–[37]. In the field of information theory, the Hel divergence is commonly used for nonparametric density estimation [38], [39], data analytics [40], and machine learning [41].

The Chi-square divergence was originally used to quantitatively assess whether an observed phenomenon tends to confirm or deny a given hypothesis [3]. It was also successfully applied in different contexts, such as information theory and signal processing, as a dissimilarity measure between two probability distributions [9], [42].

Rényi divergence was introduced as a measure of information related to the Rényi entropy [43], indicating how much a probabilistic mixture of two codes can be compressed [44]. It has been studied and applied in many areas [36], [45], [46], including image registration and alignment problems [47].

The I_α divergence was originally proposed to statistically evaluate the efficiency of an hypothesis test [48]. Subsequently, it was recognized as an instance of the φ-divergences [49] and the Bregman divergences [50], and further extended by many researchers [46], [50]–[52]. This divergence was also considered in the context of non-negative matrix factorization [53].

A. Contributions

To the best of our knowledge, existing approaches for optimizing convex criteria involving φ-divergences are often restricted to specific cases, such as performing the minimization w.r.t. one of the divergence arguments. In order to take into account both arguments, one may resort to alternating minimization schemes, but only in the case when specific assumptions are met. Otherwise, there exist some approaches that exploit the presence of additional moment constraints [54], or the equivalence between φ-divergences and some loss functions [55], but they provide little insight into the numerical procedure for solving the resulting optimization problems.

Mireille El Gheche is with Signal Processing Laboratory (LTS4), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Giovanni Chierchia is with Université Paris-Est, LIGM UMR 8049, CNRS, ENPC, ESIEE Paris, UPEM, Noisy-le-Grand, France.

Jean-Christophe Pesquet is with Center for Visual Computing, CentraleSupélec, University Paris-Saclay, Gif sur Yvette, France.

Part of the material in this paper was presented in [1], [2].

In the context of proximal methods, there exists no general approach for performing the minimization w.r.t. both the arguments of a φ-divergence. This limitation can be explained by the fact that a few number of closed-form expressions are available for the proximity operator of non-separable convex functions, as opposed to separable ones [56], [57]. Some examples of such functions are the Euclidean norm [58], the squared Euclidean norm composed with an arbitrary linear operator [58], a separable function composed with an orthonormal or semi-orthogonal linear operator [58], the max function [59], the quadratic-over-linear function [60]–[62], and the indicator function of some closed convex sets [58], [63].

In this work, we develop a novel proximal approach that allows us to address more general forms of optimization problems involving φ-divergences. Our main contribution is the derivation of new closed-form expressions for the proximity operator of such functions. This makes it possible to employ standard proximal methods [64]–[73] for efficiently solving a wide range of convex optimization problems involving φ-divergences. In addition to its flexibility, the proposed approach leads to parallel algorithms that can be efficiently implemented on both multicore and GPGPU architectures [74].

B. Organization

The remaining of the paper is organized as follows. Section II presents the general form of the optimization problem that we aim at solving. Section III studies the proximity operator of φ-divergences and some of its properties. Section IV details the closed-form expressions of the aforementioned proximity operators. Section V makes the connection with epigraphical projections. Section VI illustrates the application to selectivity estimation for query optimization in database management systems. Finally, Section VII concludes the paper.

C. Notation

Throughout the paper, $\Gamma_0(\mathcal{H})$ denotes the class of convex functions f defined on a real Hilbert space $\mathcal{H}$ and taking their values in $]-\infty,+\infty]$ which are lower-semicontinuous and proper (i.e. their domain $\operatorname{dom} f$ on which they take finite values is nonempty). $\|\cdot\|$ and $\langle\cdot\,|\,\cdot\rangle$ denote the norm and the scalar product of $\mathcal{H}$, respectively. The Moreau subdifferential of f at $x \in \mathcal{H}$ is $\partial f(x) = \big\{u \in \mathcal{H} \;\big|\; (\forall y \in \mathcal{H})\ \langle y - x \,|\, u\rangle + f(x) \leq f(y)\big\}$. If $f \in \Gamma_0(\mathcal{H})$ is Gâteaux differentiable at x, $\partial f(x) = \{\nabla f(x)\}$ where $\nabla f(x)$ denotes the gradient of f at x. The conjugate of f is $f^* \in \Gamma_0(\mathcal{H})$ such that $(\forall u \in \mathcal{H})\ f^*(u) = \sup_{x\in\mathcal{H}} \big(\langle x \,|\, u\rangle - f(x)\big)$. The proximity operator of f is the mapping $\operatorname{prox}_f\colon \mathcal{H} \to \mathcal{H}$ defined as [75]

$$(\forall x \in \mathcal{H}) \qquad \operatorname{prox}_f(x) = \operatorname*{argmin}_{y\in\mathcal{H}}\ f(y) + \frac{1}{2}\|x-y\|^2. \quad (1)$$

Let C be a nonempty closed convex subset of $\mathcal{H}$. The indicator function of C is defined as

$$(\forall x \in \mathcal{H}) \qquad \iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise.} \end{cases} \quad (2)$$

The elements of a vector $x \in \mathcal{H} = \mathbb{R}^N$ are denoted by $x = (x^{(\ell)})_{1\leq\ell\leq N}$, whereas I_N is the $N \times N$ identity matrix.

II. Problem formulation

The objective of this paper is to address convex optimization problems involving a discrete information divergence. In particular, the focus is put on the following formulation.

Problem II.1 Let D be a function in $\Gamma_0(\mathbb{R}^P \times \mathbb{R}^P)$. Let A and B be matrices in $\mathbb{R}^{P\times N}$, and let u and v be vectors in $\mathbb{R}^P$. For every $s \in \{1,\dots,S\}$, let R_s be a function in $\Gamma_0(\mathbb{R}^{K_s})$ and $T_s \in \mathbb{R}^{K_s\times N}$. We want to

$$\operatorname*{minimize}_{x\in\mathbb{R}^N}\ D(Ax+u, Bx+v) + \sum_{s=1}^{S} R_s(T_s x). \quad (3)$$

Note that the functions D and $(R_s)_{1\leq s\leq S}$ are allowed to take the value $+\infty$, so that Problem II.1 can include convex constraints by letting some of the functions R_s be equal to the indicator function ι_{C_s} of some nonempty closed convex set C_s. In inverse problems, R_s may also model some additional prior information, such as the sparsity of coefficients after some appropriate linear transform T_s.

A. Applications in information theory

A special case of interest in information theory arises by decomposing x into two vectors $p \in \mathbb{R}^{P'}$ and $q \in \mathbb{R}^{Q'}$, that is $x = [p^\top\ q^\top]^\top$ with $N = P' + Q'$. Indeed, set $u = v = 0$, $A = [A'\ 0]$ with $A' \in \mathbb{R}^{P\times P'}$, $B = [0\ B']$ with $B' \in \mathbb{R}^{P\times Q'}$ and, for every $s \in \{1,\dots,S\}$, $T_s = [U_s\ V_s]$ with $U_s \in \mathbb{R}^{K_s\times P'}$ and $V_s \in \mathbb{R}^{K_s\times Q'}$. Then, Problem II.1 takes the following form:

Problem II.2 Let A', B', $(U_s)_{1\leq s\leq S}$, and $(V_s)_{1\leq s\leq S}$ be matrices as defined above. Let D be a function in $\Gamma_0(\mathbb{R}^P\times\mathbb{R}^P)$ and, for every $s \in \{1,\dots,S\}$, let R_s be a function in $\Gamma_0(\mathbb{R}^{K_s})$. We want to

$$\operatorname*{minimize}_{(p,q)\in\mathbb{R}^{P'}\times\mathbb{R}^{Q'}}\ D(A'p, B'q) + \sum_{s=1}^{S} R_s(U_s p + V_s q). \quad (4)$$

Several tasks can be formulated within this framework, such as the computation of channel capacity and rate-distortion functions [11], [12], the selection of log-optimal portfolios [76], maximum likelihood estimation from incomplete data [77], soft-supervised learning for text classification [78], simultaneously estimating a regression vector and an additional model parameter [62] or the image gradient distribution and a parametric model distribution [79], as well as image registration [1], deconvolution [2], and recovery [17]. We next detail an important application example in source coding.

Example II.3 Assume that a discrete memoryless source E, taking its values in a finite alphabet $\{e_1,\dots,e_{P_1}\}$ with probability $\mathbb{P}(E)$, is to be encoded by a compressed signal $\widehat{E}$ in terms of a second alphabet $\{\widehat{e}_1,\dots,\widehat{e}_{P_2}\}$. Furthermore, for every $j \in \{1,\dots,P_1\}$ and $k \in \{1,\dots,P_2\}$, let $\delta^{(k,j)}$ be the distortion induced when substituting $\widehat{e}_k$ for e_j. We wish to find an encoding $\mathbb{P}(\widehat{E}|E)$ that yields a point on the rate-distortion curve at a given distortion value $\overline{\delta} \in]0,+\infty[$. It is well-known [80] that this amounts to minimizing the mutual information $\mathcal{I}$

between E and $\widehat{E}$, more precisely the rate-distortion function R is given by

$$\mathsf{R}(\overline{\delta}) = \min_{\mathbb{P}(\widehat{E}|E)} \mathcal{I}(E,\widehat{E}), \tag{5}$$

subject to the constraint

$$\sum_{j=1}^{P_1}\sum_{k=1}^{P_2} \delta^{(k,j)}\, \mathbb{P}(E=e_j)\mathbb{P}(\widehat{E}=\widehat{e}_k|E=e_j) \leq \overline{\delta}. \tag{6}$$

The mutual information can be written as [11, Theorem 4(a)]

$$\min_{\mathbb{P}(\widehat{E})} \sum_{j=1}^{P_1}\sum_{k=1}^{P_2} \mathbb{P}(E=e_j,\widehat{E}=\widehat{e}_k) \ln\left(\frac{\mathbb{P}(\widehat{E}=\widehat{e}_k,E=e_j)}{\mathbb{P}(E=e_j)\mathbb{P}(\widehat{E}=\widehat{e}_k)}\right), \tag{7}$$

subject to the constraint

$$\sum_{k=1}^{P_2} \mathbb{P}(\widehat{E}=\widehat{e}_k) = 1. \tag{8}$$

Moreover, the constraint in (6) can be reexpressed as

$$\sum_{j=1}^{P_1}\sum_{k=1}^{P_2} \delta^{(k,j)}\, \mathbb{P}(E=e_j,\widehat{E}=\widehat{e}_k) \leq \overline{\delta}, \tag{9}$$

with

$$(\forall j \in \{1,\ldots,P_1\}) \quad \sum_{k=1}^{P_2} \mathbb{P}(E=e_j,\widehat{E}=\widehat{e}_k) = \mathbb{P}(E=e_j). \tag{10}$$

The unknown variables are thus the vectors

$$p = \big(\mathbb{P}(E=e_j,\widehat{E}=\widehat{e}_k)\big)_{1\leq j\leq P_1, 1\leq k\leq P_2} \in \mathbb{R}^{P_1P_2} \tag{11}$$

and

$$q = \big(\mathbb{P}(\widehat{E}=\widehat{e}_k)\big)_{1\leq k\leq P_2} \in \mathbb{R}^{P_2}, \tag{12}$$

whose optimal values are solutions to the problem:

$$\underset{p\in C_2\cap C_3, q\in C_1}{\text{minimize}} \quad D(p, r\otimes q) \tag{13}$$

where $r = \big(\mathbb{P}(E=e_j)\big)_{1\leq j\leq P_1} \in \mathbb{R}^{P_1}$, $\otimes$ denotes the Kronecker product, D is the Kullback-Leibler divergence, and C_1, C_2, C_3 are the closed convex sets corresponding to the linear constraints (8), (9), (10), respectively. The above formulation is a special case of Problem II.2 in which $P = P' = P_1P_2$, $Q' = P_2$, $A' = I_P$, B' is such that $(\forall q \in \mathbb{R}^{Q'})\ B'q = r\otimes q$, $S = 3$, $V_1 = I_{Q'}$, $U_2 = U_3 = I_P$, U_1 and $V_2 = V_3$ are null matrices, and $(\forall s \in \{1,2,3\})\ R_s$ is the indicator function of the constraint convex set C_s.

