Playing with Duality: An Overview of Recent Primal-Dual Approaches for Solving Large-Scale Optimization Problems

Nikos Komodakis, Jean-Christophe Pesquet

Playing with Duality: An Overview of Recent Primal-Dual Approaches for Solving Large-Scale Optimization Problems

Nikos Komodakis, *Member, IEEE*, and Jean-Christophe Pesquet, *Fellow, IEEE*

Abstract

Optimization methods are at the core of many problems in signal/image processing, computer vision, and machine learning. For a long time, it has been recognized that looking at the dual of an optimization problem may drastically simplify its solution. Deriving efficient strategies jointly bringing into play the primal and the dual problems is however a more recent idea which has generated many important new contributions in the last years. These novel developments are grounded on recent advances in convex analysis, discrete optimization, parallel processing, and nonsmooth optimization with emphasis on sparsity issues. In this paper, we aim at presenting the principles of primal-dual approaches, while giving an overview of numerical methods which have been proposed in different contexts. We show the benefits which can be drawn from primal-dual algorithms both for solving large-scale convex optimization problems and discrete ones, and we provide various application examples to illustrate their usefulness.

Index Terms

Convex optimization, discrete optimization, duality, linear programming, proximal methods, inverse problems, computer vision, machine learning, big data

N. Komodakis (corresponding author) and J.-C. Pesquet are with the Laboratoire d'Informatique Gaspard Monge, UMR CNRS 8049, Université Paris-Est, 77454 Marne la Vallée Cedex 2, France. E-mail: nikos.komodakis@enpc.fr, jean-christophe.pesquet@univ-paris-est.fr.

I. Motivation and importance of the topic

Optimization is an extremely popular paradigm which constitutes the backbone of many branches of applied mathematics and engineeering, such as signal processing, computer vision, machine learning, inverse problems, and network communications, to mention just a few. The popularity of optimization approaches often stems from the fact that many problems from the above fields are typically characterized by a lack of closed form solutions and by uncertainties. In signal and image processing, for instance, uncertainties can be introduced due to noise, sensor imperfectness, or ambiguities that are often inherent in the visual interpretation. As a result, perfect or exact solutions hardly exist, whereas inexact but optimal (in a statistical or an application-specific sense) solutions and their efficient computation is what one aims at. At the same time, one important characteristic that is nowadays shared by increasingly many optimization problems encountered in the above areas is the fact that these problems are often of very large scale. A good example is the field of computer vision where one often needs to solve low level problems that require associating at least one (and typically more than one) variable to each pixel of an image (or even worse of an image sequence as in the case of video). This leads to problems that easily can contain millions of variables, which are therefore the norm rather than the exception in this context. Similarly, in fields like machine learning, due to the great ease with which data can now be collected and stored, quite often one has to cope with truly massive datasets and to train very large models, which thus naturally lead to optimization problems of very high dimensionality. Of course, a similar situation arises in many other scientific domains, including application areas such as inverse problems (e.g., medical image reconstruction or satellite image restoration) or telecommunications (e.g., network design, network provisioning) and industrial engineering. Due to this fact, computational efficiency constitutes a major issue that needs to be thoroughly addressed. This, therefore, makes mandatory the use of tractable optimization techniques that are able to properly exploit the problem structures, but which at the same time remain applicable to a class of problems as wide as possible.

A bunch of important advances that took place in this regard over the last years concerns a particular class of optimization approaches known as *primal-dual* methods. As their name implies, these approaches proceed by concurrently solving a primal problem (corresponding to the original optimization task) as well as a dual formulation of this problem. As it turns out, in doing so they are able to exploit more efficiently the problem specific properties, thus offering in many cases important computational advantages, some of which are briefly mentioned next for two very broad classes of problems.

1) Convex optimization: Primal-dual methods have been primarily employed in convex optimization problems where strong duality holds. They have been successfully applied to various types of nonlinear and nonsmooth cost functions that are prevalent in the above-mentioned application fields.

Many such applied problems can essentially be expressed under the form of a minimization of a sum of terms, where each term is given by the composition of a convex function with a linear operator. One first advantage of primal-dual methods pertains to the fact that they can yield very efficient splitting optimization schemes, according to which a solution to the original problem is iteratively computed through solving a sequence of easier subproblems,

each one involving only one of the terms appearing in the objective function.

The resulting primal-dual splitting schemes can also handle both differentiable and nondifferentiable terms, the former by use of gradient operators (i.e., through explicit steps) and the latter by use of proximity operators (i.e., through implicit steps). Depending on the target functions, either explicit or implicit steps may be easier to implement. Therefore, the derived optimization schemes exploit the properties of the input problem, in a flexible manner, thus leading to very efficient first-order algorithms.

Even more importantly, primal-dual techniques are able to achieve what is known as *full splitting* in the optimization literature, meaning that each of the operators involved in the problem (i.e., not only the gradient or proximity operators but also the involved linear operators) is used separately. As a result, no call to the inversion of a linear operator, which is an expensive operation for large scale problems, is required during the optimization process. This is an important feature which gives these methods a significant computational advantage compared with all other splitting-based approaches.

Last but not least, primal-dual methods lead to algorithms that are easily parallelizable, which is nowadays becoming increasingly important for efficiently handling high-dimensional problems.

2) Discrete optimization: Besides convex optimization, another important area where primal-dual methods play a prominent role is discrete optimization. This is of particular significance given that a large variety of tasks from signal processing, computer vision, and pattern recognition are formulated as discrete labeling problems, where one seeks to optimize some measure related to the quality of the labeling. This includes, for instance, tasks such as image segmentation, optical flow estimation, image denoising, stereo matching, to mention a few examples from image analysis. The resulting discrete optimization problems not only are of very large size, but also typically exhibit highly nonconvex objective functions, which are generally NP-hard to optimize.

Similarly to the case of convex optimization, primal-dual methods again offer many computational advantages, leading often to very fast graph-cut or message-passing-based algorithms, which are also easily parallelizable, thus providing in many cases a very efficient way for handling discrete optimization problems that are encountered in practice.

Besides being efficient, they are also successful in making little compromises regarding the quality of the estimated solutions. Techniques like the so-called *primal-dual schema* are known to provide a principled way for deriving powerful approximation algorithms to NP-hard combinatorial problems, thus allowing primal-dual methods to often exhibit theoretical (i.e., worst-case) approximation properties.

Furthermore, apart from the aforementioned worst-case guaranties, primal-dual algorithms can also provide (for free) *per-instance* approximation guaranties. This is essentially made possible by the fact that these methods are estimating not only primal but also dual solutions.

Goals of this tutorial paper. Based on the above observations, our objectives will be the following:

i) To provide a thorough introduction that intuitively explains the basic principles and ideas behind primal-dual approaches.

ii) To describe how these methods can be employed both in the context of continuous optimization and in the context of discrete optimization.

iii) To explain some of the recent advances that have taken place concerning primal-dual algorithms for solving large-scale optimization problems.

iv) To detail useful connections between primal-dual methods and some widely used optimization techniques like the alternating direction method of multipliers (ADMM).

v) Finally, to provide examples of useful applications in the context of image analysis and signal processing.

The remainder of the paper is structured as follows. In Section II, we introduce the necessary methodological background on optimization. Our presentation is grounded on the powerful notion of duality known as Frenchel's duality, from which duality properties in linear programming can be deduced. We also introduce useful tools from functional analysis and convex optimization, including the notions of subgradient and subdifferential, conjugate function, and proximity operator. The following two sections explain and describe various primal-dual methods. Section III is devoted to convex optimization problems. We discuss the merits of various algorithms and explain their connections with ADMM, that we show to be a special case of primal-dual proximal method. Section IV deals with primal-dual methods for discrete optimization. We explain how to derive algorithms of this type based on the primal-dual schema which is a well-known approximation technique in combinatorial optimization, and we also present primal-dual methods based on LP relaxations and dual decomposition. In Section V, we present applications from the domains of signal processing and image analysis, including inverse problems and computer vision tasks related to Markov Random Field energy minimization. In Section VI, we finally conclude the tutorial with a brief summary and discussion.

II. Optimization background

In this section, we introduce the necessary mathematical definitions and concepts used for introducing primal-dual algorithms in later sections.

A. Notation

In this paper, we will consider functions from $\mathbb{R}^N$ to $]-\infty,+\infty]$. The fact that we allow functions to take $+\infty$ value is useful in modern optimization to discard some "forbidden part" of the space when searching for an optimal solution (for example, in image processing problems, the components of the solution often are intensity values which must be nonnegative). The *domain* of a function $f\colon \mathbb{R}^N \to]-\infty,+\infty]$ is the subset of $\mathbb{R}^N$ where this function takes finite values, i.e. $\operatorname{dom} f = \big\{x \in \mathbb{R}^N \;\big|\; f(x) < +\infty\big\}$. A function with a nonempty domain is said to be *proper*. A function f is *convex* if

$$\big(\forall (x,y) \in (\mathbb{R}^N)^2\big)(\forall \lambda \in [0,1]) \qquad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{1}$$

The class of functions for which most of the main results in convex analysis have been established is $\Gamma_0(\mathbb{R}^N)$, the class of proper, convex, lower-semicontinuous functions from $\mathbb{R}^N$ to $]-\infty,+\infty]$. Recall that a function $f\colon \mathbb{R}^N \to$

$]-\infty,+\infty]$ is lower-semicontinuous if its *epigraph* $\operatorname{epi} f = \big\{(x,\zeta) \in \operatorname{dom} f \times \mathbb{R} \;\big|\; f(x) \leq \zeta\big\}$ is a closed set (see Fig. 1).

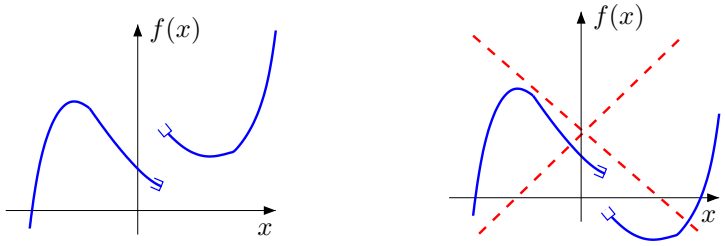


Fig. 1: Illustration of the lower-semicontinuity property.

If C is a nonempty subset of $\mathbb{R}^N$, the *indicator function* of C is defined as

$$(\forall x \in \mathbb{R}^N) \qquad \iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise.} \end{cases} \tag{2}$$

This function belongs to $\Gamma_0(\mathbb{R}^N)$ if and only if C is a nomempty closed convex set.

The Moreau *subdifferential* of a function $f\colon \mathbb{R}^N \to]-\infty,+\infty]$ at $x \in \mathbb{R}^N$ is defined as

$$\partial f(x) = \big\{u \in \mathbb{R}^N \;\big|\; (\forall y \in \mathbb{R}^N)\;\; f(y) \geq f(x) + u^\top (y-x)\big\}. \tag{3}$$

Any vector u in $\partial f(x)$ is called a *subgradient* of f at x (see Fig. 2).

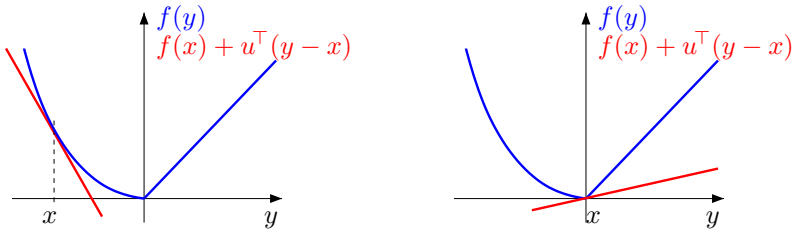


Fig. 2: Examples of subgradients u of a function f at x.

Fermat's rule states that 0 is a subgradient of f at x if and only if x belongs to the set of global minimizers of f. If f is a proper convex function which is differentiable at x, then its subdifferential at x reduces to the singleton consisting of its gradient, i.e. $\partial f(x) = \{\nabla f(x)\}$. Note that, in the nonconvex case, extended definitions of the subdifferential may be useful such as the *limiting subdifferential* [1], but this one reduces to the Moreau subdifferential when the function is convex.

A concept which has been of growing importance in recent developments in optimization is the concept of *proximity operator*. It must be pointed out that the proximity operator was introduced in the early work by J. J. Moreau (1923-2014) [2]. The proximity operator of a function $f \in \Gamma_0(\mathbb{R}^N)$ is defined as

$$\operatorname{prox}_f\colon \mathbb{R}^N \to \mathbb{R}^N\colon x \mapsto \operatorname*{argmin}_{y \in \mathbb{R}^N}\;\; f(y) + \frac{1}{2}\|y-x\|^2 \tag{4}$$

where $\|\cdot\|$ denotes the Euclidean norm. For every $x \in \mathbb{R}^N$, $\operatorname{prox}_f x$ can thus be intrepreted as the result of a regularized minimization of f in the neighborhood of x. Note that the minimization to be performed to calculate $\operatorname{prox}_f x$ always has a unique solution. An example of proximity operator is shown in Fig. 3.

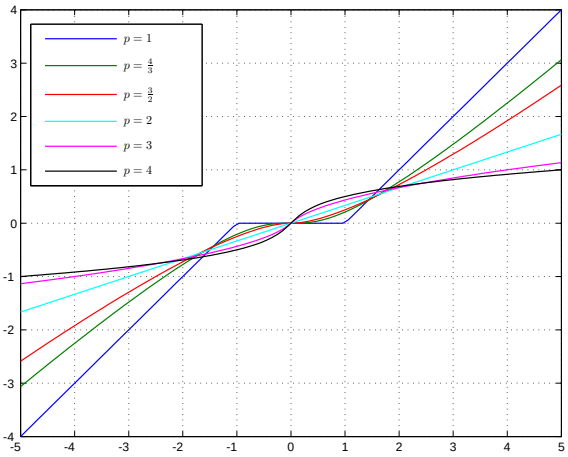


Fig. 3: Graph of $\operatorname{prox}_{|\cdot|^p}$.

In the case when f is equal to the indicator function of a nonempty closed convex set C, the proximity operator of f reduces to the projection P_C onto this set, i.e. $(\forall x \in \mathbb{R}^N)\ P_C x = \operatorname*{argmin}_{y \in C} \|y - x\|$.

This shows that proximity operators can be viewed as extensions of projections onto convex sets. The proximity operator enjoys many properties of the projection, in particular it is firmly nonexpansive which makes it successful in ensuring the convergence of fixed point algorithms grounded on its use. For more details about proximity operators and their rich properties, the reader is refered to the tutorial papers in [3]–[5]. The definition of the proximity operator can be extended to nonconvex lower-semicontinuous functions which are lower bounded by an affine function, but $\operatorname{prox}_f x$ is no longer guaranteed to be uniquely defined at any given point x.

B. Conjugate function

A fundamental notion when dealing with duality issues is the notion of *conjugate function*. The conjugate of a function $f\colon \mathbb{R}^N \to]-\infty, +\infty]$ is the function f^* defined as

$$f^*\colon \mathbb{R}^N \to]-\infty, +\infty] : u \mapsto \sup_{x \in \mathbb{R}^N} \big(x^\top u - f(x)\big). \tag{5}$$

This concept was introduced by A. M. Legendre (1752-1833) in the one-variable case, and it was generalized by M. W. Fenchel (1905-1988). A graphical illustration of the conjugate function is provided in Fig. 4. In particular, for every vector $x \in \mathbb{R}^N$ such that the supremum in (5) is attained, u is a subgradient of f at x.

It must be emphasized that, even if f is nonconvex, f^* is a (non necessarily proper) lower-semicontinuous convex function. In addition, when $f \in \Gamma_0(\mathbb{R}^N)$, then $f^* \in \Gamma_0(\mathbb{R}^N)$, and also the biconjugate of f (that it the conjugate

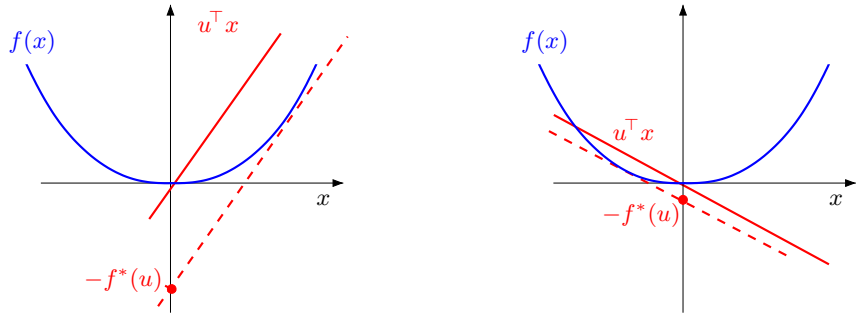


Fig. 4: Graphical interpretation of the conjugate function.

of its conjugate) is equal to f. This means that we can express any function f in $\Gamma_0(\mathbb{R}^N)$ as

$$(\forall x \in \mathbb{R}^N) \qquad f(x) = \sup_{u \in \mathbb{R}^N} \big(u^\top x - f^*(u)\big). \tag{6}$$

A geometrical interpretation of this result is that the epigraph of any proper lower-semicontinuous convex function always is an intersection of closed half-spaces.

