A proximal approach for optimization problems involving Kullback divergences

Mireille El Gheche, Jean-Christophe Pesquet, Farah Joumana

HAL Id: hal-00842705

https://hal.archives-ouvertes.fr/hal-00842705

Submitted on 9 Jul 2013

A PROXIMAL APPROACH FOR OPTIMIZATION PROBLEMS INVOLVING KULLBACK DIVERGENCES

Mireille El Gheche[1], Jean-Christophe Pesquet[1] and Joumana Farah[2],

[1] Université Paris-Est
LIGM and UMR-CNRS 8049,
77454 Marne-la-Vallée cedex, France
{mireille.elgheche, pesquet}@univ-mlv.fr

[2] Holy-Spirit University of Kaslik,
Faculty of Engineering, Telecom. dep.,
P.O. Box 446, Jounieh, Lebanon
joumanafarah@usek.edu.lb

ABSTRACT

Convex optimization problems involving information measures have been extensively investigated in source and channel coding. These measures can also be successfully used in inverse problems encountered in signal and image processing. The related optimization problems are often challenging due to their large size. In this paper, we derive closed-form expressions of the proximity operators of Kullback-Leibler and Jeffreys-Kullback divergences. Building upon these results, we develop an efficient primal-dual proximal approach. This allows us to address a wide range of convex optimization problems whose objective function expression includes one of these divergences. An image registration application serves as an example for illustrating the good performance of the proposed method.

Index Terms— Divergences, inverse problems, convex optimization, proximity operator, parallel algorithms.

1. INTRODUCTION

Kullback-Leibler (KL) and Jeffreys-Kullback (JK) divergences are often used as discrete measures in signal processing problems. They serve as dissimilarity functions in many information theoretic models (e.g. in source and channel coding), data recovery tasks (e.g. image restoration and reconstruction), machine learning (pattern recognition and clustering),.... Their popularity stems from their intimate connections with information theory concepts such as the entropy and the mutual information.

The KL divergence is known to play a prominent role in the computation of channel capacity and rate-distortion functions. One can address these problems with the celebrated alternating minimization algorithm proposed by Blahut and Arimoto [1, 2]. However, other approaches based on geometric programming [3] may provide more efficient numerical solutions.

The generalized KL divergence is also known as the I-divergence. It plays an important role in inverse problems for recovering the signal of interest in the presence of Poisson noise. In such a case, the I-divergence is usually employed as a data fidelity term. For instance, an alternating projection technique was proposed in [4], where both the data fidelity term and the regularization term are based on the KL divergence. The problem was formulated in a similar manner in [5], whereas in [6, 7, 8, 9] more general forms of the regularization functions are considered. In particular, the developments in [7, 8, 9] are grounded on proximal splitting methods. These proximal tools were recently shown to offer both efficient and flexible solutions to a wide class of possibly nonsmooth convex minimization problems (see [10] and references therein). Note that, in all of the aforementioned works, one of the two variables of the KL divergence is fixed. As the KL divergence is a Bregman distance, optimization problems involving this function can also be addressed by using the alternating minimization approach proposed in [11]. However, the required optimization steps may be difficult to implement and the convergence of the algorithm is only guaranteed under restrictive conditions.

Moreover, a Kullback-proximal algorithm generalizing the EM algorithm was investigated in [12]. The KL divergence is then used as a metric for the maximization of a log-likelihood function, rather than being one of the terms of the objective function.

In this paper, we consider the following generic form of the optimization problem:

Problem 1.1

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

where D is the KL or JK divergence defined over $\mathbb{R}^P \times \mathbb{R}^P$, u and v are two vectors in $\mathbb{R}^P$. In addition, for every $s \in \{1, \ldots, S\}$, $R_s \colon \mathbb{R}^{K_s} \to]-\infty, +\infty]$ is a convex function and T_s a matrix in $\mathbb{R}^{K_s \times N}$.

In applications to inverse problems, R_s with $s \in \{1, \ldots, S\}$ may be some regularization function serving to enforce the smoothness of the sought solution or to model some additional prior information, e.g. the sparsity of $(T_s x)$.