B. Considered class of divergences

We will focus on additive information measures of the form

$$\big(\forall (p,q) \in \mathbb{R}^P \times \mathbb{R}^P\big) \qquad D(p,q) = \sum_{i=1}^{P} \Phi(p^{(i)},q^{(i)}), \tag{14}$$

where $\Phi \in \Gamma_0(\mathbb{R}\times\mathbb{R})$ is the *perspective function* [81] on $[0,+\infty[\,\times\,]0,+\infty[$ of a function $\varphi\colon \mathbb{R}\to[0,+\infty]$ belonging to $\Gamma_0(\mathbb{R})$ and twice differentiable on $]0,+\infty[$. In other words, Φ is defined as follows: for every $(\upsilon,\xi)\in\mathbb{R}^2$,

$$\Phi(\upsilon,\xi) = \begin{cases} \xi\,\varphi\Big(\dfrac{\upsilon}{\xi}\Big) & \text{if } \upsilon\in[0,+\infty[\text{ and } \xi\in\,]0,+\infty[\\ \upsilon \lim\limits_{\zeta\to+\infty} \dfrac{\varphi(\zeta)}{\zeta} & \text{if } \upsilon\in\,]0,+\infty[\text{ and } \xi = 0 \\ 0 & \text{if } \upsilon=\xi=0 \\ +\infty & \text{otherwise,} \end{cases} \tag{15}$$

where the above limit is guaranteed to exist [82, Sec. 2.3]. Moreover, if φ is a strictly convex function such that

$$\varphi(1) = \varphi'(1) = 0, \tag{16}$$

the function D in (14) belongs to the class of φ-divergences [7], [83]. Then, for every $(p,q)\in[0,+\infty[^P\times[0,+\infty[^P$,

$$D(p,q) \geq 0 \tag{17}$$

$$D(p,q) = 0 \quad\Leftrightarrow\quad p = q. \tag{18}$$

Examples of φ-divergences will be provided in Sections IV-A, IV-B, IV-C, IV-D and IV-F. For a thorough investigation of the rich properties of φ-divergences, the reader is refered to [7], [8], [84]. Other divergences (e.g., Rényi divergence) are expressed as

$$\big(\forall (p,q) \in \mathbb{R}^P \times \mathbb{R}^P\big) \qquad D_g(p,q) = g\big(D(p,q)\big) \tag{19}$$

where g is an increasing function. Then, provided that $g\big(\varphi(1)\big) = 0$, $D_g(p,q)\geq 0$ for every $[p^\top\ q^\top]^\top \in C$ with

$$C = \Big\{x\in[0,1]^{2P} \;\Big|\; \sum_{i=1}^{P} x^{(i)} = 1 \text{ and } \sum_{i=1}^{P} x^{(P+i)} = 1\Big\}. \tag{20}$$

From an optimization standpoint, minimizing D or D_g (possibly subject to constraints) makes no difference, hence we will only address problems involving D in the rest of this paper.

C. Proximity operators

Proximity operators will be fundamental tools in this paper. We first recall some of their key properties.

Proposition II.4 [75], [81] *Let $f\in\Gamma_0(\mathcal{H})$. Then,*

(i) *For every $\overline{x}\in\mathcal{H}$, $\operatorname{prox}_f\overline{x}\in\operatorname{dom} f$.*

(ii) *For every $(x,\overline{x})\in\mathcal{H}^2$*

$$x = \operatorname{prox}_f(\overline{x}) \quad\Leftrightarrow\quad \overline{x}-x\in\partial f(x). \tag{21}$$

(iii) *For every $(\overline{x},z)\in\mathcal{H}^2$,*

$$\operatorname{prox}_{f(\cdot+z)}(\overline{x}) = \operatorname{prox}_f(\overline{x}+z) - z. \tag{22}$$

(iv) *For every $(\overline{x},z)\in\mathcal{H}^2$ and for every $\alpha\in\mathbb{R}$,*

$$\operatorname{prox}_{f+\langle z|\cdot\rangle+\alpha}(\overline{x}) = \operatorname{prox}_f(\overline{x}-z). \tag{23}$$

(v) *Let f^* be the conjugate function of f. For every $\overline{x}\in\mathcal{H}$ and for every $\gamma\in\,]0,+\infty[$,*

$$\operatorname{prox}_{\gamma f^*}(\overline{x}) = \overline{x} - \gamma\operatorname{prox}_{f/\gamma}(\overline{x}/\gamma). \tag{24}$$

(vi) *Let $\mathcal{G}$ be a real Hilbert space and let $T\colon \mathcal{G} \to \mathcal{H}$ be a bounded linear operator, with the adjoint denoted by T^*. If $TT^* = \kappa\,\mathrm{Id}$ and $\kappa \in]0,+\infty[$, then for all $\overline{x} \in \mathcal{H}$*

$$\operatorname{prox}_{f\circ T}(\overline{x}) = \overline{x} + \frac{1}{\kappa}T^*\big(\operatorname{prox}_{\kappa f}(T\overline{x}) - T\overline{x}\big). \quad (25)$$

Numerous additional properties of proximity operators are mentioned in [57], [85].

In this paper, we will be mainly concerned with the determination of the proximity operator of the function D defined in (14) with $\mathcal{H} = \mathbb{R}^P \times \mathbb{R}^P$. The next result emphasizes that this task reduces to the calculation of the proximity operator of a real function of two variables.

Proposition II.5 *Let D be defined by* (14) *where $\Phi \in \Gamma_0(\mathbb{R}^2)$ and let $\gamma \in]0,+\infty[$. Let $u \in \mathbb{R}^P$ and $v \in \mathbb{R}^P$. Then, for every $\overline{p} \in \mathbb{R}^P$ and for every $\overline{q} \in \mathbb{R}^P$,*

$$\operatorname{prox}_{\gamma D(\cdot+u,\cdot+v)}(\overline{p},\overline{q}) = (p-u, q-v) \quad (26)$$

where, for every $i \in \{1,\dots,P\}$,

$$(p^{(i)},q^{(i)}) = \operatorname{prox}_{\gamma\Phi}(\overline{p}^{(i)}+u^{(i)},\overline{q}^{(i)}+v^{(i)}). \quad (27)$$

Note that, although an extensive list of proximity operators of one-variable real functions can be found in [57], few results are available for real functions of two variables [58], [60], [61], [63]. An example of such a result is provided below.

Proposition II.6 *Let $\varphi \in \Gamma_0(\mathbb{R})$ be an even differentiable function on $\mathbb{R}\setminus\{0\}$. Let $\Phi\colon \mathbb{R}^2 \to]-\infty,+\infty]$ be defined as:* $(\forall (\nu,\xi) \in \mathbb{R}^2)$

$$\Phi(\nu,\xi) = \begin{cases} \varphi(\nu-\xi) & \text{if } (\nu,\xi) \in [0,+\infty[^2 \\ +\infty & \text{otherwise.} \end{cases} \quad (28)$$

Then, for every $(\overline{\nu},\overline{\xi}) \in \mathbb{R}^2$,

$$\operatorname{prox}_{\Phi}(\overline{\nu},\overline{\xi}) = \begin{cases} \frac{1}{2}\big(\overline{\nu}+\overline{\xi}+\pi_1, \overline{\nu}+\overline{\xi}-\pi_1\big) & \text{if } |\pi_1| < \overline{\nu}+\overline{\xi} \\ (0,\pi_2) & \text{if } \pi_2 > 0 \text{ and } \pi_2 \geq \overline{\nu}+\overline{\xi} \\ (\pi_3,0) & \text{if } \pi_3 > 0 \text{ and } \pi_3 \geq \overline{\nu}+\overline{\xi} \\ (0,0) & \text{otherwise,} \end{cases} \quad (29)$$

with $\pi_1 = \operatorname{prox}_{2\varphi}(\overline{\nu}-\overline{\xi})$, $\pi_2 = \operatorname{prox}_{\varphi}(\overline{\xi})$ and $\pi_3 = \operatorname{prox}_{\varphi}(\overline{\nu})$.

The above proposition provides a simple characterization of the proximity operators of some distances defined for nonnegative-valued vectors. However, the assumptions made in Proposition II.6 are not satisfied by the class of functions Φ considered in Section II-B.[1]

[1]Indeed, none of the considered φ-divergences can be expressed as a function of the difference between the two arguments.

III. Main result

As shown by Proposition II.5, we need to compute the proximity operator of a scaled version of a function $\Phi \in \Gamma_0(\mathbb{R}^2)$ as defined in (15). In the following, Θ denotes a primitive on $]0,+\infty[$ of the function $\zeta \mapsto \zeta\varphi'(\zeta^{-1})$. The following functions will subsequently play an important role:

$$\vartheta_-\colon]0,+\infty[\to \mathbb{R}\colon \zeta \mapsto \varphi'(\zeta^{-1}) \quad (30)$$

$$\vartheta_+\colon]0,+\infty[\to \mathbb{R}\colon \zeta \mapsto \varphi(\zeta^{-1}) - \zeta^{-1}\varphi'(\zeta^{-1}). \quad (31)$$

A first technical result is as follows.