As we have seen, the subdifferential plays an important role in the characterization of the minimizers of a function. A natural question is thus to enquire about the relations existing between the subdifferential of a function $f\colon \mathbb{R}^N \to \,]-\infty,+\infty]$ and the subdifferential of its conjugate function. An answer is provided by the following important properties:

$$\begin{aligned} u \in \partial f(x) \quad &\Rightarrow \quad x \in \partial f^*(u) \qquad \text{if } f \text{ is proper} \\ u \in \partial f(x) \quad &\Leftrightarrow \quad x \in \partial f^*(u) \qquad \text{if } f \in \Gamma_0(\mathbb{R}^N). \end{aligned} \tag{7}$$

Another important property is Moreau's decomposition formula which links the proximity operator of a function $f \in \Gamma_0(\mathbb{R}^N)$ to the proximity operator of its conjugate:

$$(\forall x \in \mathbb{R}^N)(\forall \gamma \in \,]0,+\infty[) \qquad x = \operatorname{prox}_{\gamma f} x + \gamma \operatorname{prox}_{\gamma^{-1} f^*}(\gamma^{-1} x). \tag{8}$$

Other useful properties of the conjugation operation are listed in Table I, where a parallel is drawn with the mutidimensional Fourier transform. Conjugation also makes it possible to build an insightful bridge between the main two kinds of nonsmooth convex functions encountered in signal and image processing problems, namely indicator functions of feasibility constraints and sparsity measures (see framebox on page 8).

TABLE I: Parallel between properties of the Legendre-Fenchel conjugation [3] and of the Fourier transform. f is a function defined on $\mathbb{R}^N$, f^* denotes its conjugate, $\widehat{f}$ is its Fourier transform such that $\widehat{f}(\nu) = \int_{\mathbb{R}^N} f(x) \exp(-\jmath 2\pi x^\top \nu) dx$ where $\nu \in \mathbb{R}^N$ and $\jmath$ is the imaginary unit (a similar notation is used for other functions), h, g, and $(f_m)_{1\leq m\leq M}$ are functions defined on $\mathbb{R}^N$, $(\varphi_j)_{1\leq j\leq N}$ are functions defined on $\mathbb{R}$, ψ is an even function defined on $\mathbb{R}$, $\widetilde{\psi}$ is defined as $\widetilde{\psi}(\rho) = 2\pi\rho^{(2-N)/2} \int_0^{+\infty} r^{N/2} J_{(N-2)/2}(2\pi r\rho)\psi(r)dr$ where $\rho \in \mathbb{R}$ and $J_{(N-2)/2}$ is the Bessel function of order $(N-2)/2$, and δ denotes the Dirac distribution. (Some properties of the Fourier transform may require some technical assumptions.)

		conjugate		Fourier transform	
	Property	$h(x)$	$h^*(u)$	$h(x)$	$\widehat{h}(\nu)$
i	invariant function	$\frac{1}{2}\|x\|^2$	$\frac{1}{2}\|u\|^2$	$\exp(-\pi\|x\|^2)$	$\exp(-\pi\|\nu\|^2)$
ii	translation $c \in \mathbb{R}^N$	$f(x-c)$	$f^*(u) + c^\top u$	$f(x-c)$	$\exp(-\jmath 2\pi c^\top \nu)\widehat{f}(\nu)$
iii	dual translation $c \in \mathbb{R}^N$	$f(x) + c^\top x$	$f^*(u-c)$	$\exp(\jmath 2\pi c^\top x) f(x-c)$	$\widehat{f}(\nu - c)$
iv	scalar multiplication $\alpha \in \ ]0, +\infty[$	$\alpha f(x)$	$\alpha f^*\left(\frac{u}{\alpha}\right)$	$\alpha f(x)$	$\alpha\widehat{f}(\nu)$
v	invertible linear transform $L \in \mathbb{R}^{N\times N}$ invertible	$f(Lx)$	$f^*\big((L^{-1})^\top u\big)$	$f(Lx)$	$\frac{1}{\lvert\det(L)\rvert}\widehat{f}\big((L^{-1})^\top \nu\big)$
vi	scaling $\alpha \in \mathbb{R}^*$	$f\left(\frac{x}{\alpha}\right)$	$f^*(\alpha u)$	$f\left(\frac{x}{\alpha}\right)$	$\lvert\alpha\rvert\widehat{f}(\alpha\nu)$
vii	reflection	$f(-x)$	$f^*(-u)$	$f(-x)$	$\widehat{f}(-\nu)$
viii	separability	$\sum_{j=1}^N \varphi_j(x^{(j)})$ $x = (x^{(j)})_{1\leq j\leq N}$	$\sum_{j=1}^N \varphi_j^*(u^{(j)})$ $u = (u^{(j)})_{1\leq j\leq N}$	$\prod_{j=1}^N \varphi_j(x^{(j)})$ $x = (x^{(j)})_{1\leq j\leq N}$	$\prod_{j=1}^N \widehat{\varphi}_j(\nu^{(j)})$ $\nu = (\nu^{(j)})_{1\leq j\leq N}$
ix	isotropy	$\psi(\|x\|)$	$\psi^*(\|u\|)$	$\psi(\|x\|)$	$\widetilde{\psi}(\|\nu\|)$
x	inf-convolution /convolution	$(f \,\square\, g)(x)$ $= \inf_{y\in\mathbb{R}^N} f(y) + g(x-y)$	$f^*(u) + g^*(u)$	$(f \star g)(x)$ $= \int_{\mathbb{R}^N} f(y)g(x-y)dy$	$\widehat{f}(\nu)\widehat{g}(\nu)$
xi	sum/product	$f(x) + g(x)$ $f \in \Gamma_0(\mathbb{R}^N)$, $g \in \Gamma_0(\mathbb{R}^N)$ $\operatorname{dom} f \cap \operatorname{dom} g \neq \varnothing$	$(f^* \,\square\, g^*)(u)$	$f(x)g(x)$	$(\widehat{f} \star \widehat{g})(\nu)$
xii	identity element of convolution	$\iota_{\{0\}}(x)$	0	$\delta(x)$	1
xiii	identity element of addition/product	0	$\iota_{\{0\}}(u)$	1	$\delta(\nu)$
xiv	offset $\alpha \in \mathbb{R}$	$f(x) + \alpha$	$f^*(u) - \alpha$	$f(x) + \alpha$	$\widehat{f}(\nu) + \alpha\delta(\nu)$
xv	infinum/sum	$\inf_{1\leq m\leq M} f_m(x)$	$\sup_{1\leq m\leq M} f_m^*(u)$	$\sum_{m=1}^M f_m(x)$	$\sum_{m=1}^M \widehat{f}_m(\nu)$
xvi	value at 0	$f^*(0) = -\inf f$		$\widehat{f}(0) = \int_{\mathbb{R}^N} f(x)dx$	

CONJUGATES OF SUPPORT FUNCTIONS ([3])

The support function of a set $C \subset \mathbb{R}^N$ is defined as

$$(\forall u \in \mathbb{R}^N) \qquad \sigma_C(u) = \sup_{x \in C} x^\top u. \tag{9}$$

In fact, a function f is the support function of a nonempty closed convex set C if and only if it belongs to $\Gamma_0(\mathbb{R}^N)$ and it is positively homogeneous [6], i.e.

$$(\forall x \in \mathbb{R}^N)(\forall \alpha \in \,]0, +\infty[) \qquad f(\alpha x) = \alpha f(x).$$

Examples of such functions are norms, e.g. the ℓ_1-norm:

$$\big(\forall x = (x^{(j)})_{1 \leq j \leq N} \in \mathbb{R}^N\big) \qquad f(x) = \|x\|_1 = \sum_{j=1}^{N} |x^{(j)}|$$

which is a useful convex sparsity-promoting measure in LASSO estimation [7] and in compressive sensing [8]. Another famous example is the Total Variation semi-norm [9] which is popular in image processing for retrieving constant areas with sharp contours. An important property is that, if C is a nonempty closed convex set, its conjugate function is the indicator function of C. For example, the conjugate function of the ℓ_1-norm is the indicator function of the hypercube $[-1, 1]^N$. This shows that sparsity measures are equivalent in the dual domain to some hard constraints.

C. Duality results

A wide array of problems in signal and image processing can be expressed under the following variational form:

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; f(x) + g(Lx) \tag{10}$$

where $f \colon \mathbb{R}^N \to \,]-\infty, +\infty]$, $g \colon \mathbb{R}^K \to \,]-\infty, +\infty]$, and $L \in \mathbb{R}^{K \times N}$. Problem (10) is usually referred to as the *primal problem* which is associated with the following *dual problem* [6], [10], [11]:

$$\underset{v \in \mathbb{R}^K}{\text{minimize}} \;\; f^*(-L^\top v) + g^*(v). \tag{11}$$

The latter problem may be easier to solve than the former one, especially when K is much smaller than N.

A question however is to know whether solving the dual problem may bring some information on the solution of the primal one. A first answer to this question is given by the Fenchel-Rockafellar duality theorem which basically states that solving the dual problem provides a lower bound on the minimum value which can be obtained in the primal one. More precisely, if f and g are proper functions and if μ and μ^* denote the infima of the functions minimized in the primal and dual problems, respectively, then *weak duality* holds, which means that $\mu \geq -\mu^*$. If μ is finite, $\mu + \mu^*$ is called the *duality gap*. In addition, if $f \in \Gamma_0(\mathbb{R}^N)$ and $g \in \Gamma_0(\mathbb{R}^K)$, then, under appropriate qualification conditions,[1] there always exists a solution to the dual problem and the duality gap vanishes. When the duality gap is equal to zero, it is said that *strong duality* holds.

[1]For example, this property is satisfied if the intersection of the interior of the domain of g and the image of the domain of f by L is nonempty.

CONSENSUS AND SHARING ARE DUAL PROBLEMS

Suppose that our objective is to minimize a separable function $\sum_{m=1}^{M} g_m$ where the potential $g_m \colon \mathbb{R}^N \to]-\infty,+\infty]$ is computed at the vertex of index $m \in \{1,\ldots,M\}$ of a graph. A classical technique to perform this task in a distributed or parallel manner [12] consists of reformulating this problem as a *consensus problem*, where a variable is assigned to each vertex, and the defined variables $x_1,\ldots,x_M$ are updated so as to reach a consensus: $x_1 = \ldots = x_M$. This means that, in the product space $(\mathbb{R}^N)^M$ the original optimization problem can be rewritten as

$$\underset{\boldsymbol{x}=(x_1,\ldots,x_M)\in(\mathbb{R}^N)^M}{\text{minimize}} \ \iota_D(\boldsymbol{x}) + \underbrace{\sum_{m=1}^{M} g_m(x_m)}_{g(\boldsymbol{x})}$$

where D is the vector space defined as $D = \big\{\boldsymbol{x} = (x_1,\ldots,x_M) \in (\mathbb{R}^N)^M \;\big|\; x_1 = \ldots = x_M\big\}$.
By noticing that the conjugate of the indicator function of a vector space is the indicator function of its orthogonal complement, it is easy to see that the dual of this consensus problem has the following form:

$$\underset{\boldsymbol{v}=(v_1,\ldots,v_M)\in(\mathbb{R}^N)^M}{\text{minimize}} \ \iota_{D^\perp}(\boldsymbol{v}) + \underbrace{\sum_{m=1}^{M} g_m^*(v_m)}_{g^*(\boldsymbol{v})}$$

where $D^\perp = \big\{\boldsymbol{v} = (v_1,\ldots,v_M) \in (\mathbb{R}^N)^M \;\big|\; v_1 + \cdots + v_M = 0\big\}$ is the orthogonal complement of D. By making the variable change $(\forall m \in \{1,\ldots,M\})\ v_m = u_m - u/M$ where u is some given vector in $\mathbb{R}^N$, and by setting $h_m(u_m) = g_m^*(u_m - u/M)$, the latter minimization can be reexpressed as

$$\underset{\substack{u_1\in\mathbb{R}^N,\ldots,u_M\in\mathbb{R}^N\\ u_1+\cdots+u_M=u}}{\text{minimize}} \ \sum_{m=1}^{M} h_m(u_m).$$

This problem is known as a *sharing problem* where one wants to allocate a given ressource u between M agents while maximizing the sum of their welfares evaluated through their individual utility functions $(h_m)_{1\le m\le M}$.

Another useful result follows from the fact that, by using the definition of the conjugate function of g, Problem (10) can be reexpressed as the following saddle-point problem:

$$\text{Find} \qquad \inf_{x\in\mathbb{R}^N} \sup_{v\in\mathbb{R}^K} \big(f(x) + v^\top L x - g^*(v)\big). \tag{12}$$

In order to find a saddle point $(\widehat{x},\widehat{v}) \in \mathbb{R}^N \times \mathbb{R}^K$, it thus appears natural to impose the inclusion relations:

$$-L^\top \widehat{v} \in \partial f(\widehat{x}), \qquad L\widehat{x} \in \partial g^*(\widehat{v}). \tag{13}$$

A pair $(\widehat{x},\widehat{v})$ satisfying the above conditions is called a *Kuhn-Tucker point*. Actually, under some technical assumption, by using Fermat's rule and (7), it can be proved that, if $(\widehat{x},\widehat{v})$ is a Kuhn-Tucker point, then $\widehat{x}$ is a solution to the primal problem and $\widehat{v}$ is a solution to the dual one. This property especially holds when $f \in \Gamma_0(\mathbb{R}^N)$ and $g \in \Gamma_0(\mathbb{R}^K)$.

D. Duality in linear programming

In linear programming (LP) [13], we are interested in convex optimization problems of the form:

$$\text{Primal-LP :} \qquad \underset{x\in[0,+\infty[^N}{\text{minimize}} \ c^\top x \quad \text{s.t.} \quad Lx \ge b, \tag{14}$$

where $L = (L^{(i,j)})_{1\leq i\leq K, 1\leq j\leq N} \in \mathbb{R}^{K\times N}$, $b \in \mathbb{R}^K$, and $c \in \mathbb{R}^N$.[2] The above formulation can be viewed as a special case of (10) where

$$(\forall x \in \mathbb{R}^N) \quad f(x) = c^\top x + \iota_{[0,+\infty[^N}(x), \qquad (\forall z \in \mathbb{R}^K) \quad g(z) = \iota_{[0,+\infty[^N}(z-b). \tag{15}$$

By using the properties of the conjugate function and by setting $y = -v$, it is readily shown that the dual problem (11) can be reexpressed as

$$\text{Dual-LP}: \qquad \underset{y\in[0,+\infty[^K}{\text{maximize}} \; b^\top y \quad \text{s.t.} \quad L^\top y \leq c. \tag{16}$$

Since f is a convex function, strong duality holds in LP. If $\widehat{x} = (\widehat{x}^{(j)})_{1\leq j\leq N}$ is a solution to Primal-LP, a solution $\widehat{y} = (\widehat{y}^{(i)})_{1\leq i\leq K}$ to Dual-LP can be obtained by the *primal complementary slackness condition*:

$$(\forall j \in \{1,\ldots,N\}) \quad \text{such that} \quad \widehat{x}^{(j)} > 0, \qquad \sum_{i=1}^{K} L^{(i,j)}\,\widehat{y}^{(i)} = c^{(j)}. \tag{17}$$

whereas, if $\widehat{y}$ is a solution to Dual-LP, a solution $\widehat{x}$ to Primal-LP can be obtained by the *dual complementary slackness condition*:

$$(\forall i \in \{1,\ldots,K\}) \quad \text{such that} \quad \widehat{y}^{(i)} > 0, \qquad \sum_{j=1}^{N} L^{(i,j)}\,\widehat{x}^{(j)} = b^{(i)}. \tag{18}$$

III. Convex optimization algorithms

In this section, we present several primal-dual splitting methods for solving convex optimization problems, starting from the basic forms to the more sophisticated highly parallelized ones.