Note that a special case of interest in information theory is when $u = v = 0$, $N = 2P$, $A = [\mathbf{I}_P \ \mathbf{0}]$, and $B = [\mathbf{0} \ \mathbf{I}_P]$,

where $\mathbf{I}_P$ denotes the $P \times P$ identity matrix. By decomposing the vector x as $x = [p^\top q^\top]^\top$, where $(p, q) \in (\mathbb{R}^P)^2$, and by setting $(\forall s \in \{1, \dots, S\})$ $T_s = [U_s \ \ V_s]$, where U_s and V_s are matrices in $\mathbb{R}^{K_s \times P}$, (1) simplifies to

$$\underset{p \in \mathbb{R}^P, \, q \in \mathbb{R}^P}{\text{minimize}} \ D(p, q) + \sum_{s=1}^{S} R_s(U_s p + V_s q). \quad (2)$$

The fact that the variables p and q usually correspond to probability masses can then be imposed by choosing $S \geq 2$, $U_1 = V_2 = \mathbf{I}_P$ and $V_1 = U_2 = \mathbf{0}$, and by setting R_1 and R_2 to the indicator function of the unit simplex:

$$C = \Big\{ (\pi^{(1)}, \dots, \pi^{(P)})^\top \in [0, +\infty[^P \ | \sum_{i=1}^{P} \pi^{(i)} = 1 \Big\}. \quad (3)$$

The outline of the paper is as follows. In Section 2, we derive expressions of the proximity operators of KL and JK divergences. These proximity operators constitute the building blocks of the primal-dual algorithm which is proposed in Section 3. In Section 4, an image registration problem is formulated under the form of Problem 1.1. Simulation results show the validity of the proposed method for the joint estimation of depth information and illumination variation.

2. PROXIMITY OPERATORS OF KULLBACK DIVERGENCES

2.1. Convex analysis background

Definitions Let $\mathcal{H}$ be a real Hilbert space endowed with the norm $\| \cdot \|$. In the following, $\Gamma_0(\mathcal{H})$ denotes the class of convex functions defined on $\mathcal{H}$, which are lower-semicontinuous, proper (i.e. with a nonempty domain), and which take their values in $]-\infty, +\infty]$. Let $f \in \Gamma_0(\mathcal{H})$. For every $\overline{x} \in \mathcal{H}$, there exists a unique minimizer of the function $f + \frac{1}{2}\| \cdot - \overline{x} \|^2$. This minimizer is called the proximity operator of f at $\overline{x}$ and is denoted by $\text{prox}_f \overline{x}$. The proximity operator has played a key role in recent developments in convex optimization, since it provides a natural extension of the notion of projection [10]. Indeed, if C is a nonempty closed convex subset of $\mathcal{H}$, and ι_C is the indicator function of C (equal to 0 on C and $+\infty$ otherwise), then prox_{ι_C} reduces to the projection P_C onto C.

Separability In this paper, we are mainly interested in Kullback divergences. One of their properties is that they are separable, i.e. $\big(\forall p = (p^{(i)})_{1 \leq i \leq P} \in \mathbb{R}^P\big)\big(\forall q = (q^{(i)})_{1 \leq i \leq P} \in \mathbb{R}^P\big)$

$$D(p, q) = \sum_{i=1}^{P} d(p^{(i)}, q^{(i)}) \quad (4)$$

where d is a scalar divergence measure belonging to $\Gamma_0(\mathbb{R} \times \mathbb{R})$. Hence, the proximity operator of the divergence D, calculated at points $\bar{p} = (\bar{p}^{(i)})_{1 \leq i \leq P} \in \mathbb{R}^P$ and $\bar{q} = (\bar{q}^{(i)})_{1 \leq i \leq P} \in \mathbb{R}^P$, reads

$$\text{prox}_D(\bar{p}, \bar{q}) = \Big(\text{prox}_d(\bar{p}^{(i)}, \bar{q}^{(i)}) \Big)_{1 \leq i \leq P}. \quad (5)$$

We shall now look more closely at the form of prox_d for KL and JK divergences.