Lemma III.1 *Let $\gamma \in]0,+\infty[$, let $(\overline{v},\overline{\xi}) \in \mathbb{R}^2$, and define*

$$\chi_- = \inf\big\{\zeta \in]0,+\infty[\;\big|\; \vartheta_-(\zeta) < \gamma^{-1}\overline{v}\big\} \quad (32)$$

$$\chi_+ = \sup\big\{\zeta \in]0,+\infty[\;\big|\; \vartheta_+(\zeta) < \gamma^{-1}\overline{\xi}\big\} \quad (33)$$

(with the usual convention $\inf\varnothing = +\infty$ and $\sup\varnothing = -\infty$). If $\chi_- \neq +\infty$, the function

$$\psi\colon]0,+\infty[\to \mathbb{R}\colon \zeta \mapsto \zeta\varphi(\zeta^{-1}) - \Theta(\zeta) + \frac{\gamma^{-1}\overline{v}}{2}\zeta^2 - \gamma^{-1}\overline{\xi}\zeta \quad (34)$$

is strictly convex on $]\chi_-,+\infty[$. In addition, if

(i) $\chi_- \neq +\infty$ *and* $\chi_+ \neq -\infty$

(ii) $\lim_{\substack{\zeta\to\chi_-\\ \zeta>\chi_-}} \psi'(\zeta) < 0$

(iii) $\lim_{\zeta\to\chi_+} \psi'(\zeta) > 0$

then ψ admits a unique minimizer $\widehat{\zeta}$ on $]\chi_-,+\infty[$, and $\widehat{\zeta} < \chi_+$.

Proof. The derivative of ψ is, for every $\zeta \in]0,+\infty[$,

$$\begin{aligned}\psi'(\zeta) &= \varphi(\zeta^{-1}) - (\zeta+\zeta^{-1})\varphi'(\zeta^{-1}) + \gamma^{-1}\overline{v}\zeta - \gamma^{-1}\overline{\xi} \\ &= \zeta\big(\gamma^{-1}\overline{v} - \vartheta_-(\zeta)\big) + \vartheta_+(\zeta) - \gamma^{-1}\overline{\xi}. \end{aligned} \quad (35)$$

The function ϑ_- is decreasing as the convexity of φ yields

$$(\forall \zeta \in]0,+\infty[) \qquad \vartheta_-'(\zeta) = -\zeta^{-2}\varphi''(\zeta^{-1}) \leq 0. \quad (36)$$

This allows us to deduce that

$$\begin{aligned}&\text{if } \big\{\zeta \in]0,+\infty[\;\big|\; \vartheta_-(\zeta) < \gamma^{-1}\overline{v}\big\} \neq \varnothing, \\ &\text{then }]\chi_-,+\infty[= \big\{\zeta \in]0,+\infty[\;\big|\; \vartheta_-(\zeta) < \gamma^{-1}\overline{v}\big\}. \end{aligned} \quad (37)$$

Similarly, the function ϑ_+ is increasing as the convexity of φ yields

$$(\forall \zeta \in]0,+\infty[) \qquad \vartheta_+'(\zeta) = \zeta^{-3}\varphi''(\zeta^{-1}) \geq 0 \quad (38)$$

which allows us to deduce that

$$\begin{aligned}&\text{if } \big\{\zeta \in]0,+\infty[\;\big|\; \vartheta_+(\zeta) < \gamma^{-1}\overline{\xi}\big\} \neq \varnothing, \\ &\text{then }]0,\chi_+[= \big\{\zeta \in]0,+\infty[\;\big|\; \vartheta_+(\zeta) < \gamma^{-1}\overline{\xi}\big\}. \end{aligned} \quad (39)$$

If $(\chi_-,\chi_+) \in]0,+\infty[^2$, then (35) leads to

$$\psi'(\chi_-) = \vartheta_+(\chi_-) - \gamma^{-1}\overline{\xi} \quad (40)$$

$$\psi'(\chi_+) = \chi_+\big(\gamma^{-1}\overline{v} - \vartheta_-(\chi_+)\big). \quad (41)$$

So, Conditions (ii) and (iii) are equivalent to

$$\vartheta_+(\chi_-) - \gamma^{-1}\overline{\xi} < 0 \quad (42)$$

$$\chi_+\big(\gamma^{-1}\overline{v} - \vartheta_-(\chi_+)\big) > 0. \quad (43)$$

In view of (37) and (39), these inequalities are satisfied if and only if $\chi_- < \chi_+$. This inequality is also obviously satisfied if $\chi_- = 0$ or $\chi_+ = +\infty$. In addition, we have: $(\forall \zeta \in]0,+\infty[)$

$$\psi''(\zeta) = \gamma^{-1}\overline{\upsilon} - \vartheta_-(\zeta) + \zeta^{-1}(1+\zeta^{-2})\varphi''(\zeta^{-1}). \quad (44)$$

When $\zeta > \chi_- \neq +\infty$, $\gamma^{-1}\overline{\upsilon} - \vartheta_-(\zeta) > 0$, and the convexity of φ yields $\psi''(\zeta) > 0$. This shows that ψ is strictly convex on $]\chi_-,+\infty[$.

If Conditions (i)-(iii) are satisfied, due to the continuity of ψ', there exists $\widehat{\zeta} \in]\chi_-,\chi_+[$ such that $\psi'(\widehat{\zeta}) = 0$. Because of the strict convexity of ψ on $]\chi_-,+\infty[$, $\widehat{\zeta}$ is the unique minimizer of ψ on this interval. □

The required assumptions in the previous lemma can often be simplified as stated below.

Lemma III.2 *Let $\gamma \in]0,+\infty[$ and $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$. If $(\chi_-,\chi_+) \in]0,+\infty[^2$, then Conditions (ii) and (iii) in Lemma III.1 are equivalent to: $\chi_- < \chi_+$. If $\chi_- \in]0,+\infty[$ and $\chi_+ = +\infty$ (resp. $\chi_- = 0$ and $\chi_+ \in]0,+\infty[$), Conditions (ii)-(iii) are satisfied if and only if $\lim_{\zeta\to+\infty} \psi'(\zeta) > 0$ (resp. $\lim_{\substack{\zeta\to 0\\ \zeta>0}} \psi'(\zeta) < 0$).*

Proof. If $(\chi_-,\chi_+) \in]0,+\infty[^2$, we have already shown that Conditions (ii) and (iii) are satisfied if and only $\chi_- < \chi_+$.

If $\chi_- \in]0,+\infty[$ and $\chi_+ = +\infty$ (resp. $\chi_- = 0$ and $\chi_+ \in]0,+\infty[$), we still have

$$\psi'(\chi_-) = \vartheta_+(\chi_-) - \gamma^{-1}\overline{\xi} < 0 \quad (45)$$

$$(\text{resp. } \psi'(\chi_+) = \chi_+\big(\gamma^{-1}\overline{\upsilon} - \vartheta_-(\chi_+)\big) > 0), \quad (46)$$

which shows that Condition (ii) (resp. Condition (iii)) is always satisfied. □

By using the same expressions of χ_- and χ_+ as in the previous lemmas, we obtain the following characterization of the proximity operator of any scaled version of Φ:

Proposition III.3 *Let $\gamma \in]0,+\infty[$ and $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$. $\operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}) \in]0,+\infty[^2$ if and only if Conditions (i)-(iii) in Lemma III.1 are satisfied. When these conditions hold,*

$$\operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}) = \big(\overline{\upsilon} - \gamma\,\vartheta_-(\widehat{\zeta}), \overline{\xi} - \gamma\,\vartheta_+(\widehat{\zeta})\big) \quad (47)$$

where $\widehat{\zeta} < \chi_+$ is the unique minimizer of ψ on $]\chi_-,+\infty[$.

Proof. For every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$, such that Conditions (i)-(iii) in Lemma III.1 hold, let

$$\upsilon = \overline{\upsilon} - \gamma\,\vartheta_-(\widehat{\zeta}) \quad (48)$$

$$\xi = \overline{\xi} - \gamma\,\vartheta_+(\widehat{\zeta}) \quad (49)$$

where the existence of $\widehat{\zeta} \in]\chi_-,\chi_+[$ is guaranteed by Lemma III.1. As consequences of (37) and (39), υ and ξ are positive. In addition, since

$$\psi'(\widehat{\zeta}) = 0 \quad \Leftrightarrow \quad \widehat{\zeta}\big(\gamma^{-1}\overline{\upsilon} - \vartheta_-(\widehat{\zeta})\big) = \gamma^{-1}\overline{\xi} - \vartheta_+(\widehat{\zeta}) \quad (50)$$

we derive from (48) and (49) that $\widehat{\zeta} = \xi/\upsilon > 0$. This allows us to re-express (48) and (49) as

$$\upsilon - \overline{\upsilon} + \gamma\varphi'\Big(\frac{\upsilon}{\xi}\Big) = 0 \quad (51)$$

$$\xi - \overline{\xi} + \gamma\left(\varphi\Big(\frac{\upsilon}{\xi}\Big) - \frac{\upsilon}{\xi}\varphi'\Big(\frac{\upsilon}{\xi}\Big)\right) = 0, \quad (52)$$

that is

$$\upsilon - \overline{\upsilon} + \gamma\frac{\partial\Phi}{\partial\upsilon}(\upsilon,\xi) = 0 \quad (53)$$

$$\xi - \overline{\xi} + \gamma\frac{\partial\Phi}{\partial\xi}(\upsilon,\xi) = 0. \quad (54)$$

The latter equations are satisfied if and only if [57]

$$(\upsilon,\xi) = \operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}). \quad (55)$$

Conversely, for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$, let $(\upsilon,\xi) = \operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi})$. If $(\upsilon,\xi) \in]0,+\infty[^2$, (υ,ξ) satisfies (51) and (52). By setting $\widetilde{\zeta} = \xi/\upsilon > 0$, after simple calculations, we find

$$\upsilon = \overline{\upsilon} - \gamma\,\vartheta_-(\widetilde{\zeta}) > 0 \quad (56)$$

$$\xi = \overline{\xi} - \gamma\,\vartheta_+(\widetilde{\zeta}) > 0 \quad (57)$$

$$\psi'(\widetilde{\zeta}) = 0. \quad (58)$$

According to (37) and (39), (56) and (57) imply that $\chi_- \neq +\infty$, $\chi_+ \neq -\infty$, and $\widetilde{\zeta} \in]\chi_-,\chi_+[$. In addition, according to Lemma III.1, ψ' is strictly increasing on $]\chi_-,+\infty[$ (since ψ is strictly convex on this interval). Hence, ψ' has a limit at χ_- (which may be equal to $-\infty$ when $\chi_- = -\infty$), and Condition (ii) is satisfied. Similarly, ψ' has a limit at χ_+ (possibly equal to $+\infty$ when $\chi_+ = +\infty$), and Condition (iii) is satisfied. □

Remark III.4 In (15), a special case arises when

$$(\forall \zeta \in]0,+\infty[) \qquad \varphi(\zeta) = \widetilde{\varphi}(\zeta) + \zeta\widetilde{\varphi}(\zeta^{-1}) \quad (59)$$

where $\widetilde{\varphi}$ is a twice differentiable convex function on $]0,+\infty[$. Then Φ takes a symmetric form, leading to $\mathcal{L}$-divergences. It can then be deduced from (31) that, for every $\zeta \in]0,+\infty[$,

$$\vartheta_-(\zeta) = \vartheta_+(\zeta^{-1}) = \widetilde{\varphi}(\zeta) + \widetilde{\varphi}'(\zeta^{-1}) - \zeta\widetilde{\varphi}'(\zeta). \quad (60)$$