A. Problem

A wide range of convex optimization problems can be formulated as follows:

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \;\; f(x) + g(Lx) + h(x). \tag{19}$$

where $f \in \Gamma_0(\mathbb{R}^N)$, $g \in \Gamma_0(\mathbb{R}^K)$, $L \in \mathbb{R}^{N\times K}$, and $h \in \Gamma_0(\mathbb{R}^N)$ is a differentiable function having a Lipschitzian gradient with a Lipschitz constant $\beta \in \,]0,+\infty[$. The latter assumption means that the gradient ∇h of h is such that

$$\big(\forall (x,y) \in (\mathbb{R}^N)^2\big) \qquad \|\nabla h(x) - \nabla h(y)\| \leq \beta\|x-y\|. \tag{20}$$

For examples, the functions f, $g\circ L$, and h may model various data fidelity terms and regularization functions encountered in the solution of inverse problems. In particular, the Lipschitz differentiability property is satisfied for least squares criteria.

With respect to Problem (10), we have introduced an additional smooth term h. This may be useful in offering more flexibility for taking into account the structure of the problem of interest and the properties of the involved objective function. We will however see that not all algorithms are able to possibly take advantage of the fact that h is a smooth term.

[2]The vector inequality in (14) means that $Lx - b \in [0,+\infty[^K$.

Based on the results in Section II-C and Property (xi) in Table I, the dual optimization problem reads:

$$\underset{v\in\mathbb{R}^K}{\text{minimize}} \;\; (f^* \,\square\, h^*)(-L^\top v) + g(v). \tag{21}$$

Note that, in the particular case when $h = 0$, the inf-convolution $f^* \,\square\, h^*$ (see the definition in Table I(x)) of the conjugate functions of f and h reduces to f^* and we recover the basic form (11) of the dual problem.

The common trick used in the algorithms which will be presented in this section is to solve jointly Problems (19) and (21), instead of focusing exclusively on either (19) or (21). More precisely, these algorithms aim at finding a Kuhn-Tucker point $(\widehat{x}, \widehat{v}) \in \mathbb{R}^N \times \mathbb{R}^K$ such that

$$-L^\top \widehat{v} - \nabla h(\widehat{x}) \in \partial f(x) \quad \text{and} \quad L\widehat{x} \in \partial g^*(\widehat{v}). \tag{22}$$

B. ADMM

The celebrated ADMM (Alternating Direction Method of Multipliers) can be viewed as a primal-dual algorithm. This algorithm belongs to the class of *augmented Lagrangian* methods since a possible way of deriving this algorithm consists of looking for a saddle point of an augmented version of the classical Lagrange function [12]. This augmented Lagrangian is defined as

$$\big(\forall (x, y, z) \in \mathbb{R}^N \times (\mathbb{R}^K)^2\big) \qquad \widetilde{\mathcal{L}}(x, y, z) = f(x) + h(x) + g(y) + \gamma z^\top (Lx - y) + \frac{\gamma}{2}\|Lx - y\|^2 \tag{23}$$

where $\gamma \in \,]0, +\infty[$ and γz corresponds to a Lagrange multiplier. The resulting iterations are given in Algorithm 1. This algorithm has been known for a long time [14], [15] although it has attracted recently much interest in the

Algorithm 1 ADMM

Set $y_0 \in \mathbb{R}^K$ and $z_0 \in \mathbb{R}^K$

Set $\gamma \in \,]0, +\infty[$

For $n = 0, 1, \ldots$

$$\begin{aligned}
x_n &= \underset{x\in\mathbb{R}^N}{\operatorname{argmin}} \;\; \tfrac{1}{2}\,\|Lx - y_n + z_n\|^2 + \tfrac{1}{\gamma}\big(f(x) + h(x)\big) \\
s_n &= Lx_n \\
y_{n+1} &= \operatorname{prox}_{\frac{g}{\gamma}}\,(z_n + s_n) \\
z_{n+1} &= z_n + s_n - y_{n+1}.
\end{aligned}$$

signal and image processing community (see e.g. [16]–[20]). Under the assumptions that

- $\operatorname{rank}(L) = N$,
- Problem (19) admits a solution,
- $\operatorname{int}(\operatorname{dom} g) \cap L(\operatorname{dom} f) \neq \varnothing$ or $\operatorname{dom} g \cap \operatorname{int}\big(L(\operatorname{dom} f)\big) \neq \varnothing$,[3]

[3]$\operatorname{int} S$ denotes the interior of a set S. More general qualification conditions involving the relative interiors of the domain of g and $L(\operatorname{dom} f)$ can be obtained [3].

it can be shown that $(x_n)_{n\in\mathbb{N}}$ converges to a solution to the primal problem (19) and $(\gamma z_n)_{n\in\mathbb{N}}$ converges to a solution to the dual problem (21). A convergence rate analysis is conducted in [21].

It must be emphasized that ADMM is equivalent to the application of the Douglas-Rachford algorithm [22], [23], another famous algorithm in convex optimization, to the dual problem. Other primal-dual algorithms can be deduced from the Douglas-Rachford iteration [24] or an augmented Lagrangian approach [25].

Although ADMM was observed to have a good numerical performance in many problems, its applicability may be limited by the computation of x_n at each iteration $n \in \mathbb{N}$, which may be intricate due to the presence of matrix L, especially when this matrix is high-dimensional and has no simple structure. In addition, functions f and h cannot be dealt with separately, and so the smoothness of h is not exploited here in an explicit manner.

C. Methods based on a Forward-Backward approach

The methods which will be presented in this subsection are based on a forward-backward approach: they combine a gradient descent step (forward step) with a computation step involving a proximity operator. The latter computation corresponds to a kind of subgradient step performed in an implicit (or backward) manner [3]. A deeper justification of this terminology is provided by the theory of monotone operators [6] which allows to highlight the fact that a pair $(\widehat{x}, \widehat{v}) \in \mathbb{R}^N \times \mathbb{R}^K$ satisfying (22) is a zero of a sum of two maximally monotone operators. We will not go into details which can become rather technical, but we can mention that the algorithms presented in this section can then be viewed as offsprings of the forward-backward algorithm for finding such a zero [6]. One of the most popular primal-dual method within this class is given by Algorithm 2. Two step-sizes τ and σ and relaxation factors $(\lambda_n)_{n\in\mathbb{N}}$ are involved in this algorithms, which can be adjusted by the user so as to get the best convergence profile for a given application.

Algorithm 2 FB-based primal-dual algorithm

Set $x_0 \in \mathbb{R}^N$ and $v_0 \in \mathbb{R}^K$
Set $(\tau, \sigma) \in \,]0, +\infty[^2$
For $n = 0, 1, \ldots$
$\quad p_n = \mathrm{prox}_{\tau f}\big(x_n - \tau\big(\nabla h(x_n) + L^\top v_n\big)\big)$
$\quad q_n = \mathrm{prox}_{\sigma g^*}\big(v_n + \sigma L(2p_n - x_n)\big)$
$\quad$ Set $\lambda_n \in \,]0, +\infty[$
$\quad (x_{n+1}, v_{n+1}) = (x_n, v_n) + \lambda_n\big((p_n, q_n) - (x_n, v_n)\big)$.

Note that when $L = 0$ and $g^* = 0$ the basic form of the forward-backward algorithm is recovered, a popular example of which is the iterative soft-thresholding algorithm [26].

A rescaled variant of the primal-dual method (see Algorithm 3) is sometimes preferred, which can be deduced from the previous one by using Moreau's decomposition (8) and by making the variable changes: $q'_n \equiv q_n/\sigma$ and $v'_n \equiv v_n/\sigma$. Under this form, it can be seen that, when $N = K$, $L = \mathrm{Id}$, $h = 0$, and $\tau\sigma = 1$, the algorithm reduces to the Douglas-Rachford algorithm.

Algorithm 3 Rescaled variant of Algorithm 2

Set $x_0 \in \mathbb{R}^N$ and $v'_0 \in \mathbb{R}^K$
Set $(\tau, \sigma) \in \left]0, +\infty\right[^2$
For $n = 0, 1, \ldots$
$p_n = \operatorname{prox}_{\tau f}\big(x_n - \tau\big(\nabla h(x_n) + \sigma L^\top v'_n\big)\big)$
$q'_n = (\mathrm{Id} - \operatorname{prox}_{g/\sigma})\big(v'_n + L(2p_n - x_n)\big)$
Set $\lambda_n \in \left]0, +\infty\right[$
$(x_{n+1}, v'_{n+1}) = (x_n, v'_n) + \lambda_n\big((p_n, q'_n) - (x_n, v'_n)\big)$.

Also, by using the symmetry existing between the primal and the dual problems, another variant of Algorithm 2 can be obtained (see Algorithm 4) which is often encountered in the literature. When $L^\top L = \mu \mathrm{Id}$ with $\mu \in \left]0, +\infty\right[$, $h = 0$, $\tau\sigma\mu = 1$, and $\lambda_n \equiv 1$, Algorithm 4 reduces to ADMM by setting $\gamma = \sigma$, and $z_n \equiv v_n/\sigma$ in Algorithm 1.

Algorithm 4 Symmetric form of Algorithm 2

Set $x_0 \in \mathbb{R}^N$ and $v_0 \in \mathbb{R}^K$
Set $(\tau, \sigma) \in \left]0, +\infty\right[^2$
For $n = 0, 1, \ldots$
$q_n = \operatorname{prox}_{\sigma g^*}\big(v_n + \sigma L x_n\big)$
$p_n = \operatorname{prox}_{\tau f}\big(x_n - \tau\big(\nabla h(x_n) + L^\top(2q_n - v_n)\big)\big)$
Set $\lambda_n \in \left]0, +\infty\right[$
$(x_{n+1}, v_{n+1}) = (x_n, v_n) + \lambda_n\big((p_n, q_n) - (x_n, v_n)\big)$.

In terms of convergence guarantees, it can be shown that under the following sufficient conditions:

- $\tau^{-1} - \sigma\|L\|_{\mathrm{S}}^2 > \beta/2$ where $\|L\|_{\mathrm{S}}$ is the spectral norm of L,
- $(\lambda_n)_{n\in\mathbb{N}}$ a sequence in $\left]0, \delta\right[$ such that $\sum_{n\in\mathbb{N}} \lambda_n(\delta - \lambda_n) = +\infty$ where $\delta = 2 - \beta(\tau^{-1} - \sigma\|L\|_{\mathrm{S}}^2)^{-1}/2 \in [1, 2[$,
- Problem (19) admits a solution,
- $\operatorname{int}(\operatorname{dom} g) \cap L(\operatorname{dom} f) \neq \varnothing$ or $\operatorname{dom} g \cap \operatorname{int}\big(L(\operatorname{dom} f)\big) \neq \varnothing$,

the sequences $(x_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ generated by Algorithms 2 and 4 are such that the former one converges to a solution to the primal problem (19) and the latter one converges to a solution to the dual problem (21). These results were established in [27], as well as for a more general version of this algorithm in [28]. Algorithm 2 also constitutes an extension of [29]–[31] (designated by some authors as PDHG, Primal-Dual Hybrid Gradient). Preconditioned or adaptive versions of this algorithm were proposed in [32]–[35] which may accelerate its convergence. Convergence rate results were also recently derived in [36].

Another primal-dual method (see Algorithm 5) was proposed in [37], [38] which also results from a forward-backward approach. This algorithm is restricted to the case when $f = 0$ in Problem (19).

Algorithm 5 Second FB-based primal-dual algorithm

Set $x_0 \in \mathbb{R}^N$ and $v_0 \in \mathbb{R}^K$

Set $(\tau, \sigma) \in \left]0, +\infty\right[^2$

For $n = 0, 1, \ldots$

$s_n = x_n - \tau \nabla h(x_n)$

$y_n = s_n - \tau L^\top v_n$

$q_n = \mathrm{prox}_{\sigma g^*}\Big(v_n + \sigma L y_n\Big)$

$p_n = s_n - \tau L^\top q_n$

Set $\lambda_n \in \left]0, +\infty\right[$

$(x_{n+1}, v_{n+1}) = (x_n, v_n) + \lambda_n\big((p_n, q_n) - (x_n, v_n)\big)$.

Then, under the assumptions that

- $\tau\sigma\|L\|_{\mathrm{S}}^2 < 1$ and $\tau < 2/\beta$,
- $(\lambda_n)_{n\in\mathbb{N}}$ a sequence in $]0, 1]$ such that $\inf_{n\in\mathbb{N}} \lambda_n > 0$,
- Problem (19) admits a solution,
- $\mathrm{int}\,(\mathrm{dom}\, g) \cap \mathrm{ran}\,(L) \neq \varnothing$,

the sequence $(x_n)_{n\in\mathbb{N}}$ converges to a solution to the primal problem (19) (where $f = 0$) and $(v_n)_{n\in\mathbb{N}}$ converges to a solution to the dual problem (21). The conditions on the step-sizes τ and σ are thus less restrictive than for Algorithm 2.

Note also that the dual forward-backward approach that was proposed in [39] for solving (19) in the specific case when $h = \|\cdot - r\|^2/2$ with $r \in \mathbb{R}^N$ belongs to the class of primal-dual forward-backward approaches.

It must be emphasized that Algorithms 2-5 present two interesting features which are very useful in practice. At first, they allow to deal with the functions involved in the optimization problem at hand either through their proximity operator or through their gradient. Indeed, for some functions, especially non differentiable or non finite ones, the proximity operator can be a very powerful tool but, for some smooth functions (e.g. the Poisson-Gauss neg-log-likelihood [40]) the gradient may be easier to handle. Secondly, these algorithms do not require to invert any matrix, but only to apply L and its adjoint. This advantage is of main interest when large-size problems have to be solved for which the inverse of L does not exist or it has a no tractable expression.

D. Methods based on a Forward-Backward-Forward approach

Primal-dual methods based on a forward-backward-forward approach were among the first primal-dual proximal methods proposed in the optimization literature, inspired from the seminal work in [41]. They were first developed in the case when $h = 0$ [42], then extended to more general scenarios in [43] (see also [44], [45] for further refinements).

Algorithm 6 FBF-based primal-dual algorithm

Set $x_0 \in \mathbb{R}^N$ and $v_0 \in \mathbb{R}^K$

For $n = 0, 1, \ldots$

Set $\gamma_n \in \left]0, +\infty\right[$

$y_{1,n} = x_n - \gamma_n\big(\nabla h(x_n) + L^\top v_n\big)$

$y_{2,n} = v_n + \gamma_n L x_n$

$p_{1,n} = \operatorname{prox}_{\gamma_n f} y_{1,n}, p_{2,n} = \operatorname{prox}_{\gamma_n g^*} y_{2,n}$

$q_{1,n} = p_{1,n} - \gamma_n\big(\nabla h(p_{1,n}) + L^\top p_{2,n}\big)$

$q_{2,n} = p_{2,n} + \gamma_n L p_{1,n}$

$(x_{n+1}, v_{n+1}) = (x_n - y_{1,n} + q_{1,n}, v_n - y_{2,n} + q_{2,n})$.

The convergence of the sequence $(x_n, v_n)_{n\in\mathbb{N}}$ generated by Algorithm 6 to a pair of primal-dual solutions is guaranteed under the following assumptions:

- $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence in $[\varepsilon, (1-\epsilon)/\mu]$ where $\varepsilon \in]0, 1/(1+\mu)[$ and $\mu = \beta + \|L\|_{\mathrm{S}}$,
- Problem (19) admits a solution,
- $\operatorname{int}(\operatorname{dom} g) \cap L(\operatorname{dom} f) \neq \varnothing$ or $\operatorname{dom} g \cap \operatorname{int}\big(L(\operatorname{dom} f)\big) \neq \varnothing$,

Algorithm 6 is often refered to as the M+LFBF (Monotone+Lipschitz Forward Backward Forward) algorithm. It enjoys the same advantages as FB-based primal-dual algorithms we have seen before. It however makes it possible to compute the proximity operators of scaled versions of functions f and g^* in parallel. In addition, the choice of its parameters in order to satisfy convergence conditions may appear more intuitive than for Algorithms 2-4. With respect to FB-based algorithms, an additional forward step however needs to be performed. This may lead to a slower convergence if, for example, the computational cost of the gradient is high and an iteration of a FB-based algorithm is at least as efficient as an iteration of Algorithm 6.