2.2. Proximity operators

Kullback-Leibler divergence. In this case, function d is defined as

$$d \colon (v, \xi) \mapsto \begin{cases} v \ln\left(\frac{v}{\xi}\right) + \xi - v & \text{if } (v, \xi) \in \,]0, +\infty[^2 \\ \xi & \text{if } v = 0 \text{ and } \xi \in [0, +\infty[\\ +\infty & \text{otherwise.} \end{cases} \quad (6)$$

Proposition 2.1 *The proximity operator of γd with $\gamma > 0$ is given by:* $(\forall(\overline{v}, \overline{\xi}) \in \mathbb{R}^2)$, $\text{prox}_{\gamma d}(\overline{v}, \overline{\xi}) =$

$$\begin{cases} \big(\overline{v} + \gamma \ln \widehat{\zeta}, \overline{\xi} + \gamma(\widehat{\zeta}^{-1} - 1)\big) & \text{if } \exp(\overline{v}/\gamma) > 1 - \gamma^{-1}\overline{\xi} \\ (0, 0) & \text{otherwise} \end{cases} \quad (7)$$

where, in the first case, $\widehat{\zeta}$ is the unique minimizer on $]\exp(-\overline{v}/\gamma), +\infty[$ of the function ψ given by:

$$(\forall \zeta \in \,]0, +\infty[)$$
$$\psi(\zeta) = \Big(\frac{\zeta^2}{2} - 1\Big) \ln \zeta + \frac{1}{2}\Big(\gamma^{-1}\overline{v} - \frac{1}{2}\Big)\zeta^2 + (1 - \gamma^{-1}\overline{\xi})\zeta. \quad (8)$$

Jeffreys-Kullback divergence The corresponding function d is the symmetrized form of the one in (6):

$$d \colon (v, \xi) \mapsto \begin{cases} (v - \xi)\big(\ln v - \ln \xi\big) & \text{if } (v, \xi) \in \,]0, +\infty[^2 \\ 0 & \text{if } v = \xi = 0 \\ +\infty & \text{otherwise.} \end{cases} \quad (9)$$

Proposition 2.2 *The proximity operator of γd with $\gamma > 0$ is:* $(\forall(\overline{v}, \overline{\xi}) \in \mathbb{R}^2)$, $\text{prox}_{\gamma d}(\overline{v}, \overline{\xi}) =$

$$\begin{cases} \big(\overline{v} + \gamma\big(\ln \widehat{\zeta} + \widehat{\zeta} - 1\big), \overline{\xi} - \gamma\big(\ln \widehat{\zeta} - \widehat{\zeta}^{-1} + 1\big)\big) \\ \qquad \text{if } W(e^{1-\gamma^{-1}\overline{v}})W(e^{1-\gamma^{-1}\overline{\xi}}) < 1 \quad (10) \\ (0, 0) \qquad \text{otherwise} \end{cases}$$

where, in the first case, $\widehat{\zeta}$ is the unique minimizer on $]W(e^{1-\gamma^{-1}\overline{v}}), +\infty[$ of the function ψ defined as

$$(\forall \zeta \in \,]0, +\infty[) \qquad \psi(\zeta) =$$
$$\Big(\frac{\zeta^2}{2} + \zeta - 1\Big) \ln \zeta + \frac{\zeta^3}{3} + \frac{1}{2}\Big(\gamma^{-1}\overline{v} - \frac{3}{2}\Big)\zeta^2 - \gamma^{-1}\overline{\xi}\zeta. \quad (11)$$

and W is the Lambert W function [13].

Due to some strictly convex properties of the above functions ψ, the existence of $\widehat{\zeta}$ is guaranteed, and it can be computed by standard one-dimensional search techniques which are implementable in parallel. Due to the limited space, the reader is referred to [14] for further details.

Translation property Proximity operators have many interesting properties [10]. In particular, let D be defined by (4) and let $\gamma > 0$. Let $u = (u^{(i)})_{1 \le i \le P} \in \mathbb{R}^P$, and $v = (v^{(i)})_{1 \le i \le P} \in \mathbb{R}^P$. Then, for every $\overline{p} = (\overline{p}^{(i)})_{1 \le i \le P} \in \mathbb{R}^P$ and $\overline{q} = (\overline{q}^{(i)})_{1 \le i \le P} \in \mathbb{R}^P$,

$$\mathrm{prox}_{\gamma D(\cdot + u, \cdot + v)}(\overline{p}, \overline{q}) = (p^{(i)} - u^{(i)}, q^{(i)} - v^{(i)})_{(1 \le i \le P)}$$

where

$$(\forall i \in \{1, \dots, P\})\ (p^{(i)}, q^{(i)}) = \mathrm{prox}_{\gamma d}(\overline{p}^{(i)} + u^{(i)}, \overline{q}^{(i)} + v^{(i)}).$$