IV. Examples

A. Kullback-Leibler divergence

Let us now apply the results in the previous section to the function

$$\Phi(\upsilon,\xi) = \begin{cases} \upsilon\ln\left(\dfrac{\upsilon}{\xi}\right) + \xi - \upsilon & \text{if } (\upsilon,\xi) \in]0,+\infty[^2 \\ \xi & \text{if } \upsilon = 0 \text{ and } \xi \in [0,+\infty[\\ +\infty & \text{otherwise.} \end{cases} \quad (61)$$

This is a function in $\Gamma_0(\mathbb{R}^2)$ satisfying (15) with

$$(\forall \zeta \in]0,+\infty[) \qquad \varphi(\zeta) = \zeta\ln\zeta - \zeta + 1. \quad (62)$$

Proposition IV.1 *The proximity operator of $\gamma\Phi$ with $\gamma \in]0,+\infty[$ is, for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$,*

$$\operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}) = \begin{cases} (\upsilon,\xi) & \text{if } \exp\left(\gamma^{-1}\overline{\upsilon}\right) > 1-\gamma^{-1}\overline{\xi} \\ (0,0) & \text{otherwise,} \end{cases} \tag{63}$$

where

$$\upsilon = \overline{\upsilon} + \gamma \ln \widehat{\zeta} \tag{64}$$

$$\xi = \overline{\xi} + \gamma\left(\widehat{\zeta}^{-1} - 1\right) \tag{65}$$

and $\widehat{\zeta}$ is the unique minimizer on $]\exp(-\gamma^{-1}\overline{\upsilon}),+\infty[$ of

$$\begin{aligned} &\psi\colon\ ]0,+\infty[\to \mathbb{R}\colon \\ &\quad \zeta \mapsto \left(\frac{\zeta^2}{2}-1\right)\ln\zeta + \frac{1}{2}\left(\gamma^{-1}\overline{\upsilon} - \frac{1}{2}\right)\zeta^2 + (1-\gamma^{-1}\overline{\xi})\zeta. \end{aligned} \tag{66}$$

Proof. For every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$, $(\upsilon,\xi) = \operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi})$ is such that $(\upsilon,\xi) \in \operatorname{dom}\Phi$ [86]. Let us first note that

$$\upsilon \in]0,+\infty[\quad \Leftrightarrow \quad (\upsilon,\xi) \in]0,+\infty[^2. \tag{67}$$

We are now able to apply Proposition III.3, where ψ is given by (66) and, for every $\zeta \in]0,+\infty[$,

$$\Theta(\zeta) = \frac{\zeta^2}{2}\left(\frac{1}{2} - \ln\zeta\right) - 1 \tag{68}$$

$$\vartheta_-(\zeta) = -\ln\zeta \tag{69}$$

$$\vartheta_+(\zeta) = 1 - \zeta^{-1}. \tag{70}$$

In addition,

$$\chi_- = \exp(-\gamma^{-1}\overline{\upsilon}) \tag{71}$$

$$\chi_+ = \begin{cases} (1-\gamma^{-1}\overline{\xi})^{-1} & \text{if } \overline{\xi} < \gamma \\ +\infty & \text{otherwise.} \end{cases} \tag{72}$$

According to (67) and Proposition III.3, $\upsilon \in]0,+\infty[$ if and only if Conditions (i)-(iii) in Lemma III.1 hold. Since $\chi_- \in]0,+\infty[$ and $\lim_{\zeta\to+\infty}\psi'(\zeta) = +\infty$, Lemma III.2 shows that these conditions are satisfied if and only if

$$\overline{\xi} \geq \gamma \quad \text{or} \quad \left(\overline{\xi} < \gamma \quad \text{and} \quad \exp(-\overline{\upsilon}/\gamma) < (1-\gamma^{-1}\overline{\xi})^{-1}\right), \tag{73}$$

which is equivalent to

$$\exp(\overline{\upsilon}/\gamma) > 1 - \gamma^{-1}\overline{\xi}. \tag{74}$$

Under this assumption, Proposition III.3 leads to the expressions (64) and (65) of the proximity operator, where $\widehat{\zeta}$ is the unique minimizer on $]\exp(-\overline{\upsilon}/\gamma),+\infty[$ of the function ψ.

We have shown that $\upsilon > 0 \Leftrightarrow$ (74). So, $\upsilon = 0$ when (74) is not satisfied. Then, the expression of ξ simply reduces to the asymmetric soft-thresholding rule [87]:

$$\xi = \begin{cases} \overline{\xi} - \gamma & \text{if } \overline{\xi} > \gamma \\ 0 & \text{otherwise.} \end{cases} \tag{75}$$

However, $\exp(\gamma^{-1}\overline{\upsilon}) \leq 1 - \gamma^{-1}\overline{\xi} \Rightarrow \overline{\xi} < \gamma$, so that ξ is necessarily equal to 0. □

Remark IV.2 More generally, we can derive the proximity operator of

$$\widetilde{\Phi}(\upsilon,\xi) = \begin{cases} \upsilon\ln\left(\dfrac{\upsilon}{\xi}\right) + \kappa(\xi-\upsilon) & \text{if } (\upsilon,\xi) \in]0,+\infty[^2 \\ \kappa\xi & \text{if } \upsilon = 0 \text{ and } \xi \in [0,+\infty[\\ +\infty & \text{otherwise,} \end{cases} \tag{76}$$

where $\kappa \in \mathbb{R}$. Of particular interest in the literature is the case when $\kappa = 0$ [11], [12], [21], [24]. From Proposition II.4(iv), we get, for every $\gamma \in]0,+\infty[$ and for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$,

$$\operatorname{prox}_{\gamma\widetilde{\Phi}}(\overline{\upsilon},\overline{\xi}) = \operatorname{prox}_{\gamma\Phi}(\overline{\upsilon}+\gamma\kappa-\gamma, \overline{\xi}-\gamma\kappa+\gamma), \tag{77}$$

where $\operatorname{prox}_{\gamma\Phi}$ is provided by Proposition IV.1.

Remark IV.3 It can be noticed that

$$\psi'(\widehat{\zeta}) = \widehat{\zeta}\ln\widehat{\zeta} + \gamma^{-1}\overline{\upsilon}\widehat{\zeta} - \widehat{\zeta}^{-1} + 1 - \gamma^{-1}\overline{\xi} = 0 \tag{78}$$

is equivalent to

$$\widehat{\zeta}^{-1}\exp\left(\widehat{\zeta}^{-1}\left(\widehat{\zeta}^{-1} + \gamma^{-1}\overline{\xi} - 1\right)\right) = \exp(\gamma^{-1}\overline{\upsilon}). \tag{79}$$

In the case where $\overline{\xi} = \gamma$, the above equation reduces to

$$\begin{aligned} & 2\widehat{\zeta}^{-2}\exp\left(2\widehat{\zeta}^{-2}\right) = 2\exp(2\gamma^{-1}\overline{\upsilon}) \\ \Leftrightarrow \quad & \widehat{\zeta} = \left(\frac{2}{W(2e^{2\gamma^{-1}\overline{\upsilon}})}\right)^{1/2} \end{aligned} \tag{80}$$

where W is the Lambert W function [88]. When $\overline{\xi} \neq \gamma$, although a closed-form expression of (79) is not available, efficient numerical methods to compute $\widehat{\zeta}$ can be developed.

Remark IV.4 To minimize ψ in (66), we need to find the zero on $]\exp(-\gamma^{-1}\overline{\upsilon}),+\infty[$ of the function: $(\forall\zeta \in]0,+\infty[)$

$$\psi'(\zeta) = \zeta\ln\zeta + \gamma^{-1}\overline{\upsilon}\,\zeta - \zeta^{-1} + 1 - \gamma^{-1}\overline{\xi}. \tag{81}$$

This can be performed by Algorithm 1, the convergence of which is proved in Appendix A.

Algorithm 1 Newton method for minimizing (66).

SET $\widehat{\zeta}_0 = \exp(-\gamma^{-1}\overline{\upsilon})$
FOR $n = 0,1,\ldots$
$\quad\lfloor\ \widehat{\zeta}_{n+1} = \widehat{\zeta}_n - \psi'(\widehat{\zeta}_n)/\psi''(\widehat{\zeta}_n).$

In the following, we provide expressions of the proximity operators of other standard divergences, which are derived from the results in Section III (see [89] for more technical details).

B. Jeffreys divergence

Let us now consider the symmetric form of (61) given by

$$\Phi(\upsilon,\xi) = \begin{cases} (\upsilon-\xi)\big(\ln\upsilon - \ln\xi\big) & \text{if } (\upsilon,\xi) \in]0,+\infty[^2 \\ 0 & \text{if } \upsilon = \xi = 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{82}$$

This function belongs to $\Gamma_0(\mathbb{R}^2)$ and satisfies (15) and (59) with

$$(\forall \zeta \in]0,+\infty[) \qquad \widetilde{\varphi}(\zeta) = -\ln \zeta. \tag{83}$$

Proposition IV.5 *The proximity operator of $\gamma\Phi$ with $\gamma \in]0,+\infty[$ is, for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$,*

$$\operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}) = \begin{cases} (\upsilon,\xi) & \text{if } W(e^{1-\gamma^{-1}\overline{\upsilon}})W(e^{1-\gamma^{-1}\overline{\xi}}) < 1 \\ (0,0) & \text{otherwise} \end{cases} \tag{84}$$

where

$$\upsilon = \overline{\upsilon} + \gamma\big(\ln\widehat{\zeta} + \widehat{\zeta} - 1\big) \tag{85}$$

$$\xi = \overline{\xi} - \gamma\big(\ln\widehat{\zeta} - \widehat{\zeta}^{-1} + 1\big) \tag{86}$$

and $\widehat{\zeta}$ is the unique minimizer on $]W(e^{1-\gamma^{-1}\overline{\upsilon}}),+\infty[$ of $\psi\colon\ ]0,+\infty[\to \mathbb{R}$:

$$\zeta \mapsto \Big(\frac{\zeta^2}{2} + \zeta - 1\Big)\ln\zeta + \frac{\zeta^3}{3} + \frac{1}{2}\Big(\gamma^{-1}\overline{\upsilon} - \frac{3}{2}\Big)\zeta^2 - \gamma^{-1}\overline{\xi}\zeta. \tag{87}$$

Remark IV.6 To minimize ψ in (87), we need to find the zero on $[\chi_-,\chi_+]$ of the function: $(\forall \zeta \in]0,+\infty[)$

$$\psi'(\zeta) = (\zeta+1)\ln\zeta + \frac{\zeta}{2} - \zeta^{-1} + \zeta^2 + \Big(\gamma^{-1}\overline{\upsilon} - \frac{3}{2}\Big)\zeta + 1 - \gamma^{-1}\overline{\xi}. \tag{88}$$

This can be performed by a projected Newton algorithm.