E. A projection-based primal-dual algorithm

Another primal-dual algorithm was recently proposed in [46] which relies on iterative projections onto half-spaces including the set of Kuhn-Tucker points (see Algorithm 7). We have then the following convergence result: assume that

- $(\gamma_n)_{n\in\mathbb{N}}$ and $(\mu_n)_{n\in\mathbb{N}}$ are sequences such that $\inf_{n\in\mathbb{N}} \gamma_n > 0$, $\sup_{n\in\mathbb{N}} \gamma_n < +\infty$, $\inf_{n\in\mathbb{N}} \mu_n > 0$, $\sup_{n\in\mathbb{N}} \mu_n < +\infty$,
- $(\lambda_n)_{n\in\mathbb{N}}$ a sequence in $\mathbb{R}$ such that $\inf_{n\in\mathbb{N}} \lambda_n > 0$ and $\sup_{n\in\mathbb{N}} \lambda_n < 2$,
- Problem (19) admits a solution,
- $\operatorname{int}(\operatorname{dom} g) \cap L(\operatorname{dom} f) \neq \varnothing$ or $\operatorname{dom} g \cap \operatorname{int}\big(L(\operatorname{dom} f)\big) \neq \varnothing$,

then, either the algorithm terminates in a finite number of iterations at a pair of primal-dual solutions $(\widehat{x}, \widehat{v})$, or it generates a sequence $(x_n, v_n)_{n\in\mathbb{N}}$ converging to such a point. Although few numerical experiments have been

performed with this algorithm, one of its potential advantages is that it introduces few constraints on the choice of the parameters γ_n, μ_n and λ_n at iteration n. Nonetheless, the use of this algorithm does not allow us to exploit the fact that h is a differentiable function.

Algorithm 7 Projection-based primal-dual algorithm

Set $x_0 \in \mathbb{R}^N$ and $v_0 \in \mathbb{R}^K$

For $n = 0, 1, \ldots$

- Set $(\gamma_n, \mu_n) \in \left]0, +\infty\right[$
- $a_n = \operatorname{prox}_{\gamma_n (f+h)}(x_n - \gamma_n L^\top v_n)$
- $l_n = L x_n$
- $b_n = \operatorname{prox}_{\mu_n g}(l_n + \mu_n v_n)$
- $s_n = \gamma_n^{-1}(x_n - a_n) + \mu_n^{-1} L^\top (l_n - b_n)$
- $t_n = b_n - L a_n$
- $\tau_n = \|s_n\|^2 + \|t_n\|^2$
- if $\tau_n = 0$
  - $\widehat{x} = a_n$
  - $\widehat{v} = v_n + \mu_n^{-1}(l_n - b_n)$
  - return
- else
  - Set $\lambda_n \in \left]0, +\infty\right[$
  - $\theta_n = \lambda_n (\gamma_n^{-1} \|x_n - a_n\|^2 + \mu_n^{-1} \|l_n - b_n\|^2)/\tau_n$
  - $x_{n+1} = x_n - \theta_n s_n$
  - $v_{n+1} = v_n - \theta_n t_n$.

F. Extensions

More generally, one may be interested in more challenging convex optimization problems of the form:

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; f(x) + \sum_{m=1}^{M} (g_m \,\square\, \ell_m)(L_m x) + h(x), \tag{24}$$

where $f \in \Gamma_0(\mathbb{R}^N)$, $h \in \Gamma_0(\mathbb{R}^N)$, and, for every $m \in \{1, \ldots, M\}$, $g_m \in \Gamma_0(\mathbb{R}^{K_m})$, $\ell_m \in \Gamma_0(\mathbb{R}^{K_m})$, and $L_m \in \mathbb{R}^{K_m \times N}$. The dual problem then reads

$$\underset{v_1 \in \mathbb{R}^{K_1}, \ldots, v_m \in \mathbb{R}^{K_M}}{\text{minimize}} \;\; \big(f^* \,\square\, h^*\big)\bigg(- \sum_{m=1}^{M} L_m^\top v_m \bigg) + \sum_{m=1}^{M} \big(g_m^*(v_m) + \ell_m^*(v_m)\big). \tag{25}$$

Some comments can be made on this general formulation. At first, one of its benefits is to split an original objective function in a sum of a number of simpler terms. Such splitting strategy is often the key of an efficient resolution of difficult optimization problems. For example, the proximity operator of the global objective function may be quite involved, while the proximity operators of the individual functions may have an explicit form. A second

point is that we have now introduced in the formulation, additional functions $(\ell_m)_{1\leq m\leq M}$. These functions may be useful in some models [47], but they present also the conceptual advantage to make the primal problem and its dual form quite symmetric. For instance, this fact accounts for the symmetric roles played by Algorithms 2 and 4. An assumption which is commonly adopted is to assume that, whereas h is Lipschitz differentiable, the functions $(\ell_m)_{1\leq m\leq M}$ are strongly convex, i.e. their conjugates are Lipschitz differentiable. A last point to be emphasized is that, such split forms are amenable to efficient parallel implementations. Using parallelized versions of primal-dual algorithms on multi-core architectures may render these methods even more successful for dealing with large-scale problems.

HOW TO PARALLELIZE PRIMAL-DUAL METHODS ?

Two main ideas can be used in order to put a primal-dual method under a parallel form.

Let us first consider the following simplified form of Problem (24):

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \quad \sum_{m=1}^{M} g_m(L_m x). \tag{26}$$

A possibility consists of reformulating this problem in a higher-dimensional space as

$$\underset{y_1\in\mathbb{R}^{K_1},\ldots,y_M\in\mathbb{R}^{K_M}}{\text{minimize}} \quad f(\boldsymbol{y}) + \sum_{m=1}^{M} g_m(y_m), \tag{27}$$

where $\boldsymbol{y} = [y_1^\top,\ldots,y_M^\top]^\top \in \mathbb{R}^K$ with $K = K_1 + \cdots + K_M$, and f is the indicator function of $\mathrm{ran}\,(\boldsymbol{L})$, where $\boldsymbol{L} = [L_1^\top,\ldots,L_M^\top]^\top \in \mathbb{R}^{K\times N}$. Function f serves to enforce the constraint: $(\forall m\in\{1,\ldots,M\})\ y_m = L_m x$. By defining the separable function $g\colon \boldsymbol{y}\mapsto \sum_{m=1}^{M} g_m(y_m)$, we are thus led to the minimization of $f+g$ in the space $\mathbb{R}^K$. This optimization can be performed by the algorithms described in Sections III-B-III-E. The proximity operator of f reduces to the linear projection onto $\mathrm{ran}\,(\boldsymbol{L})$, whereas the separability of g ensures that its proximity operator can be obtained by computing in parallel the proximity operators of the function $(g_m)_{1\leq m\leq M}$. Note that, when $L_1 = \ldots = L_M = \mathrm{Id}$, we recover a consensus-based approach that we have already discussed. This technique can be used to derive parallel forms of the Douglas-Rachford algorithm, namely the Parallel ProXimal Algorithm (PPXA) [48] and PPXA+ [49], as well as parallel versions of ADMM (Simultaneous Direction Method of Multipliers or SDMM) [50].

The second approach is even more direct since it requires no projection onto $\mathrm{ran}\,(\boldsymbol{L})$. For simplicity, let us consider the following instance of Problem (24):

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \quad f(x) + \sum_{m=1}^{M} g_m(L_m x) + h(x). \tag{28}$$

By defining the function g and the matrix $\boldsymbol{L}$ as in the previous approach, the problem can be recast as

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \quad f(x) + g(\boldsymbol{L}x) + h(x). \tag{29}$$

Once again, under appropriate assumptions on the involved functions, this formulation allows us to employ the algorithms proposed in Sections III-B-III-E and we still have the ability to compute the proximity operator of g in a parallel manner.

IV. Discrete optimization algorithms

A. Background on discrete optimization

As already mentioned in the introduction, another common class of problems in signal processing and image analysis are discrete optimization problems, for which primal-dual algorithms also play an important role. Problems of this type are often stated as *integer linear programs* (ILPs), which will thus be the general setting that we

will explore hereafter. Among the problems encountered in practice, many of them involve ILPs that are NP-hard to solve. In such cases, a principled approach for finding an approximate solution is through the use of convex *relaxations* (see framebox), where the original NP-hard problem is approximated with a surrogate one (the so-called relaxed problem), which is convex and thus much easier to solve. The premise is the following: to the extent that the surrogate problem provides a reasonably good approximation to the original optimization task, one can expect to obtain an approximately optimal solution for the latter by essentially making use of or solving the former.

RELAXATIONS AND DISCRETE OPTIMIZATION

Relaxations are very useful for solving approximately discrete optimization problems. Formally, given a problem

$$(\mathcal{P}) : \underset{x \in C}{\text{minimize}} \;\; f(x)$$

where C is a subset of $\mathbb{R}^N$, we say that

$$(\mathcal{P}') : \underset{x \in C'}{\text{minimize}} \;\; f'(x)$$

with $C' \subset \mathbb{R}^N$ is a relaxation of $(\mathcal{P})$ if and only if (*i*) $C \subset C'$, and (*ii*) $(\forall x \in C')\; f(x) \geq f'(x)$.

For instance, let us consider the integer linear program defined by $(\forall x \in \mathbb{R}^N)\; f(x) = c^\top x$ and $C = S \cap \mathbb{Z}^N$, where $c \in \mathbb{R}^N \setminus \{0\}$ and S is a nonempty closed polyhedron defined as

$$S = \left\{ x \in \mathbb{R}^N \;\middle|\; Lx \geq b \right\}$$

with $L \in \mathbb{R}^{K \times N}$ and $b \in \mathbb{R}^K$. One possible linear programming relaxation of $(\mathcal{P})$ is obtained by setting $f' = f$ and $C' = S$, which is typically much easier than $(\mathcal{P})$ (which is generally NP-hard). The quality of $(\mathcal{P}')$ is quantified by its so-called integrality gap defined as $\frac{\inf f(C)}{\inf f'(C')} \geq 1$ (provided that $-\infty < \inf f'(C') \neq 0$).

Hence, for approximation algorithms, LP relaxations are not all of equal value. If

$$(\mathcal{P}'') : \underset{x \in C''}{\text{minimize}} \;\; c^\top x$$

is another relaxation of $(\mathcal{P})$ with $C'' \subset C'$, then relaxation $(\mathcal{P}'')$ is tighter. Interestingly, $(\mathcal{P})$ always has a tight LP relaxation (with integrality gap 1) given by $C'' = \mathrm{conv}(S \cap \mathbb{Z}^N)$, where $\mathrm{conv}(C)$ is the convex hull polyhedron of C. Note, however, that if $(\mathcal{P})$ is NP-hard, polyhedron $\mathrm{conv}(S \cap \mathbb{Z}^N)$ will involve exponentially many inequalities.

The relaxations in all of the previous examples involve expanding the original feasible set. But, as mentioned, we can also derive relaxations by modifying the original objective function. For instance, in so-called submodular relaxations [51], [52], one uses as new objective a maximum submodular function that lower bounds the original objective.

The type of relaxations that are typically preferred in large scale discrete optimization are based on linear programming, involving the minimization of a linear function subject to linear inequality constraints. These can be naturally obtained by simply relaxing the integrality constraints of the corresponding ILP, and their use is often dictated by the need of maintaining a reasonable computational cost. More powerful convex relaxations do exist in many cases, but these may become intractable as the number of variables grows larger, especially for Semidefinite Programming (SDP) or Second-Order Cone Programming (SOCP) relaxations.

Based on the above observations, in the following we aim to present some very general primal-dual optimization strategies that can be used in this context, focusing a lot on their underlying principles, which are based on two powerful techniques the so-called *primal-dual schema* and *dual decomposition*. As we shall see, in order to estimate an approximate solution to the ILP, both approaches make heavy use of the dual of the underlying LP relaxation. But their strategies for doing so is quite different: the second one essentially aims at solving this dual LP (and

then converting the fractional solution into an integral one, trying not to increase the cost too much in the process), whereas the other simply uses it in the design of the algorithm.

B. The primal-dual schema

The primal-dual schema is a technique well known in the combinatorial optimization community that has its origins in LP duality theory. It is worth noting that it started as an exact method for solving linear programs. As such, it had initially been used in deriving exact polynomial-time algorithms for many cornerstone problems in combinatorial optimization that have a tight LP relaxation. Its first use probably goes back to Edmond's famous Blossom algorithm for constructing maximum matchings on graphs, but it had been also applied to many other combinatorial problems including max-flow (e.g., Ford and Fulkerson's augmenting path-based techniques for max-flow can essentially be understood in terms of this schema), shortest path, minimum branching, and minimum spanning tree [53]. In all of these cases, the primal-dual schema is driven by the fact that optimal LP solutions should satisfy the *complementary slackness conditions* (see (17), (18)). Starting with an initial primal-dual pair of feasible solutions, it therefore iteratively steers them towards satisfying these complementary slackness conditions (by trying at each step to minimize their total violation). Once this is achieved, both solutions (the primal and the dual) are guaranteed to be optimal. Moreover, since the primal is always chosen to be updated integrally during the iterations, it is ensured that an integral optimal solution is obtained at the end. A notable feature of the primal-dual method is that it often reduces the original LP, which is a weighted optimization problem, to a series of purely combinatorial unweighted ones (related to minimizing the violation of complementary slackness conditions at each step).

Interestingly, today the primal-dual schema is no longer used for providing exact algorithms. Instead, its main use concerns deriving approximation algorithms to NP-hard combinatorial problems, for which it has proved to be a very powerful and widely applicable tool. As such, it has been applied to many NP-hard problems up to now, including set-cover, Steiner-network, scheduling, Steiner tree, feedback vertex set, facility location, to mention only a few [54], [55]. With regard to problems from the domains of computer vision and image analysis, the primal-dual schema has been introduced recently in [56], [57], and has been used for modeling a broad class of tasks from these fields.

It should be noted that for NP-hard problems an integral solution is no longer guaranteed to satisfy the complementary slackness conditions (since the LP-relaxation is not exact). How could it then be possible to apply this schema to such problems? It turns out that the answer to this question consists of using an appropriate relaxation of the above conditions. To understand exactly how we need to proceed in this case, let us consider the following integer linear program

$$\text{Primal-ILP:} \quad \underset{x\in\mathbb{R}^N}{\text{minimize}} \;\; c^\top x$$
$$\text{s.t.} \quad Lx \geq b, \quad x \in \mathcal{N} \subset \mathbb{N}^N,$$

where $L = (L^{(i,j)})_{1\leq i\leq K, 1\leq j\leq N}$ represents a matrix of size $K \times N$, and $b = (b^{(i)})_{1\leq i\leq K}$, $c = (c^{(j)})_{1\leq j\leq N}$ are column vectors of size K and N, respectively. We wish to compute an optimal solution to Primal-ILP, but, due to the integrality constraints $x \in \mathcal{N}$, this is assumed to be NP-hard, and so we can only estimate an approximate solution. To achieve that, we will first need to relax the integrality constraints, thus giving rise to the relaxed primal problem in (14) as well as it dual (16). A primal-dual algorithm attempts to compute an approximate solution to Primal-ILP by relying on the following principle (see framebox for an explanation):

Primal-dual principle in the discrete case: *Let $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^K$ be integral-primal and dual feasible solutions (i.e. $x \in \mathcal{N}$ and $Lx \geq b$, and $y \in [0, +\infty[^K$ and $L^\top y \leq c$). Assume that there exists $\nu \in [1, +\infty[$ such that*

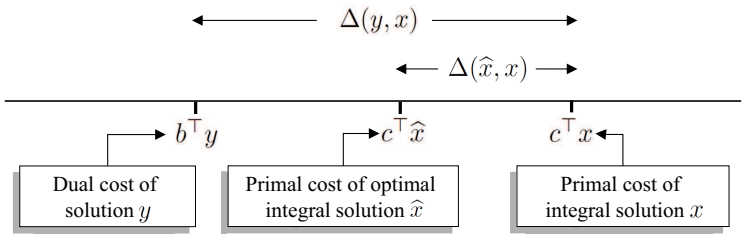
$$c^\top x \leq \nu\, b^\top y. \tag{30}$$

Then, x is a ν-approximation to an unknown optimal integral solution $\widehat{x}$, i.e.

$$c^\top \widehat{x} \leq c^\top x \leq \nu\, c^\top \widehat{x}. \tag{31}$$

PRIMAL-DUAL PRINCIPLE IN THE DISCRETE CASE

Essentially, this principle relies on the fact that the sequence of optimal costs of problems Dual-LP, Primal-LP, and Primal-ILP is increasing.