3. PRIMAL-DUAL ALGORITHM

Now that we know how to compute the proximity operators of KL and JK divergences, various parallel proximal splitting methods ([15, 16, 17, 18, 19, 20]) can be employed to solve Problem 1.1 when, for every $s \in \{1, \dots, S\}$, R_s belongs to $\Gamma_0(\mathbb{R}^{K_s})$. We propose here to make use of a primal-dual proximal algorithm whose main advantage, for large-size problems, is the absence of any matrix inversion. More precisely, the considered M+LFBF (Monotone+Lipschitz Forward Backward Forward) method [19] takes the following form:

Initialization
$$\left\lfloor \begin{array}{l} t_{0,0} \in \mathbb{R}^P, t_{1,0} \in \mathbb{R}^P, t_{2,0} \in \mathbb{R}^{K_1}, \dots, t_{S+1,0} \in \mathbb{R}^{K_S}, \\ x_0 \in \mathbb{R}^{2P} \\ \beta = \left(\|A\|^2 + \|B\|^2 + \sum_{s=1}^{S} \|T_s\|^2 \right)^{1/2}, \ \varepsilon \in\ ]0, 1/(\beta+1)[\end{array} \right.$$
For $n = 0, 1, \dots$
$$\left\lfloor \begin{array}{l} \gamma_n \in [\varepsilon, (1-\varepsilon)/\beta] \\ \widehat{x}_n = x_n - \gamma_n(A^\top t_{0,n} + B^\top t_{1,n} + \sum_{s=1}^{S} T_s^\top t_{s+1,n}) \\ \widehat{t}_{0,n} = t_{0,n} + \gamma_n A x_n, \ \widehat{t}_{1,n} = t_{1,n} + \gamma_n B x_n \\ (r_{0,n}, r_{1,n}) = (\widehat{t}_{0,n}, \widehat{t}_{1,n}) \\ \qquad\qquad - \gamma_n \mathrm{prox}_{\gamma_n^{-1} D(\cdot + u, \cdot + v)}(\gamma_n^{-1} \widehat{t}_{0,n}, \gamma_n^{-1} \widehat{t}_{1,n}) + e_{0,n} \\ \widetilde{t}_{0,n} = r_{0,n} + \gamma_n A \widehat{x}_n, \ \widetilde{t}_{1,n} = r_{1,n} + \gamma_n B \widehat{x}_n \\ t_{0,n+1} = t_{0,n} - \widehat{t}_{0,n} + \widetilde{t}_{0,n}, \ t_{1,n+1} = t_{1,n} - \widehat{t}_{1,n} + \widetilde{t}_{1,n} \\ \text{For } s = 1, \dots, S \\ \quad \left\lfloor \begin{array}{l} \widehat{t}_{s+1,n} = t_{s+1,n} + \gamma_n T_s x_n \\ r_{s+1,n} = \widehat{t}_{s+1,n} - \gamma_n \mathrm{prox}_{\gamma_n^{-1} R_s}(\gamma_n^{-1} \widehat{t}_{s+1,n}) + e_{s,n} \\ \widetilde{t}_{s+1,n} = r_{s+1,n} + \gamma_n T_s \widehat{x}_n \\ t_{s+1,n+1} = t_{s+1,n} - \widehat{t}_{s+1,n} + \widetilde{t}_{s+1,n} \end{array} \right. \\ \widetilde{x}_n = \widehat{x}_n - \gamma_n(A^\top r_{0,n} + B^\top r_{1,n} + \sum_{s=1}^{S} T_s^\top r_{s+1,n}) \\ x_{n+1} = x_n - \widehat{x}_n + \widetilde{x}_n. \end{array} \right.$$
$$\tag{12}$$

In this algorithm, $(\gamma_n)_{n \ge 0}$ is a sequence of step-sizes, and $e_{0,n} \in (\mathbb{R}^P)^2$, $e_{1,n} \in \mathbb{R}^{K_1}, \dots, e_{S,n} \in \mathbb{R}^{K_S}$ correspond to some possible additive summable errors in the computation

of the proximity operators.

If the set of solutions to Problem 1.1 is nonempty, then any sequence $(x_n)_{n \in \mathbb{N}}$ generated by Algorithm (12) converges to an element of this set (under a suitable weak qualification condition). Note that this algorithm has two interesting features. First, several operations (e.g. the loop on s) can be implemented in a parallel manner. Second, it is error-tolerant with respect to the computation of the proximity operators.