Remark IV.7 From a numerical standpoint, to avoid the arithmetic overflow in the exponentiations when $\gamma^{-1}\overline{\upsilon}$ or $\gamma^{-1}\overline{\xi}$ tend to $-\infty$, one can use the asymptotic approximation of the Lambert W function for large values: for every $\tau \in [1,+\infty[$,

$$\tau - \ln\tau + \frac{1}{2}\frac{\ln\tau}{\tau} \leq W\big(e^{\tau}\big) \leq \tau - \ln\tau + \frac{e}{e-1}\frac{\ln\tau}{\tau}, \tag{89}$$

with equality only if $\tau = 1$ [90].

C. Hellinger divergence

Let us now consider the function of $\Gamma_0(\mathbb{R}^2)$ given by

$$\Phi(\upsilon,\xi) = \begin{cases} (\sqrt{\upsilon} - \sqrt{\xi})^2 & \text{if } (\upsilon,\xi) \in [0,+\infty[^2 \\ +\infty & \text{otherwise.} \end{cases} \tag{90}$$

This symmetric function satisfies (15) and (59) with

$$(\forall \zeta \in]0,+\infty[) \qquad \widetilde{\varphi}(\zeta) = \zeta - \sqrt{\zeta}. \tag{91}$$

Proposition IV.8 *The proximity operator of $\gamma\Phi$ with $\gamma \in]0,+\infty[$ is, for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$,*

$$\operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}) = \begin{cases} (\upsilon,\xi) & \text{if } \overline{\upsilon} \geq \gamma \text{ or } \Big(1 - \frac{\overline{\upsilon}}{\gamma}\Big)\Big(1 - \frac{\overline{\xi}}{\gamma}\Big) < 1 \\ (0,0) & \text{otherwise,} \end{cases} \tag{92}$$

where

$$\upsilon = \overline{\upsilon} + \gamma(\rho - 1) \tag{93}$$

$$\xi = \overline{\xi} + \gamma\left(\rho^{-1} - 1\right) \tag{94}$$

and ρ is the unique solution on $\big]\max(1-\gamma^{-1}\overline{\upsilon},0),+\infty\big[$ of

$$\rho^4 + \left(\gamma^{-1}\overline{\upsilon} - 1\right)\rho^3 + \left(1 - \gamma^{-1}\overline{\xi}\right)\rho - 1 = 0. \tag{95}$$

D. Chi square divergence

Let us now consider the function of $\Gamma_0(\mathbb{R}^2)$ given by

$$\Phi(\upsilon,\xi) = \begin{cases} \dfrac{(\upsilon-\xi)^2}{\xi} & \text{if } \upsilon \in [0,+\infty[\text{ and } \xi \in]0,+\infty[\\ 0 & \text{if } \upsilon = \xi = 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{96}$$

This function satisfies (15) with

$$(\forall \zeta \in]0,+\infty[) \qquad \varphi(\zeta) = (\zeta-1)^2. \tag{97}$$

Proposition IV.9 *The proximity operator of $\gamma\Phi$ with $\gamma \in]0,+\infty[$ is, for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$,*

$$\operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}) = \begin{cases} (\upsilon,\xi) & \text{if } \overline{\upsilon} > -2\gamma \text{ and} \\ & \overline{\xi} > -\left(\overline{\upsilon} + \dfrac{\overline{\upsilon}^2}{4\gamma}\right) \\ \big(0,\max\{\overline{\xi} - \gamma, 0\}\big) & \text{otherwise,} \end{cases} \tag{98}$$

where

$$\upsilon = \overline{\upsilon} + 2\gamma(1-\rho) \tag{99}$$

$$\xi = \overline{\xi} + \gamma(\rho^2 - 1) \tag{100}$$

and ρ is the unique solution on $]0, 1 + \gamma^{-1}\overline{\upsilon}/2[$ of

$$\rho^3 + \left(1 + \gamma^{-1}\overline{\xi}\right)\rho - \gamma^{-1}\overline{\upsilon} - 2 = 0. \tag{101}$$

E. Renyi divergence

Let $\alpha \in]1,+\infty[$ and consider the below function of $\Gamma_0(\mathbb{R}^2)$

$$\Phi(\upsilon,\xi) = \begin{cases} \dfrac{\upsilon^\alpha}{\xi^{\alpha-1}} & \text{if } \upsilon \in [0,+\infty[\text{ and } \xi \in]0,+\infty[\\ 0 & \text{if } \upsilon = \xi = 0 \\ +\infty & \text{otherwise,} \end{cases} \tag{102}$$

which corresponds to the case when

$$(\forall \zeta \in]0,+\infty[) \qquad \varphi(\zeta) = \zeta^\alpha. \tag{103}$$

Note that the above function Φ allows us to generate the Rényi divergence up to a log transform and a multiplicative constant.

Proposition IV.10 *The proximity operator of $\gamma\Phi$ with $\gamma \in]0,+\infty[$ is, for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$,*

$$\operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}) = \begin{cases} (\upsilon,\xi) & \text{if } \overline{\upsilon} > 0 \text{ and} \\ & \dfrac{\gamma^{\frac{1}{\alpha-1}}\overline{\xi}}{1-\alpha} < \left(\dfrac{\overline{\upsilon}}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \\ \big(0,\max\{\overline{\xi},0\}\big) & \text{otherwise,} \end{cases} \tag{104}$$

where

$$\upsilon = \overline{\upsilon} - \gamma\alpha\widehat{\zeta}^{1-\alpha} \tag{105}$$

$$\xi = \overline{\xi} + \gamma(\alpha-1)\widehat{\zeta}^{-\alpha} \tag{106}$$

and $\widehat{\zeta}$ is the unique solution on $](\alpha\gamma\overline{\upsilon}^{-1})^{\frac{1}{\alpha-1}},+\infty[$ of

$$\gamma^{-1}\overline{\upsilon}\,\widehat{\zeta}^{1+\alpha} - \gamma^{-1}\overline{\xi}\,\widehat{\zeta}^{\alpha} - \alpha\widehat{\zeta}^2 + 1 - \alpha = 0. \tag{107}$$

Note that (107) becomes a polynomial equation when α is a rational number. In particular, when $\alpha = 2$, it reduces to the cubic equation:

$$\rho^3 + \left(2 + \gamma^{-1}\overline{\xi}\right)\rho - \gamma^{-1}\overline{\upsilon} = 0 \tag{108}$$

with $\widehat{\zeta} = \rho^{-1}$.

F. I_α divergence

Let $\alpha \in]0,1[$ and consider the function of $\Gamma_0(\mathbb{R}^2)$ given by

$$\Phi(\upsilon,\xi) = \begin{cases} \alpha\upsilon + (1-\alpha)\xi - \upsilon^\alpha \xi^{1-\alpha} & \text{if } (\upsilon,\xi) \in [0,+\infty[^2 \\ +\infty & \text{otherwise} \end{cases} \tag{109}$$

which corresponds to the case when

$$(\forall \zeta \in]0,+\infty[) \qquad \varphi(\zeta) = 1 - \alpha + \alpha\zeta - \zeta^\alpha. \tag{110}$$

Proposition IV.11 *The proximity operator of $\gamma\Phi$ with $\gamma \in]0,+\infty[$ is, for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$,*

$$\operatorname{prox}_{\gamma\Phi}(\overline{\upsilon},\overline{\xi}) = \begin{cases} (\upsilon,\xi) & \text{if } \overline{\upsilon} \geq \gamma\alpha \text{ or} \\ & 1 - \dfrac{\overline{\xi}}{\gamma(1-\alpha)} < \left(1 - \dfrac{\overline{\upsilon}}{\gamma\alpha}\right)^{\frac{\alpha}{\alpha-1}} \\ (0,0) & \text{otherwise}, \end{cases} \tag{111}$$

where

$$\upsilon = \overline{\upsilon} + \gamma\alpha(\widehat{\zeta}^{1-\alpha} - 1) \tag{112}$$

$$\xi = \overline{\xi} + \gamma(1-\alpha)(\widehat{\zeta}^{-\alpha} - 1) \tag{113}$$

and $\widehat{\zeta}$ is the unique solution on $\left]\left(\max\{1 - \frac{\overline{\upsilon}}{\gamma\alpha}, 0\}\right)^{\frac{1}{1-\alpha}}, +\infty\right[$ of

$$\alpha\widehat{\zeta}^2 + (\gamma^{-1}\overline{\upsilon} - \alpha)\widehat{\zeta}^{\alpha+1} + (1 - \alpha - \gamma^{-1}\overline{\xi})\widehat{\zeta}^\alpha = 1 - \alpha. \tag{114}$$

As for the Renyi divergence, (114) becomes a polynomial equation when α is a rational number.

Remark IV.12 We can also derive the proximity operator of

$$\widetilde{\Phi}(\upsilon,\xi) = \begin{cases} \kappa\big(\alpha\upsilon + (1-\alpha)\xi\big) - \upsilon^\alpha\xi^{1-\alpha} & \text{if } (\upsilon,\xi) \in [0,+\infty[^2 \\ +\infty & \text{otherwise,} \end{cases} \tag{115}$$

where $\kappa \in \mathbb{R}$. From Proposition II.4(iv), we get, for every $\gamma \in]0,+\infty[$ and for every $(\overline{\upsilon},\overline{\xi}) \in \mathbb{R}^2$,

$$\operatorname{prox}_{\gamma\widetilde{\Phi}}(\overline{\upsilon},\overline{\xi}) = \operatorname{prox}_{\gamma\Phi}\left(\overline{\upsilon} + \gamma(1-\kappa)\alpha, \overline{\xi} + \gamma(1-\kappa)(1-\alpha)\right), \tag{116}$$

where $\operatorname{prox}_{\gamma\Phi}$ is provided by Proposition IV.11.

V. Connection with epigraphical projections

Proximal methods iterate a sequence of steps in which proximity operators are evaluated. The efficient computation of these operators is thus essential for dealing with high-dimensional convex optimization problems. In the context of constrained optimization, at least one of the additive terms of the global cost to be minimized consists of the indicator function of a closed convex set, whose proximity operator reduces to the projections onto this set. However, if we except a few well-known cases, such projection does not admit a closed-form expression. The resolution of large-scale optimization problems involving non trivial constraints is thus quite challenging. This difficulty can be circumvented when the constraint can be expressed as the lower-level set of some separable function, by making use of epigraphical projection techniques. Such approaches have attracted interest in the last years [27], [63], [91]–[94]. The idea consists of decomposing the constraint of interest into the intersection of a half-space and a number of epigraphs of simple functions. For this approach to be successful, it is mandatory that the projection onto these epigraphs can be efficiently computed.