Specifically, by weak LP duality, the cost $c^\top \widehat{x}$ (of an unknown optimal integral solution $\widehat{x}$) is guaranteed to lie between costs $b^\top y$ and $c^\top x$ of any pair of integral-primal and dual feasible solutions (x, y). Therefore, if the gap $\Delta(y, x)$ between the costs of y and x is small (e.g., it holds $c^\top x \leq \nu\, b^\top y$), the same will be true for the gap $\Delta(\widehat{x}, x)$ between the costs of $\widehat{x}$ and x (i.e., $c^\top x \leq \nu\, c^\top \widehat{x}$), thus proving that x is a ν-approximation to optimal solution $\widehat{x}$.

Although the above principle lies at the heart of many primal-dual techniques (i.e., in one way or another, primal-dual methods often try to fulfill the assumptions imposed by this principle), it does not directly specify how to estimate a primal-dual pair of solutions (x, y) that satisfies these assumptions. This is where the so-called *relaxed complementary slackness conditions* come into play, as they typically provide an alternative and more convenient (from an algorithmic viewpoint) way for generating such a pair of solutions. These conditions generalize the complementary slackness conditions associated with an arbitrary pair of primal-dual linear programs (see Section II-D). The latter conditions apply only in cases when there is no duality gap, like between Primal-LP and Dual-LP, but they are not applicable to cases like Primal-ILP and Dual-LP, where a duality gap exists as a result of the integrality constraint imposed on variable x. As in the exact case, two types of relaxed complementary slackness conditions exist, depending on whether the primal or dual variables are checked for being zero.

Relaxed Primal Complementary Slackness Conditions with relaxation factor $\nu_{\text{primal}} \leq 1$: for a given $x = (x^{(j)})_{1\leq j\leq N} \in \mathbb{R}^N$, $y = (y^{(i)})_{1\leq i\leq K} \in \mathbb{R}^K$ is such that

$$(\forall j \in J_x) \qquad \nu_{\text{primal}}\, c^{(j)} \leq \sum_{i=1}^{K} L^{(i,j)} y^{(i)} \leq c^{(j)} \tag{32}$$

where $J_x = \big\{j \in \{1,\ldots,N\} \;\big|\; x^{(j)} > 0\big\}$.

Relaxed Dual Complementary Slackness Conditions with relaxation factor $\nu_{\text{dual}} \geq 1$: for a given $y = (y^{(i)})_{1\leq i\leq K} \in \mathbb{R}^K$, $x = (x^{(j)})_{1\leq j\leq N} \in \mathbb{R}^N$ is such that

$$(\forall i \in I_y) \qquad b^{(i)} \leq \sum_{j=1}^{N} L^{(i,j)} x^{(j)} \leq \nu_{\text{dual}}\, b^{(i)} \tag{33}$$

where $I_y = \big\{i \in \{1,\ldots,K\} \;\big|\; y^{(i)} > 0\big\}$.

When both $\nu_{\text{primal}} = 1$ and $\nu_{\text{dual}} = 1$, we recover the exact complementary slackness conditions in (17) and (18). The use of the above conditions in the context of a primal-dual approximation algorithm becomes clear by the following result:

If $x = (x^{(j)})_{1\leq j\leq N}$ and $y = (y^{(i)})_{1\leq i\leq K}$ are feasible with respect to Primal-ILP *and* Dual-LP *respectively, and satisfy the relaxed complementary slackness conditions* (32) *and* (33)*, then the pair (x, y) satisfies the primal-dual principle in the discrete case with $\nu = \frac{\nu_{\text{dual}}}{\nu_{\text{primal}}}$. Therefore, x is a ν-approximate solution to* Primal-ILP.

This result simply follows from the inequalities

$$\begin{aligned} c^\top x = \sum_{j=1}^{N} c^{(j)} x^{(j)} &\overset{(32)}{\leq} \sum_{j=1}^{N} \Big(\frac{1}{\nu_{\text{primal}}} \sum_{i=1}^{K} L^{(i,j)} y^{(i)}\Big) x^{(j)} = \frac{1}{\nu_{\text{primal}}} \sum_{i=1}^{N} \Big(\sum_{j=1}^{K} L^{(i,j)} x^{(j)}\Big) y^{(i)} \\ &\overset{(33)}{\leq} \frac{\nu_{\text{dual}}}{\nu_{\text{primal}}} \sum_{i=1}^{m} b^{(i)} y^{(i)} = \frac{\nu_{\text{dual}}}{\nu_{\text{primal}}} b^\top y. \end{aligned} \tag{34}$$

Based on the above result, iterative schemes can be devised yielding a primal-dual ν-approximation algorithm. For example, we can employ the following algorithm:

Primal-dual schema: *Generate a sequence $(x_n, y_n)_{n\in\mathbb{N}}$ of elements of $\mathbb{R}^N \times \mathbb{R}^K$ as follows*

Set $\nu_{\text{primal}} \leq 1$ and $\nu_{\text{dual}} \geq 1$

Set $y_0 \in [0, +\infty[^K$ such that $L^\top y_0 \leq c$

For $n = 0, 1, \ldots$

Find $x_n \in \big\{x \in \mathcal{N} \;\big|\; Lx \geq b\big\}$ minimizing

$\quad \sum_{i\in I_{y_n}} q^{(i)}$ s.t. $(\forall i \in I_{y_n}) \quad \sum_{j=1}^{N} L^{(i,j)} x^{(j)} \leq \nu_{\text{dual}}\, b^{(i)} + q^{(i)},\ q^{(i)} \geq 0$

Find $y_{n+1} \in \big\{y \in [0, +\infty[^K \;\big|\; L^\top y \leq c\big\}$ minimizing

$\quad \sum_{j\in J_{x_n}} r^{(j)}$ s.t. $(\forall j \in J_{x_n}) \quad \sum_{i=1}^{K} L^{(i,j)} y^{(i)} + r^{(j)} \geq \nu_{\text{primal}}\, c^{(j)},\ r^{(j)} \geq 0.$ (35)

Note that, in this scheme, primal solutions are always updated integrally. Also, note that, when applying the primal-dual schema, different implementation strategies are possible. The strategy described in (35) is to maintain feasible primal-dual solutions (x_n, y_n) at iteration n, and iteratively improve how tightly the (primal or dual)

complementary slackness conditions get satisfied. This is performed through the introduction of slackness variables $(q^{(i)})_{i\in I_{y_n}}$ and $(r^{(j)})_{j\in J_{x_n}}$ the sums of which measure the degrees of violation of each relaxed slackness condition and have thus to be minimized. Alternatively, for example, we can opt to maintain solutions (x_n, y_n) that satisfy the relaxed complementary slackness conditions, but may be infeasible, and iteratively improve the feasibility of the generated solutions. For instance, if we start with a feasible dual solution but with an infeasible primal solution, such a scheme would result into improving the feasibility of the primal solution, as well as the optimality of the dual solution at each iteration, ensuring that a feasible primal solution is obtained at the end. No matter which one of the above two strategies we choose to follow, the end result will be to gradually bring the primal and dual costs $c^\top x_n$ and $b^\top y_n$ closer and closer together so that asymptotically the primal-dual principle gets satisfied with the desired approximation factor. Essentially, at each iteration, through the coupling by the complementary slackness conditions the current primal solution is used to improve the dual, and vice versa.

Three remarks are worth making at this point: the first one relates to the fact that the two processes, i.e. the primal and the dual, make only local improvements to each other. Yet, in the end they manage to yield a result that is almost globally optimal. The second point to emphasize is that, for computing this approximately optimal result, the algorithm requires no solution to the Primal-LP or Dual-LP to be computed, which are replaced by simpler optimization problems. This is an important advantage from a computational standpoint since, for large scale problems, solving these relaxations can often be quite costly. In fact, in most cases where we apply the primal-dual schema, purely combinatorial algorithms can be obtained that contain no sign of linear programming in the end. A last point to be noticed is that these algorithms require appropriate choices of the relaxation factors ν_{primal} and ν_{dual}, which are often application-guided.

C. Primal-dual schema applied to the set cover problem

To provide a simple illustration of the primal-dual schema, let us consider the problem of set-cover, which is known to be NP-hard. In this problem, we are given as input a finite set $\mathcal{V}$ of K elements $(v^{(i)})_{1\le i\le K}$, a collection of (non disjoint) subsets $\mathcal{S} = \{S_j\}_{1\le j\le N}$ where, for every $j \in \{1,\dots,N\}$, $S_j \subset \mathcal{V}$, and $\bigcup_{j=1}^N S_j = \mathcal{V}$. Let $\varphi\colon \mathcal{S} \to \mathbb{R}$ be a function that assigns a cost $\varphi(S_j)$ for each subset S_j. The goal is to find a set cover (i.e. a subcollection of $\mathcal{S}$ that covers all elements of $\mathcal{V}$) that has minimum cost. The above problem can be expressed as the following ILP:

$$\underset{x=(x^{(j)})_{1\le j\le N}}{\text{minimize}} \quad \sum_{j=1}^{N} \varphi(S_j)\, x^{(j)} \tag{36}$$

$$\text{s.t. } (\forall i \in \{1,\dots,K\}) \quad \sum_{\substack{j\in\{1,\dots,N\}\\ v^{(i)}\in S_j}} x^{(j)} \ge 1, \quad x \in \{0,1\}^N, \tag{37}$$

where indicator variables $(x^{(j)})_{1\le j\le N}$ are used for determining if a set in $\mathcal{S}$ has been included in the set cover or not, and Constraints (37) ensure that each one of the elements of $\mathcal{V}$ is contained in at least one of the sets that were chosen for participating to the set cover.

An LP-relaxation for this problem is obtained by simply replacing the Boolean constraint with the constraint $x \in [0, +\infty[^N$. The dual of this LP relaxation is given by the following linear program:

$$\underset{y=(y^{(i)})_{1\leq i\leq K}\in[0,+\infty[^K}{\text{maximize}} \quad \sum_{i=1}^{K} y^{(i)} \tag{38}$$

$$\text{s.t.} \quad (\forall j \in \{1,\ldots,N\}) \quad \sum_{\substack{i\in\{1,\ldots,K\}\\ \upsilon^{(i)}\in S_j}} y^{(i)} \leq \varphi(S_j). \tag{39}$$

Let us denote by $F_{\max}$ the maximum frequency of an element in $\mathcal{V}$, where by the term *frequency* we mean the number of sets this element belongs to. In this case, we will use the primal-dual schema to derive an $F_{\max}$-approximation algorithm by choosing $\nu_{\text{primal}} = 1$, $\nu_{\text{dual}} = F_{\max}$. This results into the following complementary slackness conditions, which we will need to satisfy:

Primal Complementary Slackness Conditions

$$(\forall j \in \{1,\ldots,N\}) \text{ if } x^{(j)} > 0 \text{ then } \sum_{\substack{i\in\{1,\ldots,K\}\\ \upsilon^{(i)}\in S_j}} y^{(i)} = \varphi(S_j) \tag{40}$$

Relaxed Dual Complementary Slackness Conditions (with relaxation factor $F_{\max}$)

$$(\forall i \in \{1,\ldots,K\}) \text{ if } y^{(i)} > 0 \text{ then } \sum_{\substack{j\in\{1,\ldots,N\}\\ \upsilon^{(i)}\in S_j}} x^{(j)} \leq F_{\max}. \tag{41}$$

A set S_j with $j \in \{1,\ldots,N\}$ for which $\sum_{\substack{i\in\{1,\ldots,K\}\\ \upsilon^{(i)}\in S_j}} y^{(i)} = \varphi(S_j)$ will be called *packed*. Based on this definition, and given that the primal variables $(x^{(j)})_{1\leq j\leq N}$ are always kept integral (i.e., either 0 or 1) during the primal-dual schema, Conditions (40) basically say that only packed sets can be included in the set cover (note that overpacked sets are already forbidden by feasibility constraints (39)). Similarly, Conditions (41) require that an element $\upsilon^{(i)}$ with $i \in \{1,\ldots,K\}$ associated with a nonzero dual variable $y^{(i)}$ should not be covered more than $F_{\max}$ times, which is, of course, trivially satisfied given that $F_{\max}$ represents the maximum frequency of any element in $\mathcal{V}$.

Algorithm 8 Primal-dual schema for set-cover.

Set $x_0 \leftarrow 0, y_0 \leftarrow 0$

Declare all elements in $\mathcal{V}$ as uncovered

While $\mathcal{V}$ contains uncovered elements

- Select an uncovered element $\upsilon^{(i)}$ with $i \in \{1,\ldots,K\}$ and increase $y^{(i)}$ until some set becomes packed
- For every packed set S_j with $j \in \{1,\ldots,N\}$, set $x^{(j)} \leftarrow 1$
  - (include all the sets that are packed in the cover)
- Declare all the elements belonging to at least one set S_j with $x^{(j)} = 1$ as covered.

Based on the above observations, the iterative method whose pseudocode is shown in Algorithm 8 emerges naturally as a simple variant of (35). Upon its termination, both x and y will be feasible given that there will be no

uncovered elements and no set that violates (39). Furthermore, given that the final pair (x, y) satisfies the relaxed complementary slackness conditions with $\nu_{\text{primal}} = 1$, $\nu_{\text{dual}} = F_{\max}$, the set cover defined by x will provide an $F_{\max}$-approximate solution.

D. Dual decomposition

We will next examine a different approach for discrete optimization, which is based on the principle of dual decomposition [58]–[60]. The core idea behind this principle essentially follows a divide and conquer strategy: that is, given a difficult or high-dimensional optimization problem, we decompose it into smaller easy-to-handle subproblems and then extract an overall solution by cleverly combining the solutions from these subproblems.

To explain this technique, we will consider the general problem of minimizing the energy of a discrete Markov Random Field (MRF), which is a ubiquitous problem in the fields of computer vision and image analysis (applied with great success on a wide variety of tasks from these domains) [61]. This problem involves a graph G with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$ (i.e., $G = (\mathcal{V}, \mathcal{E})$) plus a finite label set $\mathcal{L}$. The goal is to find a labeling $z = (z^{(p)})_{p\in\mathcal{V}} \in \mathcal{L}^{|\mathcal{V}|}$ for the graph vertices that has minimum cost, that is

$$\underset{z\in\mathcal{L}^{|\mathcal{V}|}}{\text{minimize}} \;\; \sum_{p\in\mathcal{V}} \varphi_p(z^{(p)}) + \sum_{e\in\mathcal{E}} \varphi_e(\mathsf{z}^{(e)}) \tag{42}$$

where, for every $p \in \mathcal{V}$ and $e \in \mathcal{E}$, $\varphi_p \colon \mathcal{L} \to \,]-\infty, +\infty[$ and $\varphi_e \colon \mathcal{L}^2 \to \,]-\infty, +\infty[$ represent the unary and pairwise costs (also known connectively as MRF potentials $\varphi = \{\{\varphi_p\}_{p\in\mathcal{V}}, \{\varphi_e\}_{e\in\mathcal{E}}\}$), and $\mathsf{z}^{(e)}$ denotes the pair of components of z defined by the variables corresponding to vertices connected by e (i.e., $\mathsf{z}^{(e)} = (z^{(p)}, z^{(q)})$ for $e = (p, q) \in \mathcal{E}$).