4. APPLICATION TO IMAGE REGISTRATION

4.1. Problem formulation

The objective of image registration is to determine spatial transformations that maximize a similarity metric between two images resulting from two different acquisitions. Let the two original images be represented by data vectors $I_1 \in \mathbb{R}^P$ and $I_2 \in \mathbb{R}^P$. For every $j \in \{1, 2\}$, let F_j be an image mapping operator such that

$$F_j \colon \mathbb{R}^P \times \mathbb{R}^{N_j} \to \mathbb{R}^P \colon (I_j, z_j) \mapsto F_j(I_j, z_j) \tag{13}$$

where $F_j(I_j, z_j)$ is the mapped image, and $z_j \in \mathbb{R}^{N_j}$ is a vector of mapping parameters [21, 22].

When a Kullback metric is adopted, the image registration criterion to be minimized w.r.t. $x = \begin{bmatrix} z_1^\top, & z_2^\top \end{bmatrix}^\top$ reads

$$D\big(F_1(I_1, z_1), F_2(I_2, z_2)\big).$$

To determine the optimal parameter vector $x \in \mathbb{R}^N$ with $N = N_1 + N_2$, one can proceed by supposing that F_1 (resp. F_2) is differentiable w.r.t. its second argument and by performing a first-order Taylor expansion around an initial estimate $\bar{z}_1$ (resp. $\bar{z}_2$) as follows: $(\forall j \in \{1, 2\})$

$$F_j(I_j, z_j) \simeq F_j(I_j, \bar{z}_j) + \frac{\partial F_j}{\partial z_j}(I_j, \bar{z}_j)(z_j - \bar{z}_j) \tag{14}$$

where $\frac{\partial F_j}{\partial z_j}(I_j, \bar{z}_j)$ is a Jacobian matrix. With the linearization expressed in (14),[1] the image registration problem can be reformulated under the form of Problem 1.1, where $A = \begin{bmatrix} \frac{\partial F_1}{\partial z_1}(I_1, \bar{z}_1) & \mathbf{0} \end{bmatrix}$, $B = \begin{bmatrix} \mathbf{0} & \frac{\partial F_2}{\partial z_2}(I_2, \bar{z}_2) \end{bmatrix}$, $u = F_1(I_1, \bar{z}_1) - \frac{\partial F_1}{\partial z_1}(I_1, \bar{z}_1)\bar{z}_1$, and $v = F_2(I_2, \bar{z}_2) - \frac{\partial F_2}{\partial z_2}(I_2, \bar{z}_2)\bar{z}_2$. The additional regularization terms $(R_s \circ T_s)_{1 \le s \le S}$ allow the incorporation of prior knowledge about the sought parameter vector x.

4.2. Experiments

We conducted experiments for disparity estimation under illumination variation, using a stereoscopic pair of grayscale images I_1 and I_2 with size $P_1 \times P_2$ ($P = P_1 P_2$). The related image mapping operators are given by

$$F_1(I_1, z_1) = \mathrm{vec}\left[(I_1^{(i_1 - z_1^{(i_1, i_2)}, i_2)})_{\substack{1 \le i_1 \le P_1 \\ 1 \le i_2 \le P_2}} \right] \tag{15}$$

$$F_2(I_2, z_2) = \mathrm{vec}\left[(z_2^{(i_1, i_2)} I_2^{(i_1, i_2)})_{\substack{1 \le i_1 \le P_1 \\ 1 \le i_2 \le P_2}} \right] \tag{16}$$

---

[1] In our experiments, the linearization is performed 3 times (by refining the initial estimate from a previous one).

a) Right image   c) True disparity   e) MAE=0.8287, Err= 3.36   g) MAE=0.8348, Err= 3.44   i) MAE=0.8385, Err= 3.62