The next proposition shows that the expressions of the projection onto the epigraph of a wide range of functions can be deduced from the expressions of the proximity operators of φ-divergences. In particular, in Table I, for each of the φ-divergences presented in Section III, we list the associated functions φ^* for which such projections can thus be derived.

Proposition V.1 *Let $\varphi\colon \mathbb{R} \to [0,+\infty]$ be a function in $\Gamma_0(\mathbb{R})$ which is twice differentiable on $]0,+\infty[$. Let Φ be the function defined by* (15) *and $\varphi^* \in \Gamma_0(\mathbb{R})$ the Fenchel-conjugate function of the restriction of φ on $[0,+\infty[$, defined as*

$$(\forall \zeta^* \in \mathbb{R}) \qquad \varphi^*(\zeta^*) = \sup_{\zeta \in [0,+\infty[} \zeta\zeta^* - \varphi(\zeta). \tag{117}$$

Let the epigraph of φ^* be defined as

$$\operatorname{epi}\varphi^* = \left\{(\upsilon^*,\xi^*) \in \mathbb{R}^2 \;\middle|\; \varphi^*(\upsilon^*) \leq \xi^*\right\}. \tag{118}$$

Then, the projection onto $\operatorname{epi}\varphi^*$ is: for every $(\upsilon^*,\xi^*) \in \mathbb{R}^2$,

$$\mathsf{P}_{\operatorname{epi}\varphi^*}(\upsilon^*,\xi^*) = (\upsilon^*,-\xi^*) - \operatorname{prox}_\Phi(\upsilon^*,-\xi^*). \tag{119}$$

Proof. The conjugate function of Φ is, for every $(\upsilon,\xi) \in \mathbb{R}^2$,

$$\Phi^*(\upsilon^*,\xi^*) = \sup_{(\upsilon,\xi)\in\mathbb{R}^2} \upsilon\upsilon^* + \xi\xi^* - \Phi(\upsilon,\xi). \tag{120}$$

From the definition of Φ, we deduce that, for all $(\upsilon,\xi) \in \mathbb{R}^2$,

$$\begin{aligned} &\Phi^*(\upsilon^*,\xi^*) \\ &= \sup\Bigg\{ \sup_{(\upsilon,\xi)\in[0,+\infty[\times]0,+\infty[} \left(\upsilon\upsilon^* + \xi\xi^* - \xi\varphi\Big(\frac{\upsilon}{\xi}\Big)\right), \\ &\qquad \sup_{\upsilon\in]0,+\infty[} \left(\upsilon\upsilon^* - \lim_{\substack{\xi\to 0 \\ \xi>0}} \xi\varphi\Big(\frac{\upsilon}{\xi}\Big)\right), 0\Bigg\} \qquad (121) \\ &= \sup\Bigg\{ \sup_{(\upsilon,\xi)\in[0,+\infty[\times]0,+\infty[} \left(\upsilon\upsilon^* + \xi\xi^* - \xi\varphi\Big(\frac{\upsilon}{\xi}\Big)\right), 0\Bigg\} \qquad (122) \\ &= \sup\{\iota_{\operatorname{epi}\varphi^*}(\upsilon^*,-\xi^*), 0\} \qquad (123) \\ &= \iota_{\operatorname{epi}\varphi^*}(\upsilon^*,-\xi^*), \qquad (124) \end{aligned}$$

where the equality in (123) stems from [81, Example 13.8]. Then, (119) follows from the conjugation property of the proximity operator (see Proposition II.4 (v)). □

VI. Experimental results

To illustrate the potential of our results, we consider a query optimization problem in database management systems where the optimal query execution plan depends on the accurate estimation of the proportion of tuples that satisfy the predicates

TABLE I
CONJUGATE FUNCTION φ^* OF THE RESTRICTION OF φ TO $[0,+\infty[$.

Divergence	$\varphi(\zeta)$ $\zeta > 0$	$\varphi^*(\zeta^*)$ $\zeta^* \in \mathbb{R}$
Kullback-Leibler	$\zeta \ln \zeta - \zeta + 1$	$e^{\zeta^*} - 1$
Jeffreys	$(\zeta - 1) \ln \zeta$	$W(e^{1-\zeta^*}) + \big(W(e^{1-\zeta^*})\big)^{-1} + \zeta^* - 2$
Hellinger	$1 + \zeta - 2\sqrt{\zeta}$	$\begin{cases} \frac{\zeta^*}{1-\zeta^*} & \text{if } \zeta^* < 1 \\ +\infty & \text{otherwise} \end{cases}$
Chi square	$(\zeta - 1)^2$	$\begin{cases} \frac{\zeta^*(\zeta^*+4)}{4} & \text{if } \zeta^* \geq -2 \\ -1 & \text{otherwise} \end{cases}$
Renyi, $\alpha \in]1, +\infty[$	ζ^α	$\begin{cases} (\alpha-1)\Big(\frac{\zeta^*}{\alpha}\Big)^{\frac{\alpha}{\alpha-1}} & \text{if } \zeta^* \geq 0 \\ 0 & \text{otherwise} \end{cases}$
I_α, $\alpha \in]0, 1[$	$1 - \alpha + \alpha\zeta - \zeta^\alpha$	$\begin{cases} (1-\alpha)\Big(\Big(1 - \frac{\zeta^*}{\alpha}\Big)^{\frac{\alpha}{\alpha-1}} - 1\Big) & \text{if } \zeta^* \leq \alpha \\ +\infty & \text{otherwise} \end{cases}$

in the query. More specifically, every request formulated by a user can be viewed as an event in a probability space $(\Omega, \mathcal{T}, \mathcal{P})$, where Ω is a finite set of size N. In order to optimize request fulfillment, it is useful to accurately estimate the probabilities, also called *selectivities*, associated with each element of Ω. To do so, rough estimations of the probabilities of a certain number P of events can be inferred from the history of formulated requests and some a priori knowledge.

Let $x = (x^{(n)})_{1 \leq n \leq N} \in \mathbb{R}^N$ be the vector of sought probabilities, and let $z = (z^{(i)})_{1 \leq i \leq P} \in [0,1]^P$ be the vector of roughly estimated probabilities. The problem of selectivity estimation is equivalent to the following constrained entropy maximization problem [95]:

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \quad \sum_{n=1}^{N} x^{(n)} \ln x^{(n)} \quad \text{s.t.} \quad \begin{cases} Ax = z, \\ \displaystyle\sum_{n=1}^{N} x^{(n)} = 1, \\ x \in [0,1]^N, \end{cases} \tag{125}$$

where $A \in \mathbb{R}^{P \times N}$ is a binary matrix establishing the theoretical link between the probabilities of each event and the probabilities of the elements of Ω belonging to it.

Unfortunately, due to the inaccuracy of the estimated probabilities, the intersection between the affine constraints $Ax = z$ and the other ones may be empty, making the above problem infeasible. In order to overcome this issue, we propose to jointly estimate the selectivities and the feasible probabilities. Our idea consists of reformulating Problem (125) by introducing the divergence between Ax and an additional vector y of feasible probabilities. This allows us to replace the constraint $Ax = z$ with an ℓ_k-ball centered in z, yielding

$$\underset{(x,y) \in \mathbb{R}^N \times \mathbb{R}^P}{\text{minimize}} \quad D(Ax, y) + \lambda \sum_{n=1}^{N} x^{(n)} \ln x^{(n)} \quad \text{s.t.} \quad \begin{cases} \|y - z\|_k \leq \eta, \\ \displaystyle\sum_{n=1}^{N} x^{(n)} = 1, \\ x \in [0,1]^N, \end{cases} \tag{126}$$

where D is defined in (14), λ and η are some positive constants, whereas $k \in [1, +\infty[$ (the choice $k = 2$ yields the Euclidean ball).

To demonstrate the validity of this approach, we compare it with the following methods:

(i) a relaxed version of Problem (125), in which the constraint $Ax = z$ is replaced with a squared Euclidean distance:

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \quad \|Ax - z\|^2 + \lambda \sum_{n=1}^{N} x^{(n)} \ln x^{(n)} \quad \text{s.t.} \quad \begin{cases} \displaystyle\sum_{n=1}^{N} x^{(n)} = 1, \\ x \in [0,1]^N, \end{cases} \tag{127}$$

or with φ-divergence D:

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \quad D(Ax, z) + \lambda \sum_{n=1}^{N} x^{(n)} \ln x^{(n)} \quad \text{s.t.} \quad \begin{cases} \displaystyle\sum_{n=1}^{N} x^{(n)} = 1, \\ x \in [0,1]^N, \end{cases} \tag{128}$$

where λ is some positive constant;

TABLE II
COMPARISON OF Q_∞-SCORES

Problem (126) φ	(128)	(129)+(125) [94]	(127)
KL	**2.23**	2.95	2.45	25.84
Jef	2.44	3.41		
Hel	2.42	89.02		
Chi	2.34	3.20		
$I_{1/2}$	2.42	89.02		

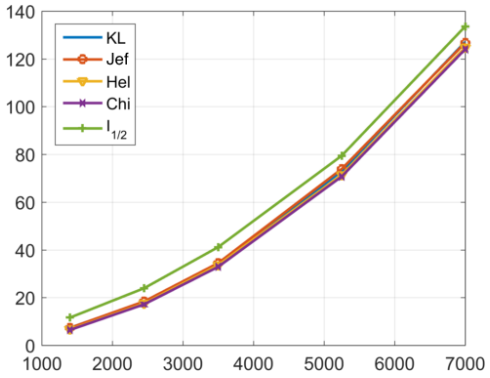

Fig. 1. Execution time (in seconds) versus size N in Problem (126).

(ii) the two-step procedure in [94], which consists of finding a solution $\widehat{x}$ to

$$\underset{x\in\mathbb{R}^N}{\text{minimize}}\ Q_1(Ax,z) \quad \text{s.t.} \quad \begin{cases} \sum_{n=1}^{N} x^{(n)} = 1, \\ x \in [0,1]^N, \end{cases} \tag{129}$$

and then solving (125) by replacing z with $\widehat{z} = A\widehat{x}$. Hereabove, for every $y \in \mathbb{R}^P$, $Q_1(y,z) = \sum_{i=1}^{P} \phi(y^{(i)}/z^{(i)})$ is a sum of quotient functions, i.e.

$$\phi(\xi) = \begin{cases} \xi, & \text{if } \xi \geq 1, \\ \xi^{-1}, & \text{if } 0 < \xi < 1, \\ +\infty, & \text{otherwise.} \end{cases} \tag{130}$$

For the numerical evaluation, we adopt an approach similar to [94], and we first consider the following low-dimensional setting:

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \qquad z = \begin{bmatrix} 0.2114 \\ 0.6331 \\ 0.6312 \\ 0.5182 \\ 0.9337 \\ 0.0035 \end{bmatrix}, \tag{131}$$

for which there exists no $x \in [0,+\infty[^N$ such that $Ax = z$. To assess the quality of the solutions x^* obtained with the different methods, we evaluate the max-quotient between Ax^* and z, that is [94]

$$Q_\infty(Ax^*, z) = \max_{1\leq i\leq P} \phi\left(\frac{[Ax^*]^{(i)}}{z^{(i)}}\right). \tag{132}$$

Table II collects the Q_∞-scores (lower is better) obtained with the different approaches. For all the considered φ-divergences,[2] the proposed approach performs favorably with respect to the state-of-the-art, the KL divergence providing the best performance among the panel of considered φ-divergences. For the sake of fairness, the hyperparameters λ and η were hand-tuned in order to get the best possible score for each compared method. The good performance of our approach is related to the fact that φ-divergences are well suited for the estimation of probability distributions.