The above problem is NP-hard, and much of the recent work on MRF optimization revolves around the following equivalent ILP formulation of (42) [62], which is the one that we will also use here:

$$\underset{x\in C_G}{\text{minimize}} \;\; f(x;\varphi) = \sum_{p\in\mathcal{V},\, z^{(p)}\in\mathcal{L}} \varphi_p(z^{(p)})\, x_p(z^{(p)}) + \sum_{e\in\mathcal{E},\, \mathsf{z}^{(e)}\in\mathcal{L}^2} \varphi_e(\mathsf{z}^{(e)})\, \mathsf{x}_e(\mathsf{z}^{(e)}), \tag{43}$$

where the set C_G is defined for any graph $G = (\mathcal{V}, \mathcal{E})$ as follows

$$C_G = \left\{ x \in \{0,1\}^N \;\middle|\; \begin{array}{ll} (\forall p \in \mathcal{V}) & \sum_{z^{(p)}\in\mathcal{L}} x_p(z^{(p)}) = 1 \\ (\forall e = (p,q) \in \mathcal{E})(\forall z^{(q)} \in \mathcal{L}) & \sum_{\mathsf{z}^{(e)}\in\mathcal{L}\times\{z^{(q)}\}} \mathsf{x}_e(\mathsf{z}^{(e)}) = x_q(z^{(q)}) \\ (\forall e = (p,q) \in \mathcal{E})(\forall z^{(p)} \in \mathcal{L}) & \sum_{\mathsf{z}^{(e)}\in\{z^{(p)}\}\times\mathcal{L}} \mathsf{x}_e(\mathsf{z}^{(e)}) = x_p(z^{(p)}) \end{array} \right\}. \tag{44}$$

In the above formulation, for every $p \in \mathcal{V}$ and $e \in \mathcal{E}$, the unary binary variable $x_p(\cdot)$ and the pairwise binary variables $\mathsf{x}_e(\cdot)$ are indicator variables, determining the labels assigned to vertex p and to the pair of vertices connected by edge $e = (p', q')$ respectively, i.e.,

$$(\forall z^{(p)} \in \mathcal{L}) \qquad x_p(z^{(p)}) = 1 \quad \Leftrightarrow \quad p \text{ is assigned label } z^{(p)} \tag{45}$$

$$(\forall \mathsf{z}^{(e)} = (z^{(p')}, z^{(q')}) \in \mathcal{L}^2) \qquad \mathsf{x}_e(\mathsf{z}^{(e)}) = 1 \quad \Leftrightarrow \quad p', q' \text{ are assigned labels } z^{(p')}, z^{(q')}. \tag{46}$$

The vector x regrouping all these binary variables is of dimension $N = |\mathcal{V}||\mathcal{L}| + |\mathcal{E}||\mathcal{L}|^2$. The first constraints in (44) simply encode the fact that each vertex must be assigned exactly one label, whereas the rest of the constraints

enforces consistency between unary variables $x_p(\cdot)$, $x_q(\cdot)$ and the pairwise variable $\mathsf{x}_e(\cdot)$ for edge $e = (p, q)$, ensuring essentially that if $x_p(z^{(p)}) = x_q(z^{(q)}) = 1$, then $\mathsf{x}_e(z^{(p)}, z^{(q)}) = 1$.

As mentioned above, our goal will be to decompose the MRF problem (43) into easier subproblems (called *slaves*), which, in this case, involve optimizing MRFs defined on subgraphs of G. More specifically, let $\{G_m = (\mathcal{V}_m, \mathcal{E}_m)\}_{1 \leq m \leq M}$ be a set of subgraphs that form a decomposition of $G = (\mathcal{V}, \mathcal{E})$ (i.e., $\cup_{m=1}^{M} \mathcal{V}_m = \mathcal{V}$, $\cup_{m=1}^{M} \mathcal{E}_m = \mathcal{E}$). On each of these subgraphs, we define a local MRF with corresponding (unary and pairwise) potentials $\varphi^m = \big\{\{\varphi_p^m\}_{p \in \mathcal{V}_m}, \{\varphi_e^m\}_{e \in \mathcal{E}_m}\big\}$, whose cost function $f^m(x; \varphi^m)$ is thus given by

$$f^m(x; \varphi^m) = \sum_{p \in \mathcal{V}_m,\, z^{(p)} \in \mathcal{L}} \varphi_p^m(z^{(p)})\, x_p(z^{(p)}) + \sum_{e \in \mathcal{E}_m,\, \mathsf{z}^{(e)} \in \mathcal{L}^2} \varphi_e^m(\mathsf{z}^{(e)})\, \mathsf{x}_e(\mathsf{z}^{(e)}). \tag{47}$$

Moreover, the sum (over m) of the potential functions φ^m is ensured to give back the potentials φ of the original MRF on G, i.e.,[4]

$$(\forall p \in \mathcal{V})(\forall e \in \mathcal{E}) \quad \sum_{\substack{m \in \{1,\ldots,M\} \\ p \in \mathcal{V}_m}} \varphi_p^m = \varphi_p, \qquad \sum_{\substack{m \in \{1,\ldots,M\} \\ e \in \mathcal{E}_m}} \varphi_e^m = \varphi_e. \tag{48}$$

This guarantees that $f = \sum_{m=1}^{M} f^m$, thus allowing us to re-express problem (43) as follows

$$\underset{x \in C_G}{\text{minimize}} \;\; \sum_{m=1}^{M} f^m(x; \varphi^m). \tag{49}$$

An assumption that often holds in practice is that minimizing separately each of the f^m (over x) is easy, but minimizing their sum is hard. Therefore, to take advantage of this fact, we introduce, for every $m \in \{1, \ldots, M\}$, an *auxiliary copy* $x^m \in C_{G_m}$ for the variables of the local MRF defined on G_m, which are thus constrained to coincide with the corresponding variables in vector x, i.e., it holds $x^m = x_{|G_m}$, where $x_{|G_m}$ is used to denote the subvector of x containing only those variables associated with vertices and edges of subgraph G_m. In this way, Problem (49) can be transformed into

$$\begin{aligned} \underset{x \in C_G,\, \{x^m \in C_{G_m}\}_{1 \leq m \leq M}}{\text{minimize}} \;\; & \sum_{m=1}^{M} f^m(x^m; \varphi^m) \\ \text{s.t.} \;\; & \big(\forall m \in \{1, \ldots, M\}\big) \qquad x^m = x_{|G_m}. \end{aligned} \tag{50}$$

By considering the dual of (50), using a technique similar to the one described in framebox on page 9, and noticing that

$$x \in C_G \quad \Leftrightarrow \quad (\forall m \in \{1, \ldots, M\}) \quad x^m \in C_{G_m}, \tag{51}$$

we finally end up with the following problem:

$$\underset{(v^m)_{1 \leq m \leq M} \in \Lambda}{\text{maximize}} \;\; \sum_{m=1}^{M} h^m(v^m) \;\; , \tag{52}$$

[4]For instance, to ensure (48) we can simply set: $(\forall m \in \{1, \ldots, M\})\; \varphi_p^m = \frac{\varphi_p}{|\{|m' | p \in \mathcal{V}_{m'}\}|}$ and $\varphi_e^m = \frac{\varphi_e}{|\{m' | e \in \mathcal{E}_{m'}\}|}$.

where, for every $m \in \{1, \ldots, M\}$, the dual variable v^m consists of $\left\{\{v_p^m\}_{p \in \mathcal{V}_m}, \{\mathsf{v}_e^m\}_{e \in \mathcal{E}_m}\right\}$ similarly to φ^m, and function h^m is related to the following optimization of a slave MRF on G_m:

$$h^m(v^m) = \min_{x^m \in C_{G_m}} f^m(x^m; \varphi^m + v^m). \tag{53}$$

The feasible set Λ is defined as

$$\Lambda = \left\{ (v^m)_{1 \leq m \leq M} \middle| (\forall p \in \mathcal{V})(\forall z^{(p)} \in \mathcal{L}) \sum_{\substack{m \in \{1,\ldots,M\} \\ p \in \mathcal{V}_m}} v_p^m(z^{(p)}) = 0,\ (\forall e \in \mathcal{E})(\forall \mathsf{z}^{(e)} \in \mathcal{L}^2) \sum_{\substack{m \in \{1,\ldots,M\} \\ e \in \mathcal{E}_m}} \mathsf{v}_e(\mathsf{z}^{(e)}) = 0 \right\}. \tag{54}$$

The above dual problem provides a relaxation to the original problem (43)-(44). Furthermore, note that this relaxation leads to a convex optimization problem, although the original one is not. As such, it can always be solved in an optimal manner. A possible way of doing this consists of using a projected subgradient method. Exploiting the form of the projection onto the vector space Λ yields Algorithm 9 where $(\gamma_n)_{n \in \mathbb{N}}$ is a summable sequence of positive step-sizes and $\left\{\{\widehat{x}_{p,n}^m\}_{p \in \mathcal{V}_m}, \{\widehat{\mathsf{x}}_{e,n}^m\}_{e \in \mathcal{E}_m}\right\}$ corresponds to a subgradient of function h^m computed at iteration n [58]. Note that this algorithm requires *only solutions to local subproblems* to be computed, which is, of course, a task much easier that furthermore can be executed in a parallel manner. The solution to the master MRF is filled in from local solutions $\left\{\{\widehat{x}_{p,n}^m\}_{p \in \mathcal{V}_m}, \{\widehat{\mathsf{x}}_{e,n}^m\}_{e \in \mathcal{E}_m}\right\}_{1 \leq m \leq M}$ after convergence of the algorithm.

Algorithm 9 Dual decomposition for MRF optimization.

Choose a decomposition $\{G_m = (\mathcal{V}_m, \mathcal{E}_m)\}_{1 \leq m \leq M}$ of G

Initialize potentials of slave MRFs:

$(\forall m \in \{1, \ldots, m\})(\forall p \in \mathcal{V}_m)\ \varphi_{p,0}^m = \frac{\varphi_p}{|\{m' | p \in \mathcal{V}_{m'}\}|}, (\forall e \in \mathcal{E}_m)\ \varphi_{e,0}^m = \frac{\varphi_e}{|\{m' | e \in \mathcal{E}_{m'}\}|}$

for $n = 0, \ldots$

- Compute minimizers of slave MRF problems: $(\forall m \in \{1, \ldots, M\})\ \left\{\{\widehat{x}_{p,n}^m\}_{p \in \mathcal{V}_m}, \{\widehat{\mathsf{x}}_{e,n}^m\}_{e \in \mathcal{E}_m}\right\} \in \operatorname*{argmin}_{x^m \in C_{G_m}} f^m(x^m; \varphi_n^m)$
- Update potentials of slave MRFs:

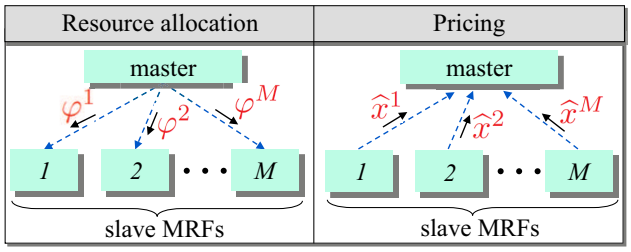
$(\forall m \in \{1, \ldots, m\})(\forall p \in \mathcal{V}_m)\ \varphi_{p,n+1}^m = \varphi_{p,n+1}^m + \gamma_n \left(\widehat{x}_{p,n}^m - \frac{\sum_{m:\, p \in \mathcal{V}_m} \widehat{x}_{p,n}^m}{|\{m' | p \in \mathcal{V}_{m'}\}|}\right)$

$(\forall m \in \{1, \ldots, m\})(\forall e \in \mathcal{E}_m)\ \varphi_{e,n+1}^m = \varphi_{e,n}^m + \gamma_n \left(\widehat{\mathsf{x}}_{e,n}^m - \frac{\sum_{m:\, e \in \mathcal{E}_m} \widehat{\mathsf{x}}_{e,n}^m}{|\{m' | p \in \mathcal{V}_{m'}\}|}\right).$

For a better intuition for the updates of variables $\left\{\{\varphi_{p,n}^m\}_{p \in \mathcal{V}_m}, \{\varphi_{e,n}^m\}_{e \in \mathcal{E}_m}\right\}_{1 \leq m \leq M, n \in \mathbb{N}}$ in Algorithm 9, we should note that their aim is essentially to bring a consensus among the solutions of the local subproblems. In other words, they try to adjust the potentials of the slave MRFs so that in the end the corresponding local solutions are consistent with each other, i.e., all variables corresponding to a common vertex or edge are assigned the same value by the different subproblems. If this condition is satisfied (i.e., there is a full consensus) then the overall solution that results from combining the consistent local solutions is guaranteed to be optimal. In general, though, this might not always be true given that the above procedure is solving only a *relaxation* of the original NP-hard problem.

MASTER-SLAVE COMMUNICATION

During dual decomposition a communication between a master process and the slaves (local subproblems) can be thought of as taking place, which can also be interpreted as a resource allocation/pricing stage.

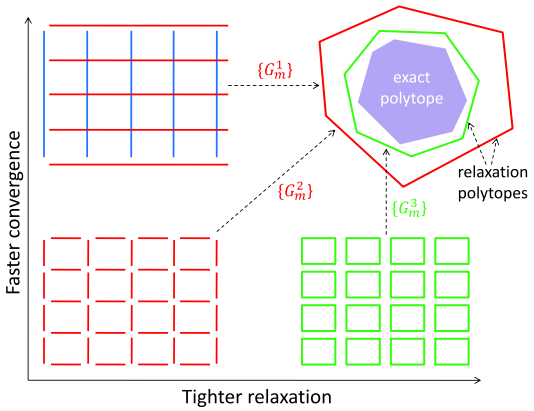


Resource allocation: At each iteration, the master assigns new MRF potentials (i.e., resources) $(\varphi^m)_{1\leq m\leq M}$ to the slaves based on the current local solutions $(\widehat{x}^m)_{1\leq m\leq M}$.

Pricing: The slaves respond by adjusting their local solutions $(\widehat{x}^m)_{1\leq m\leq M}$ (i.e., the prices) so as to maximize their welfares based on the newly assigned resources $(\widehat{x}^m)_{1\leq m\leq M}$.

DECOMPOSITIONS AND RELAXATIONS

Different decompositions can lead to different relaxations and/or can affect the speed of convergence. We show below, for instance, 3 possible decompositions for an MRF assumed to be defined on a 5×5 image grid.



Decompositions $\{G_m^1\}, \{G_m^2\}, \{G_m^3\}$ consist respectively of one suproblem per row and column, one subproblem per edge, and one subproblem per 2×2 subgrid of the original 5×5 grid. Both $\{G_m^1\}$ and $\{G_m^2\}$ (due to using solely subgraphs that are trees) lead to the same LP relaxation of (43), whereas $\{G_m^3\}$ leads to a relaxation that is tighter (due to containing loopy subgraphs).

On the other hand, decomposition $\{G_m^1\}$ leads to faster convergence compared with $\{G_m^2\}$ due to using larger subgraphs that allow a faster propagation of information during message-passing.

Interestingly, if we choose to use a decomposition consisting only of subgraphs that are trees then the resulting relaxation can be shown to actually coincide with the standard LP-relaxation of linear integer program (43) (generated by replacing the integrality constraints with non-negativity constraints on the variables). This also means that when this LP-relaxation is tight, an optimal MRF solution is computed. This, for instance, leads to the result that *dual decomposition approaches can estimate a globally optimal solution for binary submodular MRFs* (although it should be noted that much faster graph-cut based techniques exist for submodular problems of this type - see framebox on page 28). Furthermore, when using subgraphs that are trees, a minimizer to each slave problem can be computed efficiently by applying the Belief Propagation algorithm [63], which is a message-passing method. Therefore, in

this case, Algorithm 9 essentially reduces to a continuous exchange of messages between the nodes of graph G. Such an algorithm relates to or generalizes various other message-passing approaches [64]–[69]. In general, besides tree-structured subgraphs, other types of decompositions or subproblems can be used as well (such as binary planar problems, or problems on loopy subgraphs with small tree-width, for which MRF optimization can still be solved efficiently), which can lead to even tighter relaxations (see framebox on page 27) [70]–[75].

GRAPH-CUTS AND MRF OPTIMIZATION

For certain MRFs, optimizing their cost is known to be equivalent to solving a polynomial mincut problem [76], [77]. These are exactly all the binary MRFs ($|\mathcal{L}| = 2$) with submodular pairwise potentials such that, for every $e \in \mathcal{E}$,

$$\varphi_e(0,0) + \varphi_e(1,1) \leq \varphi_e(0,1) + \varphi_e(1,0). \tag{55}$$

Due to (55), the cost $f(x)$ of a binary labeling $x = (x^{(p)})_{1 \leq p \leq |\mathcal{V}|} \in \{0,1\}^{|\mathcal{V}|}$ for such MRFs can always be written (up to an additive constant) as

$$f(x) = \sum_{p \in \mathcal{V}_P} a_p x^{(p)} + \sum_{p \in \mathcal{V}_N} a^{(p)}(1 - x^{(p)}) + \sum_{(p,q) \in \mathcal{E}} a_{p,q} x^{(p)}(1 - x^{(q)}), \tag{56}$$

where all coefficients $(a_p)_{p \in \mathcal{V}}$ and $(a_{p,q})_{(p,q) \in \mathcal{E}}$ are nonnegative ($\mathcal{V}_P \subset \mathcal{V}$, $\mathcal{V}_N \subset \mathcal{V}$).
In this case, we can associate to f a capacitated network that has vertex set $\mathcal{V}_f = \mathcal{V} \cup \{s,t\}$. A source vertex s and a sink one t have thus been added. The new edge set $\mathcal{E}_f$ is deduced from the one used to express f:

$$\mathcal{E}_f = \{(p,t) \mid p \in \mathcal{V}_P\} \cup \{(s,p) \mid p \in \mathcal{V}_N\} \cup \mathcal{E},$$

and its edge capacities are defined as $(\forall p \in \mathcal{V}_P \cup \mathcal{V}_N)\ c_{p,t} = c_{s,p} = a_p$, and $(\forall (p,q) \in \mathcal{E})\ c_{p,q} = a_{p,q}$.
A one-to-one correspondence between s-t cuts and MRF labelings then exists:

$$x \in \{0,1\}^{|\mathcal{V}|} \leftrightarrow \mathrm{cut}(x) = \{s\} \cup \{p \mid x^{(p)} = 1\}$$

for which it is easy to see that

$$f(x) = \sum_{u \in \mathrm{cut}(x), v \notin \mathrm{cut}(x)} c_{u,v} = \text{cost of } \mathrm{cut}(x)\ .$$

Computing a mincut, in this case, solves the LP relaxation of (43), which is tight, whereas computing a max-flow solves the dual LP.