MS − SSIM=0.9857   MS − SSIM= 0.9856   MS − SSIM= 0.9854

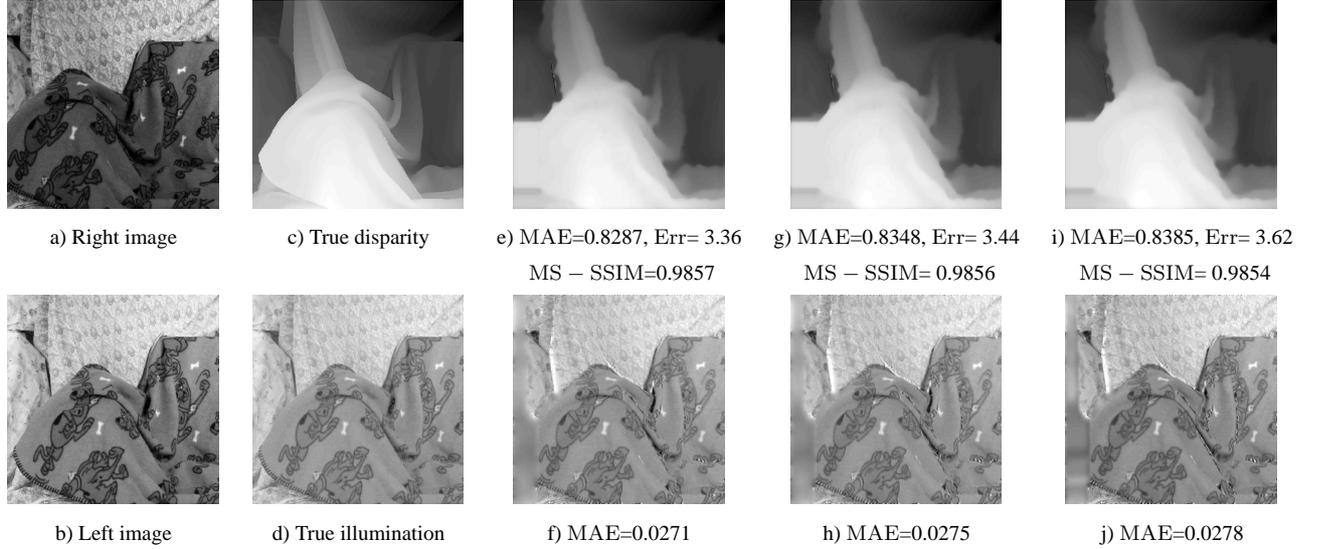b) Left image   d) True illumination   f) MAE=0.0271   h) MAE=0.0275   j) MAE=0.0278

Fig. 1. Results for "Cloth" stereo pair: a)-b) stereo images, c)-d) ground truths, M+LFBF algorithm: e)-f) JK divergence, g)-h) KL divergence and i)-j) Euclidean distance. Parameters: $((z_1)_{\min}, (z_1)_{\max}) = (15, 55)$, $((z_2)_{\min}, (z_2)_{\max}) = (0, 0.58)$, $(\kappa_1, \kappa_2) = (45000, 157)$.

where $I_1^{(i_1,i_2)}$ (resp. $I_2^{(i_1,i_2)}$) is the intensity value in the right (resp. left) view at pixel (i_1, i_2), $z_1^{(i_1,i_2)}$ denotes the disparity at (i_1, i_2), and $z_2^{(i_1,i_2)}$ is the associated illumination variation coefficient. The parameter vectors are here $z_1 = \text{vec}\left[(z_1^{(i_1,i_2)})_{1 \leq i_1 \leq P_1, 1 \leq i_2 \leq P_2}\right]$ and $z_2 = \text{vec}\left[(z_2^{(i_1,i_2)})_{1 \leq i_1 \leq P_1, 1 \leq i_2 \leq P_2}\right]$. It can be observed that

$$\frac{\partial F_1}{\partial z_1}(I_1, \bar{z}_1) = -\text{Diag}\left[(\nabla^{(1)} I_1^{(i_1 - z_1^{(i_1,i_2)}, i_2)})_{\substack{1 \leq i_1 \leq P_1 \\ 1 \leq i_2 \leq P_2}}\right]$$

$$\frac{\partial F_2}{\partial z_2}(I_2, \bar{z}_2) = \text{Diag}\left[(I_2^{(i_1,i_2)})_{\substack{1 \leq i_1 \leq P_1 \\ 1 \leq i_2 \leq P_2}}\right] \quad (17)$$

where $\nabla^{(1)}$ is the gradient operator w.r.t. the first spatial coordinate. It is useful to incorporate prior information about the solution so as to convert the original matching problem to a well-posed one. Similarly to [23], a first constraint set is introduced that takes into account the range of possible values for the disparity and the illumination variation:

$$C_1 = \left\{[z_1^\top, z_2^\top]^\top \in \mathbb{R}^{2P} \mid (z_1)_{\min} \leq z_1 \leq (z_1)_{\max},\right.$$
$$\left.(z_2)_{\min} \leq z_2 \leq (z_2)_{\max}\right\}. \quad (18)$$