Figure 1 next shows the computational time for solving Problem (126) for various dimensions N of the selectivity vector to be estimated, with A and z randomly generated so as to keep the ratio N/P equal to $7/6$. To make this comparison, the primal-dual proximal method proposed in [69] was implemented in MATLAB R2015, by using the stopping criterion $\|x_{n+1} - x_n\| < 10^{-7}\|x_n\|$. We then measured the execution times on an Intel i5 CPU at 3.20 GHz with 12 GB of RAM. The results show that all the considered φ-divergences can be efficiently optimized, with no significant computational time differences between them.

VII. CONCLUSION

In this paper, we have shown how to solve convex optimization problems involving discrete information divergences by using proximal methods. We have carried out a thorough study of the properties of the proximity operators of φ-divergences, which has led us to derive new tractable expressions of them. In addition, we have related these expressions to the projection onto the epigraph of a number of convex functions.

Finally, we have illustrated our results on a selectivity estimation problem. In this context, φ-divergences appear to be well suited for the estimation of the sought probability distributions. Moreover, computational time evaluations allowed us to show that the proposed numerical methods provide efficient solutions for solving such large-scale optimization problems.

APPENDIX A
CONVERGENCE PROOF OF ALGORITHM 1

We aim at finding the unique zero on $]\exp(-\gamma^{-1}\overline{v}),+\infty[$ of the function ψ' given by (81) along with its derivatives:

$$(\forall \zeta \in]0,+\infty[) \quad \psi''(\zeta) = 1 + \ln\zeta + \gamma^{-1}\overline{v} + \zeta^{-2}, \tag{133}$$

$$\psi'''(\zeta) = \zeta^{-1} - 2\,\zeta^{-3}. \tag{134}$$

To do so, we employ the Newton method given in Algorithm 1, the convergence of which is here established. Assume that

- $(\overline{v},\overline{\xi}) \in \mathbb{R}^2$ are such that $\exp(\gamma^{-1}\overline{v}) > 1 - \gamma^{-1}\overline{\xi}$,
- $\widehat{\zeta}$ is the zero on $]\exp(-\gamma^{-1}\overline{v}),+\infty[$ of ψ',
- $(\widehat{\zeta}_n)_{n\in\mathbb{N}}$ is the sequence generated by Algorithm 1,
- $\epsilon_n = \widehat{\zeta}_n - \widehat{\zeta}$ for every $n \in \mathbb{N}$.

We first recall a fundamental property of the Newton method, and then we proceed to the actual convergence proof.

[2]Note that the Renyi divergence is not suitable for the considered application, because it tends to favor sparse solutions.

Lemma A.1 *For every $n \in \mathbb{N}$,*

$$\epsilon_{n+1} = \epsilon_n^2 \frac{\psi'''(\varrho_n)}{2\psi''(\widehat{\zeta}_n)} \tag{135}$$

where ϱ_n is between $\widehat{\zeta}_n$ and $\widehat{\zeta}$.

Proof. The definition of ϵ_{n+1} yields

$$\epsilon_{n+1} = \widehat{\zeta}_n - \frac{\psi'(\widehat{\zeta}_n)}{\psi''(\widehat{\zeta}_n)} - \widehat{\zeta} = \frac{\epsilon_n \psi''(\widehat{\zeta}_n) - \psi'(\widehat{\zeta}_n)}{\psi''(\widehat{\zeta}_n)}. \tag{136}$$

Moreover, for every $\widehat{\zeta}_n \in]0, +\infty[$, the second-order Taylor expansion of ψ' around $\widehat{\zeta}_n$ is

$$\psi'(\widehat{\zeta}) = \psi'(\widehat{\zeta}_n) + \psi''(\widehat{\zeta}_n)(\widehat{\zeta} - \widehat{\zeta}_n) + \frac{1}{2}\psi'''(\varrho_n)(\widehat{\zeta} - \widehat{\zeta}_n)^2, \tag{137}$$

where ϱ_n is between $\widehat{\zeta}_n$ and $\widehat{\zeta}$. From the above equality, we deduce that $\psi'(\widehat{\zeta}) = \psi'(\widehat{\zeta}_n) - \psi''(\widehat{\zeta}_n)\epsilon_n + \frac{1}{2}\psi'''(\varrho_n)\epsilon_n^2 = 0$. □

Proposition A.2 *The sequence $(\widehat{\zeta}_n)_{n\in\mathbb{N}}$ converges to $\widehat{\zeta}$.*

Proof. The assumption $\exp(\gamma^{-1}\overline{v}) > 1 - \gamma^{-1}\overline{\xi}$ implies that ψ' is negative at the initial value $\widehat{\zeta}_0 = \exp(-\gamma^{-1}\overline{v})$, that is

$$\psi'(\widehat{\zeta}_0) = -\exp(\gamma^{-1}\overline{v}) + 1 - \gamma^{-1}\overline{\xi} < 0. \tag{138}$$

Moreover, ψ' is increasing on $\left[\exp(-\gamma^{-1}\overline{v}), +\infty\right[$, since

$$\left(\forall \zeta \in \left[\exp(-\gamma^{-1}\overline{v}), +\infty\right[\right) \qquad \psi''(\zeta) > 0, \tag{139}$$

and $\sqrt{2}$ is a non-critical inflection point for ψ', since

$$\left(\forall \zeta \in \left]\sqrt{2}, +\infty\right[\right) \qquad \psi'''(\zeta) > 0, \tag{140}$$

$$\left(\forall \zeta \in \left]0, \sqrt{2}\right[\right) \qquad \psi'''(\zeta) < 0. \tag{141}$$

To prove the convergence, we consider the following cases:

- *Case* $\widehat{\zeta} \leq \sqrt{2}$: ψ' is increasing and concave on $[\widehat{\zeta}_0, \sqrt{2}]$. Hence, Newton method initialized at the lower bound of interval $[\widehat{\zeta}_0, \widehat{\zeta}]$ monotonically increases to $\widehat{\zeta}$ [96].
- *Case* $\sqrt{2} \leq \widehat{\zeta}_0 < \widehat{\zeta}$: ψ' is increasing and convex on $[\widehat{\zeta}_0, +\infty[$. Hence, Lemma A.1 yields $\epsilon_1 = \widehat{\zeta}_1 - \widehat{\zeta} > 0$. It then follows from standard properties of Newton algorithm for minimizing an increasing convex function that $(\widehat{\zeta}_n)_{n\geq 1}$ monotonically decreases to $\widehat{\zeta}$ [96].
- *Case* $\widehat{\zeta}_0 < \sqrt{2} < \widehat{\zeta}$: as ψ' is negative and increasing on $[\widehat{\zeta}_0, \widehat{\zeta}[$, the quantity $-\psi'/\psi''$ is positive and lower bounded on $[\widehat{\zeta}_0, \sqrt{2}]$, that is

$$\left(\forall \zeta \in [\widehat{\zeta}_0, \sqrt{2}]\right) \quad -\frac{\psi'(\zeta)}{\psi''(\zeta)} \geq -\frac{\psi'(\sqrt{2})}{\psi''(\widehat{\zeta}_0)} > 0. \tag{142}$$

  There thus exists $k > 0$ such that $\widehat{\zeta}_0 < \ldots < \widehat{\zeta}_k$ and $\widehat{\zeta}_k > \sqrt{2}$. Then, the convergence of $(\widehat{\zeta}_n)_{n\geq k}$ follows from the same arguments as in the previous case.

□

REFERENCES

[1] M. El Gheche, A. Jezierska, J.-C. Pesquet, and J. Farah, "A proximal approach for signal recovery based on information measures," in *Proc. Eur. Signal Process. Conf.*, Marrakech, Maroc, Sept. 2013, pp. 1–5.

[2] M. El Gheche, J.-C. Pesquet, and J. Farah, "A proximal approach for optimization problems involving kullback divergences," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Vancouver, Canada, May 2013, pp. 5984–5988.

[3] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonable be supposed to have arisen from random sampling," *Phil. Mag.*, vol. 50, pp. 157–175, 1900.

[4] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.," *J. für die reine und angewandte Mathematik*, vol. 136, pp. 210–271, 1909.

[5] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 623–656, Jul. 1948.

[6] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[7] I. Csiszár, "Eine informations theoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten.," *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, vol. 8, pp. 85–108, 1963.

[8] A. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 28, no. 1, pp. 131–142, 1966.

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

[10] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[11] R.E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.

[12] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14–20, Jan. 1972.

[13] H. H. Bauschke, P. L. Combettes, and D. Noll, "Joint minimization with alternating Bregman proximity operators," *Pac. J. Optim*, vol. 2, no. 3, pp. 401–424, Sep. 2006.

[14] P. L. Combettes and Q. V. Nguyen, "Solving composite monotone inclusions in reflexive Banach spaces by constructing best Bregman approximations from their Kuhn-Tucker set," *J. Convex Anal.*, vol. 23, no. 2, May 2016.

[15] M. Chiang and S. Boyd, "Geometric programming duals of channel capacity and rate distortion," *IEEE Trans. Inf. Theory*, vol. 50, no. 2, pp. 245 – 258, Feb. 2004.

[16] S. Chrétien and A. O. Hero, "On EM algorithms and their proximal generalizations," *Control Optim. Calc. Var.*, vol. 12, pp. 308–326, Jan. 2008.

[17] C. L. Byrne, "Iterative image reconstruction algorithms based on cross-entropy minimization," *IEEE Trans. Image Process.*, vol. 2, no. 1, pp. 96 –103, Jan. 1993.

[18] W. Richardson, "Bayesian-based iterative method of image restoration," *J. Opt. Soc. Am. A*, vol. 62, no. 1, pp. 55–59, Jan. 1972.

[19] L. B. Lucy, "An iterative technique for the rectification of observed distributions," *Astron. J.*, vol. 79, no. 6, pp. 745–754, 1974.

[20] J.A. Fessler, "Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans," *IEEE Trans. Image Process.*, vol. 4, no. 10, pp. 1439 –1450, Oct. 1995.

[21] F.-X. Dupé, M. J. Fadili, and J.-L. Starck, "A proximal iteration for deconvolving Poisson noisy images using sparse representations," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 310–321, Feb. 2009.