Furthermore, besides the projected subgradient method, one can alternatively apply an ADMM scheme for solving relaxation (52) (see Section III-B). The main difference, in this case, is that the optimization of a slave MRF problem is performed by solving a (usually simple) local quadratic problem, which can again be solved efficiently for an appropriate choice of the decomposition (see Section III-F). This method again penalizes disagreements among slaves, but it does so even more aggressively than the subgradient method since there is no longer a requirement for step-sizes $(\gamma_n)_{n \in \mathbb{N}}$ converging to zero. Furthermore, alternative smoothed accelerated schemes exist and can be applied as well [78]–[80].

V. Applications

Although the presented primal-dual algorithms can be applied virtually to any area where optimization problems have to be solved, we now mention a few common applications of these techniques.

A. Inverse problems

For a long time, convex optimization approaches have been successfully used for solving inverse problems such as signal restoration, signal reconstruction, or interpolation of missing data. Most of the time, these problems are ill-posed and, in order to recover the signal of interest in a satisfactory manner, some prior information needs to be introduced. To do this, an objective function can be minimized which includes a data fidelity term modelling knowledge about the noise statistics and possibly involves a linear observation matrix (e.g. a convolutive blur), and a regularization (or penalization) term which corresponds to the additional prior information. This formulation can also often be justified statistically as the determination of a Maximum A Posteriori (MAP) estimate. In early developed methods, in particular in Tikhonov regularization, a quadratic penalty function is employed. Alternatively, hard constraints can be imposed on the solution (for example, bounds on the signal values), leading to signal feasibility problems. Nowadays, a hybrid regularization [81] may be prefered so as to combine various kinds of regularity measures, possibly computed for different representations of the signal (Fourier, wavelets,...), some of them like total variation [9] and its nonlocal extensions [82] being taylored for preserving discontinuities such as image edges. In this context, constraint sets can be translated into penalization terms being equal to the indicator functions of these sets (see (2)). Altogether, these lead to global cost functions which can be quite involved, often with many variables, for which the splitting techniques described in Section III-F are very useful. An extensive literature exists on the use of ADMM methods for solving inverse problems (e.g., see [16]–[20]). With the advent of more recent primal-dual algorithms, many works have been mainly focused on image recovery applications [29]–[32], [34], [37], [38], [40], [45], [47], [83]–[85]. Two illustrations are now provided.

In [86], a generalization of the total variation is defined for an arbitrary graph in order to solve a variety of inverse problems. For denoising applications, the cost function which is minimized reads

$$(\forall x \in \mathbb{R}^N) \qquad f(x) = \sigma_C(Lx) + \frac{1}{2}\|x - y\|^2 \tag{57}$$

where x is a vector of variables associated with each vertex of a weighted graph, and $y \in \mathbb{R}^N$ is a vector of data observed at each vertex. The matrix $L \in \mathbb{R}^{K\times N}$ is equal to $\mathrm{Diag}(\sqrt{\omega}_1, \ldots, \sqrt{\omega}_K)\, A$ where $(\omega_1, \ldots \omega_K) \in [0, +\infty[^K$ is the vector of edge weights and $A \in \mathbb{R}^{K\times N}$ is the graph incidence matrix playing a role similar to a gradient operator on the graph. The set C is defined as an intersection of closed semi-balls in such a way that its support function σ_C (see (9)) allows us to define a class of functions extending the total variation semi-norm (see [86] for more details). Good image denoising results can be obtained by building the graph in a nonlocal manner following the strategy in [82]. Results obtained for Barbara image are displayed in Fig. 5. Interestingly, the ability of methods such as those presented in Section III-D to circumvent matrix inversions leads to a significant decrease of the convergence time for irregular graphs in comparison with algorithms based on the Douglas-Rachford iteration or ADMM (see Fig. 6).

Another application example of primal-dual proximal algorithms is Parallel Magnetic Resonance Imaging (PMRI) reconstruction. A set of measurement vectors $(z_j)_{1\leq j\leq J}$ is acquired from J coils. These observations are related to the original full FOV (Field Of View) image $\overline{x} \in \mathbb{C}^N$ corresponding to a spin density. An estimate of $\overline{x}$ is obtained

(a) Original image (b) Noisy SNR = 14.47 dB (c) Nonlocal TV SNR = 20.78 dB

Fig. 5: Nonlocal denoising (additive white zero-mean Gaussian noise with variance $\sigma^2 = 20$).

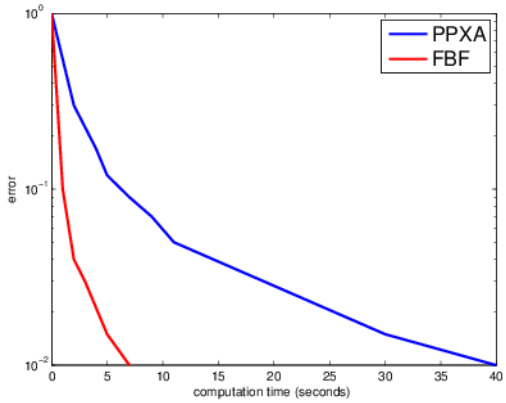


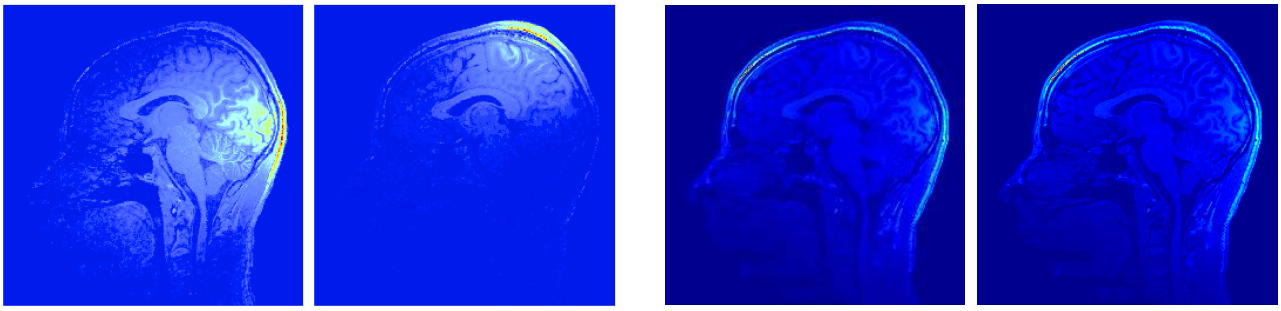
Fig. 6: Comparison of the convergence speed of a Douglas-Rachford based algorithm (PPXA [48]) (blue) and an FBF-based primal-dual algorithm (red) for image denoising using a non-regular graph, Matlab implementation on an Intel Xeon 2.5GHz 8-core system.

by solving the following problem:

$$\underset{x\in\mathbb{C}^N}{\text{minimize}} \;\; \iota_C(x) + g(Lx) + \sum_{j=1}^{J} \|\Sigma F S_j x - z_j\|^2_{\Lambda_j^{-1}} \tag{58}$$

where $(\forall j \in \{1,\dots,J\})\ \|\cdot\|^2_{\Lambda_j^{-1}} = (\cdot)^{\mathrm{H}}\Lambda_j^{-1}(\cdot)$, Λ_j is the noise covariance matrix for the j-the channel, $S_j \in \mathbb{C}^{N\times N}$ is a diagonal matrix modelling the sensitivity of the coil, $F \in \mathbb{C}^{N\times N}$ is a 2D discrete Fourier transform, $\Sigma \in \{0,1\}^{\lfloor \frac{N}{R} \rfloor \times N}$ is a subsampling matrix, $g \in \Gamma_0(\mathbb{C}^K)$ is a sparsity measure (e.g. a weighted ℓ_1-norm), $L \in \mathbb{C}^{K\times N}$ is a (possibly redundant) frame analysis operator, and C is a vector subspace of $\mathbb{C}^N$ serving to set to zero the image areas corresponding to the background.[5] Combining suitable subsampling strategies in the k-space with the use of an array of coils allows us to reduce the acquisition time while maintaining a good image quality. The subsampling factor $R > 1$ thus corresponds to an *acceleration factor*. For a more detailed account on the considered approach,

[5] $(\cdot)^{\mathrm{H}}$ denotes the transconjugate operation and $\lfloor\cdot\rfloor$ designates the lower rounding operation.



(a) (b)

Fig. 7: **(a)** Effects of the sensitivity matrices in the spatial domain in the absence of subsampling: the moduli of the images corresponding to $(S_j\overline{x})_{2\leq j\leq 3}$ are displayed for 2 channels out of 32. **(b)** Reconstruction quality: moduli of the original slice $\overline{x}$ and the reconstructed one with SNR = 20.57 dB (from left to right) using polynomial sampling of order 1 with $R = 5$, a wavelet frame, and an ℓ_1 regularization.

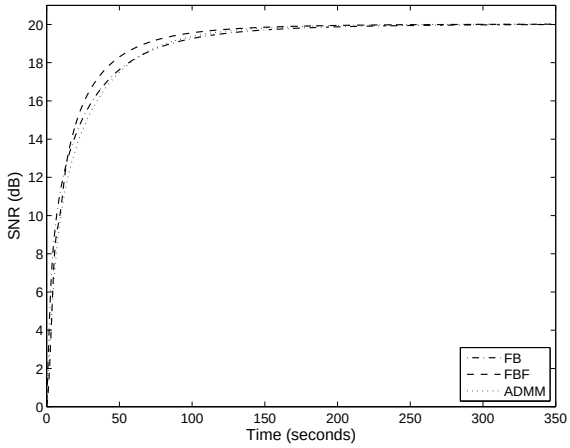


Fig. 8: Signal-to-Noise Ratio as a function of computation time using ADMM, and FB or FBF-based primal-dual methods for a given slice, Matlab implementation on an Intel i7-3520M CPU@2.9 GHz system.

the reader is refered to [87], [88] and the references therein. Reconstruction results are shown in Fig. 7. Fig. 8 also allows us to evaluate the convergence time for various algorithms. It can be observed that smaller differences between the implemented primal-dual strategies are apparent in this example. Due to the form of the subsampling matrix, the matrix inversion involved at each iteration of ADMM however requires to make use of a few subiterations of a linear conjugate gradient method.

Note that convex primal-dual proximal optimization algorithms have been applied to other fields than image recovery, in particular to machine learning [4], system identification [89], audio processing [90], optimal transport [91], empirical mode decomposition [92], seimics [93], and database management [94].

B. Computer vision and image analysis

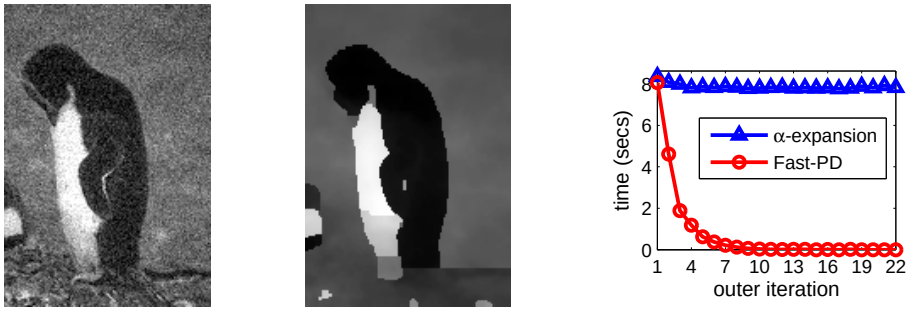
The great majority of problems in computer vision involve image observation data that are of very high dimensionality, inherently ambiguous, noisy, incomplete and often only provide a partial view of the desired space.

Hence, any successful model that aims to explain such data usually requires a reasonable regularization, a robust data measure, and a compact structure between the variables of interest to efficiently characterize their relationships. Probabilistic graphical models, and in particular discrete Markov Random Fields, have led to a suitable methodology for solving such visual perception problems [95], [96]. This type of models offer great representational power, and are able to take into account dependencies in the data, encode prior knowledge, and model (soft or hard) contextual constraints in a very efficient and modular manner. Furthermore, they offer the important ability to make use of very powerful data likelihood terms consisting of arbitrary nonconvex and non-continuous functions that are often crucial for accurately representing the problem at hand. As a result, MAP-inference for these models leads to discrete optimization problems that are (in most cases) highly nonconvex (NP-hard) and also of very large scale [97], [98].

Primal-dual approaches can offer important computational advantages when dealing with such problems. One such characteristic example is the FastPD algorithm [57], which currently provides one of the most efficient methods for solving generic MRF optimization problems of this type, also guaranteeing at the same time the convergence to solutions that are approximately optimal. The theoretical derivation of this method relies on the use of the primal-dual schema described in Section IV, which results, in this case, into a very fast graph-cut based inference scheme that generalizes previous state-of-the-art approaches such as the α-expansion algorithm [99] (see Fig. 9). More generally, due to the very wide applicability of MRF models to computer vision or image analysis problems, primal-dual approaches can and have been applied to a broad class of both low-level and high-level problems from these domains, including image segmentation [100]–[103], stereo matching and 3D multi-view reconstruction [104], [105], graph-matching [106], 3D surface tracking [107], optical flow estimation [108], scene understanding [109], image deblurring [110], panoramic image stitching [111], category-level segmentation [112], and motion tracking [113]. In the following we mention very briefly just a few examples.

A primal-dual based optimization framework has been recently proposed in [114], [115] for the problem of deformable registration/fusion, which forms one of the most central and challenging tasks in medical image analysis. This problem consists of recovering a nonlinear dense deformation field that aligns two signals that have in general an unknown relationship both in the spatial and intensity domain. In this framework, towards dimensionality reduction on the variables, the dense registration field is first expressed using a set of control points (registration grid) and an interpolation strategy. Then, the registration cost is expressed using a discrete sum over image costs projected on the control points, and a smoothness term that penalizes local deviations on the deformation field according to a neighborhood system on the grid. One advantage of the resulting optimization framework is that it is able to encode even very complex similarity measures (such as normalized mutual information and Kullback-Leibler divergence) and therefore can be used even when seeking transformations between different modalities (inter-deformable registration). Furthermore, it admits a broad range of regularization terms, and can also be applied to both 2D-2D and 3D-3D registration, as an arbitrary underlying graph structure can be readily employed (see Fig. 10 for a result on 3D inter-subject brain registration).

Another application of primal-dual methods is in stereo reconstruction [117], where given as input a pair of left



(a) 'Penguin' image denoising (from left to right: noisy input image, FastPD output, time comparison plot)

(b) 'Tsukuba' stereo matching (from left to right: left image, FastPD output, time comparison plot)

Fig. 9: FastPD [57] results for an image denoising (top) and stereo-matching (bottom) problem. The plot in each row compares the corresponding running time per iteration of the above primal-dual algorithm with the α-expansion algorithm, which is a primal-based method (experiments conducted on a 1.6 GHz CPU).