A second constraint is employed to enforce the smoothness of the estimated fields:

$$C_2 = \{[z_1^\top, z_2^\top]^\top \in \mathbb{R}^{2P} \mid \text{TV}(z_1) \leq \kappa_1, \|\nabla z_2\|_{\ell_2}^2 \leq \kappa_2\} \quad (19)$$

where TV is the discrete total variation semi-norm, $\| \cdot \|_{\ell_2}$ is the ℓ_2 norm, and ∇ is the spatial gradient operator. Consequently, we have now $S = 2$ in Problem 1.1, where R_1 and R_2 reduce to indicator functions of convex sets, $T_1 = \mathbf{I}_{2P}$, and T_2 is a block diagonal matrix with diagonal terms set to gradient operators.

We now illustrate the practical performance of our method on the "Cloth" stereo pair downloaded from MiddleBury website.[2] The results are provided in Fig. 1. The quality of the results was evaluated using three different metrics based on the ground truth: the MAE (Mean Absolute Error) evaluated between the computed maps $(z_j)_{j \in \{1,2\}}$ and the ground truth $(\widetilde{z}_j)_{j \in \{1,2\}}$, the percentage of bad matching pixels $\text{B} = \frac{1}{P} \sum_{i_1=1}^{P_1} \sum_{i_2=1}^{P_2} \mathbf{1}(|z_j^{(i_1,i_2)} - \widetilde{z}_j^{(i_1,i_2)}| > 2)$, and the MS-SSIM (multi-scale structure similarity index) [24]. The JK divergence leads to better results than the KL divergence which itself is better than the standard Euclidean distance.

5. CONCLUSION

Existing approaches for optimizing convex criteria involving the KL divergence are often restricted to problems where the minimization is performed w.r.t. one of the arguments of the divergence, or they are based on an alternating minimization process which requires specific assumptions to be valid. In this work, we have developed a novel proximal method that allows us to address more general forms of optimization problems. Few results exist concerning the expressions of the proximity operators of two-variable convex functions. The expressions we have derived for KL and JK divergences enrich the list of functions for which such proximity operators can be easily computed. In addition to its flexibility, the proposed approach leads to parallel proximal algorithms that can be efficiently implemented on multicore architectures. An application to image registration has been provided in this work so as to demonstrate the good performance of the proposed proximal algorithm. Further applications of this new approach will be investigated in our future work.

---

[2]http://vision.middlebury.edu/stereo/

6. REFERENCES

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

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

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

[4] 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.

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

[6] 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.

[7] F.-X. Dupé, J.M. 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.

[8] 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, Sep. 2011.

[9] T. Teuber, G. Steidl, and R. H. Chan, "Minimization and parameter estimation for seminorm regularization models with I-divergence constraints," *Preprint University of Kaiserslautern*, 2012.

[10] 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. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds., pp. 185–212. Springer-Verlag, New York, 2011.

[11] 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.

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

[13] 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.

[14] J.-C. Pesquet and M. El-Gheche, "Proximity operators of some discrete information divergences," Tech. Rep., Paris-Est University, 2012.

[15] M.A.T. Figueiredo and R.D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906 – 916, Aug. 2003.

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

[17] H. Raguet, J. Fadili, and G. Peyré, "Generalized Forward-Backward splitting," Tech. Rep., Preprint Hal-00613637, 2011.

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

[19] 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.

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

[21] Yun He, A. B. Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1211 – 1220, May 2003.

[22] A. B. Hamza and H. Krim, "Image registration and segmentation by maximizing the Jensen-Rényi divergence," vol. 2683 of *Lecture Notes in Computer Science*, pp. 147–163. 2003.

[23] C. Chaux, M. El-Gheche, J. Farah, J.-C. Pesquet, and B. Pesquet-Popescu, "A parallel proximal splitting method for disparity estimation from multicomponent images under illumination variation," *J. Math. Imaging Vision*, pp. 1–12, Jun. 2012.

[24] W. S. Malpica and A. C. Bovik, "Range image quality assessment by structural similarity," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Washington, DC, USA, 19-24 Apr. 2009, pp. 1149–1152.