[22] R. Zanella, P. Boccacci, L. Zanni, and M. Bertero, "Efficient gradient projection methods for edge-preserving removal of poisson noise," *Inverse Probl.*, vol. 25, no. 4, 2009.

[23] P. Piro, S. Anthoine, E. Debreuve, and M. Barlaud, "Combining spatial and temporal patches for scalable video indexing," *Multimed. Tools Appl.*, vol. 48, no. 1, pp. 89–104, May 2010.

[24] N. Pustelnik, C. Chaux, and J.-C. Pesquet, "Parallel proximal algorithm for image restoration using hybrid regularization," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2450 –2462, Nov. 2011.

[25] T. Teuber, G. Steidl, and R. H. Chan, "Minimization and parameter estimation for seminorm regularization models with I-divergence constraints," *Inverse Probl.*, vol. 29, pp. 1–28, 2013.

[26] M. Carlavan and L. Blanc-Féraud, "Sparse poisson noisy image deblurring," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1834–1846, Apr. 2012.

[27] S. Harizanov, J.-C. Pesquet, and G. Steidl, "Epigraphical projection for solving least squares anscombe transformed constrained optimization problems," in *Scale-Space and Variational Methods in Computer Vision*, A. Kuijper et al., Ed., vol. 7893 of *Lect. Notes Comput. Sc.*, pp. 125–136. 2013.

[28] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. R. Soc. Lond. A Math. Phys. Sci.*, vol. 186, no. 1007, pp. 453–461, 1946.

[29] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2882–2904, 2009.

[30] F. Nielsen, "Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 657–660, July 2013.

[31] I. Sason, "Tight bounds for symmetric divergence measures and a refined bound for lossless source coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 701–707, 2015.

[32] G. L. Gilardoni, "On the minimum f-divergence for given total variation," *Comptes Rendus Mathematique*, vol. 343, no. 11-12, pp. 763–766, 2006.

[33] R. Beran, "Minimum hellinger distance estimates for parametric models," *Ann. Stat.*, vol. 5, no. 3, pp. 445–463, 1977.

[34] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall International, 1982.

[35] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Stat. Rev.*, vol. 70, no. 3, pp. 419–435, 2002.

[36] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.

[37] T.W. Rauber, T. Braun, and K. Berns, "Probabilistic distance measures of the dirichlet and beta distributions.," *Pattern Recogn.*, vol. 41, no. 2, pp. 637–645, 2008.

[38] L. LeCam, "Convergence of estimates under dimensionality restrictions," *Ann. Stat.*, vol. 1, no. 1, pp. 38–53, 1973.

[39] S. van de Geer, "Hellinger-consistency of certain nonparametric maximum likelihood estimators," *Ann. Stat.*, vol. 21, no. 1, pp. 14–44, 1993.

[40] L. Chang-Hwan, "A new measure of rule importance using hellinger divergence," in *Int. Conf. on Data Analytics*, Barcelona, Spain, Sep. 2012, pp. 103–106.

[41] D. Cieslak, T. Hoens, N. Chawla, and W. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," *Data Min. Knowl. Discov.*, vol. 24, no. 1, pp. 136–158, 2012.

[42] I. Park, S. Seth, M. Rao, and J.C. Principe, "Estimation of symmetric chi-square divergence for point processes," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Prague, Czech Republic, May 2011, pp. 2016–2019.

[43] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. on Math. Statist. and Prob.*, California, Berkeley, Jun. 1961, vol. 1, pp. 547–561.

[44] P. Harremoës, "Interpretations of rényi entropies and divergences," *Physica A: Statistical Mechanics and its Applications*, vol. 365, no. 1, pp. 57–62, 2006.

[45] I. Vajda, *Theory of Statistical Inference and Information*, Kluwer, Dordrecht, 1989.

[46] F. Liese and I. Vajda, *Convex Statistical Distances*, Treubner, 1987.

[47] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Process. Mag.*, vol. 19, no. 5, pp. 85–95, 2002.

[48] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations," *Ann. Stat.*, vol. 23, pp. 493–507, 1952.

[49] I. Csiszár, "Information measures: A critical survey.," *IEEE Trans. Inf. Theory*, vol. A, pp. 73–86, 1974.

[50] S.-I. Amari, "Alpha-divergence is unique, belonging to both f-divergence and Bregman divergence classes," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4925–4931, 2009.

[51] J. Zhang, "Divergence function, duality, and convex analysis," *Neural Comput.*, vol. 16, pp. 159–195, 2004.

[52] T. Minka, "Divergence measures and message passing," Tech. Rep., Microsoft Research Technical Report (MSR-TR-2005), 2005.

[53] A. Cichocki, L. Lee, Y. Kim, and S. Choi, "Non-negative matrix factorization with α-divergence," *Pattern Recogn. Lett.*, vol. 29, no. 9, pp. 1433–1440, 2008.

[54] I. Csiszár and F. Matúš , "On minimization of multivariate entropy functionals," in *IEEE Information Theory Workshop*, Jun. 2009, pp. 96–100.

[55] X. L. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *Ann. Stat.*, vol. 37, no. 2, pp. 876–904, 2009.

[56] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, "A variational formulation for frame-based inverse problems," *Inverse Probl.*, vol. 23, no. 4, pp. 1495–1518, Jun. 2007.

[57] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds., pp. 185–212. Springer-Verlag, New York, 2011.

[58] P. L. Combettes and J.-C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inverse Probl.*, vol. 24, no. 6, pp. 065014, Dec. 2008.

[59] L. Condat, "Fast projection onto the simplex and the l1 ball," *Math. Program. Series A*, 2015, to appear.

[60] J.-D. Benamou and Y. Brenier, "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem," *Numer. Math.*, vol. 84, no. 3, pp. 375–393, 2000.

[61] J.-D. Benamou, Y. Brenier, and K. Guittet, "The Monge-Kantorovich mass transfer and its computational fluid mechanics formulation," *Int. J. Numer. Meth. Fluids*, vol. 40, pp. 21–30, 2002.

[62] P. L Combettes and C. L. Müller, "Perspective functions: Proximal calculus and applications in high-dimensional statistics," *J. Math. Anal. Appl.*, 2016.

[63] G. Chierchia, N. Pustelnik, J.-C. Pesquet, and B. Pesquet-Popescu, "Epigraphical projection and proximal tools for solving constrained convex optimization problems," *Signal Image Video P.*, vol. 9, no. 8, pp. 1737–1749, Nov. 2015.

[64] G. Chen and M. Teboulle, "A proximal-based decomposition method for convex minimization problems," *Math. Program.*, vol. 64, no. 1–3, pp. 81–101, Mar. 1994.

[65] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, Mar. 2011.

[66] S. Setzer, G. Steidl, and T. Teuber, "Deblurring Poissonian images by split Bregman techniques," *J. Visual Communication and Image Representation*, vol. 21, no. 3, pp. 193–199, Apr. 2010.

[67] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, May 2011.

[68] L. M. Briceño-Arias and P. L. Combettes, "A monotone + skew splitting model for composite monotone inclusions in duality," *SIAM J. Opt.*, vol. 21, no. 4, pp. 1230–1250, Oct. 2011.

[69] P. L. Combettes and J.-C. Pesquet, "Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators," *Set-Valued Var. Anal.*, vol. 20, no. 2, pp. 307–330, Jun. 2012.

[70] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Adv. Comput. Math.*, vol. 38, no. 3, pp. 667–681, Apr. 2013.

[71] L. Condat, "A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Optimiz. Theory App.*, vol. 158, no. 2, pp. 460–479, Aug. 2013.

[72] J.-C. Pesquet and N. Pustelnik, "A parallel inertial proximal optimization method," *Pac. J. Optim.*, vol. 8, no. 2, pp. 273–305, Apr. 2012.

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

[74] R. Gaetano, G. Chierchia, and B. Pesquet-Popescu, "Parallel implementations of a disparity estimation algorithm based on a proximal splitting method," in *Proc. Int. Conf. Visual Commun. Image Process.*, San Diego, USA, 2012.

[75] J. J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.

[76] T. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inf. Theory*, vol. 30, no. 2, pp. 369–373, 1984.

[77] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. B*, pp. 1–38, Dec. 1977.

[78] A. Subramanya and J. Bilmes, "Soft-supervised learning for text classification," in *Proc. of EMNLP*, 2008, pp. 1090–1099.

[79] G. Tartavel, G. Peyré, and Y. Gousseau, "Wasserstein loss for image synthesis and restoration," *SIAM J. Imaging Sci.*, vol. 9, no. 4, pp. 1726–1755, 2016.

[80] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2nd edition, 2006.

[81] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, Jan. 2011.

[82] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms, Part I : Fundamentals*, vol. 305 of *Grundlehren der mathematischen Wissenschaften*, Springer-Verlag, Berlin, Heidelberg, N.Y., 2nd edition, 1996.

[83] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[84] M. Basseville, "Distance measures for signal processing and pattern recognition," *European J. Signal Process.*, vol. 18, no. 4, pp. 349–369, 1989.

[85] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2014.

[86] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *C. R. Acad. Sci.*, vol. 255, pp. 2897–2899, 1962.

[87] P. L. Combettes and J.-C. Pesquet, "Proximal thresholding algorithm for minimization over orthonormal bases," *SIAM J. Optim.*, vol. 18, no. 4, pp. 1351–1376, Nov. 2007.

[88] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, 1996.

[89] M. El Gheche, G. Chierchia, and J. C. Pesquet, "Proximity operators of discrete information divergences – extended version," *Preprint arXiv:1606.09552*, 2017.

[90] A. Hoorfar and M. Hassani, "Inequalities on the Lambert W function and hyperpower function," *J. Inequal. Pure Appl. Math*, vol. 9, no. 2, pp. 5–9, 2008.

[91] M. Tofighi, K. Kose, and A. E. Cetin, "Signal reconstruction framework based on Projections onto Epigraph Set of a Convex cost function (PESC)," *Preprint arXiv:1402.2088*, 2014.

[92] S. Ono and I. Yamada, "Second-order Total Generalized Variation constraint," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Florence, Italy, May 2014, pp. 4938–4942.

[93] P.-W. Wang, M. Wytock, and J. Z. Kolter, "Epigraph projections for fast general convex programming," in *Proc. of ICML*, 2016.

[94] G. Moerkotte, M. Montag, A. Repetti, and G. Steidl, "Proximal operator of quotient functions with application to a feasibility problem in query optimization," *J. Comput. Appl. Math.*, vol. 285, pp. 243–255, Sept. 2015.

[95] V. Markl, P. Haas, M. Kutsch, N. Megiddo, U. Srivastava, and T. Tran, "Consistent selectivity estimation via maximum entropy," *VLDB J.*, vol. 16, no. 1, pp. 55–76, 2007.

[96] D. R. Kincaid and E. W. Cheney, *Numerical analysis: mathematics of scientific computing*, vol. 2, American Mathematical Soc., 2002.