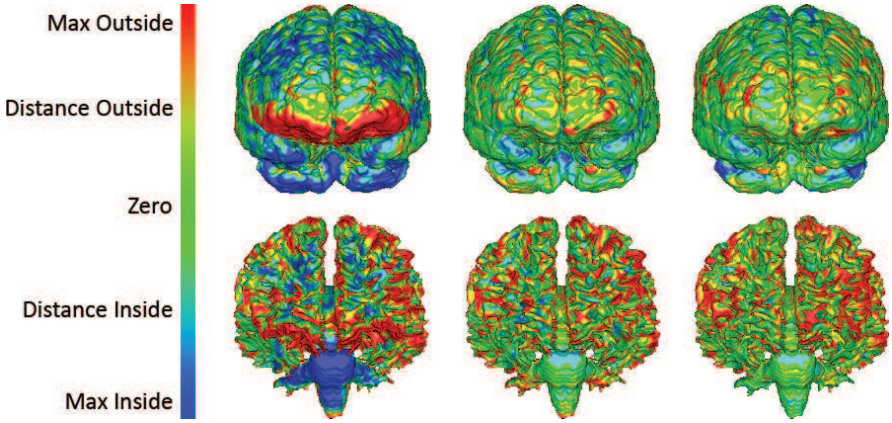


Fig. 10: Color encoded visualization of the surface distance between warped and expert segmentation after affine (left), FFD-based [116] (middle), and primal-dual based registration (right) for the Brain 1 data set. The color range is scaled to a maximum and minimum distance of 3 mm. The average surface distance (ASD) after registration for the gray matter is 1.66, 1.14, and 1.00 mm for affine, FFD-based, and primal-dual method, respectively. For the white matter the resulting ASD is 1.92, 1.31, and 1.06 mm. Note also that the FFD-based method is more than 30 times slower than the primal-dual approach.

and right images I_L, I_R we seek to estimate a function $u : \Omega \rightarrow \Gamma$ representing the depth $u(s)$ at a point s in the domain $\Omega \subset \mathbb{R}^2$ of the left image (here $\Gamma = [\upsilon_{\min}, \upsilon_{\max}]$ denotes the allowed depth range). To accomplish this, the following variational problem is proposed in [117]:

$$\underset{u}{\text{minimize}} \quad \int_{\Omega} f(u(s), s) ds + \int_{\Omega} |\nabla u(s)| ds, \tag{59}$$

Fig. 11: Estimated depth map (right) for a large aerial stereo data set of Graz using the primal-dual approach in [117]. One of the images of the corresponding stereoscopic pair (of size 1500×1400) is shown on the left.

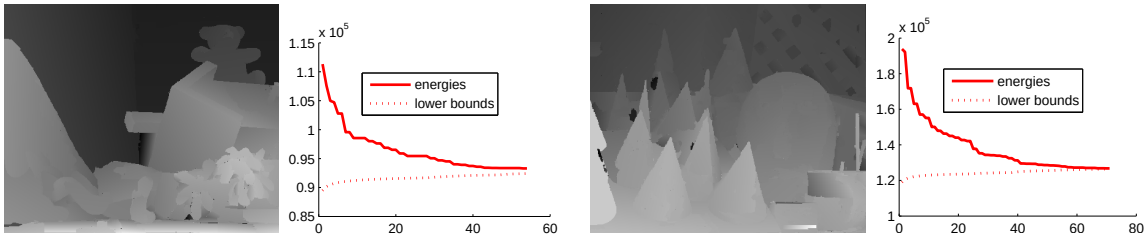
where $f(u(s), s)$ is a data term favoring different depth values by measuring the absolute intensity differences of respective patches projected in the two input images, and the second term is a TV regularizer that promotes spatially smooth depth fields. The above problem is nonconvex (due to the use of the data term f), but it turns out that there exists an equivalent convex formulation obtained by lifting the original problem to a higher-dimensional space, in which u is represented in terms of its level sets

$$\underset{\phi \in D}{\text{minimize}} \quad \int_{\Sigma} (|\nabla \phi(s, \upsilon)| + f(s, \upsilon) |\partial_{\upsilon} \phi(s, \upsilon)|) ds d\upsilon. \tag{60}$$

In the above formulation, $\Sigma = \Omega \times \Gamma$, $\phi \colon \Sigma \to \{0, 1\}$ is a binary function such that $\phi(s, \upsilon)$ equals 1 if $u(s) > \upsilon$ and 0 otherwise, and the feasible set is defined as $D = \{\phi \colon \Sigma \to \{0, 1\} \mid (\forall s \in \Omega)\, \phi(s, \upsilon_{\min}) = 1, \phi(s, \upsilon_{\max}) = 0\}$. A convex relaxation of the latter problem is obtained by using $D' = \big\{\phi \colon \Sigma \to [0, 1] \mid (\forall s \in \Omega)\ \phi(s, \upsilon_{\min}) = 1, \phi(s, \upsilon_{\max}) = 0\big\}$ instead of D. A discretized form of the resulting optimization problem can be solved with the algorithms described in Section III-C. Fig. 11 shows a sample result of this approach.

Recently, primal-dual approaches have also been developed for discrete optimization problems that involve higher-order terms [118]–[120]. They have been applied successfully to various tasks, like, for instance, in stereo matching [118]. In this case, apart from a data term that measures similarity between corresponding pixels in two images, a discontinuity-preserving smoothness prior of the form $\varphi(s_1, s_2, s_3) = \min(|s_1 - 2s_2 + s_3|, \kappa)$ with $\kappa \in \,]0, +\infty[$ has been employed as a regularizer that penalizes depth surfaces of high curvature. Indicative stereo matching results from an algorithm based on the dual decomposition principle described in Section IV-D are shown in Fig. 12.

It should be also mentioned that an advantage of all primal-dual algorithms (which is especially important for NP-hard problems) is that they also provide (for free) per-instance approximation bounds, specifying how far the cost of an estimated solution can be from the unknown optimal cost. This directly follows from the fact that these methods are computing both primal and dual solutions, which (in the case of a minimization task) provide respectively upper and lower limits to the true optimum. These approximation bounds are continuously updated throughout an algorithm execution, and thus can be directly used for assessing the performance of a primal-dual method with respect to any particular problem instance (and without essentially any extra computational cost). Moreover, often in practice, these sequences converge to a common value, which means that the corresponding estimated solutions are almost optimal (see, e.g., the plots in Fig. 12).



(a) 'Teddy' (b) 'Cones'

Fig. 12: Stereo matching results for 'Teddy' (left) and 'Cones' (right) when using a higher-order discontinuity preserving smoothness prior. We also show plots for the corresponding sequences of upper and lower bounds generated during the primal-dual method. Notice that these sequences are converging to the same limit, meaning that the estimated solution converges to the optimal value.

VI. Conclusion

In this paper, we have reviewed a number of primal-dual optimization methods which can be employed for solving signal and image processing problems. The links existing between convex approaches and discrete ones were little explored in the literature and one of the contributions of this paper is to put them in a unifying perspective. Although the presented algorithms have been proved to be quite effective in numerous problems, there remains much room for extending their scope to other application fields, and also for improving them so as to accelerate their convergence. Another issue to pay attention to is the robustness to numerical errors but, for example, it can be mentioned that most of the existing proximal algorithms are tolerant to summable errors. Various techniques can be devised for designing faster variants of these methods (preconditioning, activation of blocks of variables, combination with stochastic strategies, distributed implementations,...). Concerning discrete optimization methods, we have shown that the key to success lies in tight relaxations of combinatorial NP hard problems. Extending these methods to more challenging problems, e.g. those involving higher-order Markov fields or extremely large label sets, appears to be of main interest in this area. More generally, developing primal-dual strategies that further bridge the gap between continuous and discrete approaches, as well as for solving other kinds of nonconvex optimization problems such as those encountered in phase reconstruction or blind deconvolution opens the way to appealing investigations. So, the ground is yours now to play with duality !

References

[1] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation. Vol. I: Basic theory*, vol. 330 of *Series of Comprehensive Studies in Mathematics*, Springer-Verlag, Berlin-Heidelberg, 2006.

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

[3] 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.

[4] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends in Machine Learn.*, vol. 4, no. 1, pp. 1–106, 2012.

[5] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2013.

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

[7] R Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[8] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[9] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1-4, pp. 259–268, Nov. 1992.

[10] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[11] R. I. Boţ, *Conjugate Duality in Convex Optimization*, vol. 637 of *Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, Berlin Heidelberg, 2010.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Machine Learn.*, vol. 8, no. 1, pp. 1–122, 2011.

[13] D. Bertsimas and J. N. John N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Nashua, NH, USA, 1997.

[14] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite elements approximations," *Comput. Math. Appl.*, vol. 2, pp. 17–40, 1976.

[15] M. Fortin and R. Glowinski, Eds., *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, Elsevier Science Ltd, Amsterdam: North-Holland, 1983.

[16] J.-F. Giovannelli and A. Coulais, "Positive deconvolution for superimposed extended source and point sources," *Astron. Astrophys.*, vol. 439, pp. 401412, 2005.

[17] T. Goldstein and S. Osher, "The split Bregman method for ℓ_1-regularized problems," *SIAM J. Imaging Sci.*, vol. 2, pp. 323–343, 2009.

[18] M. A. T. Figueiredo and R. D. Nowak, "Deconvolution of Poissonian images using variable splitting and augmented Lagrangian optimization," in *IEEE Work. on Stat. Sig. Proc.*, Cardiff, United Kingdom, Aug. 31 - Sept. 3 2009, pp. x+4.

[19] M. A. T. Figueiredo and J. M. Bioucas-Dias, "Restoration of Poissonian images using alternating direction optimization," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3133–3145, Dec. 2010.

[20] 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," *j-ieee-tip*, vol. 20, no. 3, pp. 681–695, Mar. 2011.

[21] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," 2013, http://arxiv.org/abs/1208.3922.

[22] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Programming*, vol. 55, pp. 293–318, 1992.

[23] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. Selected Topics Signal Process.*, vol. 1, no. 4, pp. 564–574, Dec. 2007.

[24] R. I. Boţ and C. Hendrich, "A Douglas-Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2541–2565, Dec. 2013.

[25] G. Chen and M. Teboulle, "A proximal-based decomposition method for convex minimization problems," *Math. Program.*, vol. 64, pp. 81–101, 1994.

[26] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.

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

[28] 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.

[29] 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, 2011.

[30] E. Esser, X. Zhang, and T. Chan, "A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science," *SIAM J. Imaging Sci.*, vol. 3, no. 4, pp. 1015–1046, 2010.

[31] B. He and X. Yuan, "Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective," *SIAM J. Imaging Sci.*, vol. 5, no. 1, pp. 119–149, 2012.

[32] T. Pock and A. Chambolle, "Diagonal preconditioning for first order primal-dual algorithms in convex optimization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 6-13 2011, pp. 1762–1769.

[33] P. L. Combettes and B. C. Vũ, "Variable metric forward-backward splitting with applications to monotone inclusions in duality," *Optimization*, 2012, Published online DOI:10.1080/02331934.2012.733883.

[34] T. Goldstein, E. Esser, and R. Baraniuk, "Adaptive primal-dual hybrid gradient methods for saddle-point problems," 2013, http://arxiv.org/abs/1305.0546.

[35] P. L Combettes, L Condat, J.-C. Pesquet, and B. C. Vũ, "A forward-backward view of some primal-dual optimization methods in image recovery," in *Proc. Int. Conf. Image Process.*, Paris, France, 27-30 Oct. 2014.

[36] J Liang, J. Fadili, and G. Peyré, "Convergence rates with inexact nonexpansive operators," 2014, http://arxiv.org/abs/1404.4837.

[37] I. Loris and C. Verhoeven, "On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty," *Inverse Problems*, vol. 27, no. 12, pp. 125007, 2011.

[38] P. Chen, J. Huang, and X. Zhang, "A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration," *Inverse Problems*, vol. 29, no. 2, pp. 025011, 2013.

[39] P. L. Combettes, D. Dũng, and B. C. Vũ, "Dualization of signal recovery problems," *Set-Valued Var. Anal.*, vol. 18, pp. 373–404, Dec. 2010.

[40] A. Jezierska, E. Chouzenoux, J.-C. Pesquet, and H. Talbot, "A primal-dual proximal splitting approach for restoring data corrupted with Poisson-Gaussian noise," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Kyoto, Japan, 25-30 Mar. 2012, pp. 1085–1088.

[41] P. Tseng, "A modified forward-backward splitting method for maximal monotone mappings," *SIAM J. Control Optim.*, vol. 38, pp. 431–446, 2000.

[42] 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.

[43] 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, June 2012.

[44] P. L. Combettes, "Systems of structured monotone inclusions: duality, algorithms, and applications," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2420–2447, Dec. 2013.

[45] R. I. Boţ and C. Hendrich, "Convergence analysis for a primal-dual monotone + skew splitting algorithm with applications to total variation minimization," *J. Math. Imaging Vision*, 2014, Accepted, http://www.mat.univie.ac.at/∼rabot/publications/jour13-18.pdf.

[46] A. Alotaibi, P. L. C Patrick L. Combettes, and N. Shahzad, "Solving coupled composite monotone inclusions by successive Fejér approximations of their Kuhn-Tucker set," 2014, http://arxiv.org/abs/1312.6696.

[47] S. R. Becker and P. L. Combettes, "An algorithm for splitting parallel sums of linearly composed monotone operators, with applications to signal recovery," *Nonlinear Convex Anal.*, vol. 15, no. 1, pp. 137–159, Jan. 2014.

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

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

[50] 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.

[51] V. Kolmogorov, "Generalized roof duality and bisubmodular functions," in *Proc. Ann. Conf. Neur. Inform. Proc. Syst.*, Vancouver, Canada., 6-9 Dec. 2010, pp. 1144–1152.

[52] F. Kahl and P. Strandmark, "Generalized roof duality," *Discrete Appl. Math.*, vol. 160, no. 16-17, pp. 2419–2434, 2012.

[53] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewoods Cliffs, N.J., 1982.

[54] V. V. Vazirani, *Approximation Algorithms*, Springer-Verlag, New York, NY, USA, 2001.

[55] D. S. Hochbaum, Ed., *Approximation Algorithms for NP-hard Problems*, PWS Publishing Co., Boston, MA, USA, 1997.

[56] N. Komodakis and G. Tziritas, "Approximate labeling via graph-cuts based on linear programming," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 29, no. 8, pp. 1436–1453, Aug. 2007.

[57] N. Komodakis, G. Tziritas, and N. Paragios, "Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art with primal-dual strategies," *Computer Vision and Image Understanding*, vol. 112, no. 1, pp. 14–29, Oct. 2008.

[58] N. Komodakis, N. Paragios, and G. Tziritas, "MRF energy minimization and beyond via dual decomposition," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 33, no. 3, pp. 531–552, Jan. 2011.

[59] N. Komodakis, N. Paragios, and G. Tziritas, "MRF optimization via dual decomposition: Message-passing revisited," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 14-21 Oct. 2007, pp. 1–8.

[60] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[61] A. Blake, P. Kohli, and C. Rother, *Markov Random Fields for Vision and Image Processing*, MIT Press, 2011.

[62] C. Chekuri, S. Khanna, J. Naor, and L. Zosin, "Approximation algorithms for the metric labeling problem via a new linear programming formulation," in *12^{th} Annual ACM-SIAM Symposium on Discrete Algorithms*, Washington D.C., USA, 7-9 Jan. 2001, pp. 109–118.

[63] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, San Francisco, CA, USA, 1988.

[64] M. Wainwright, T. Jaakkola, and A. Willsky, "MAP estimation via agreement on trees: message-passing and linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 11, pp. 3697–3717, Nov. 2005.

[65] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 28, no. 10, pp. 1568–1583, Aug. 2006.

[66] T. Werner, "A linear programming approach to max-sum problem: A review," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 29, no. 7, pp. 1165–1179, Jul. 2007.

[67] A. Globerson and T. Jaakkola, "Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations," in *Proc. Ann. Conf. Neur. Inform. Proc. Syst.*, Vancouver and Whistler, Canada, 3-6 Dec. 2007, pp. 553–560.

[68] C. Yanover, T. Talya Meltzer, and Y. Weiss, "Linear programming relaxations and belief propagation – an empirical study," *Journal of Machine Learning Research*, vol. 7, pp. 1887–1907, Sep. 2006.

[69] T. Hazan and A. Shashua, "Norm-product belief propagation: Primal-dual message-passing for approximate inference," *IEEE Trans. Inform. Theory*, vol. 56, no. 12, pp. 6294–6316, Dec. 2010.

[70] S. Jegelka, F. Bach, and S. Sra, "Reflection methods for user-friendly submodular optimization," in *Proc. Ann. Conf. Neur. Inform. Proc. Syst.*, Lake Tahoe, NV, USA, 5-10 Dec. 2013, pp. 1313–1321.

[71] N. N. Schraudolph, "Polynomial-time exact inference in NP-hard binary MRFs via reweighted perfect matching," in *13-th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Chia Laguna Resort, Sardinia, Italy, 13-15 May 2010, pp. 717–724.

[72] A. Osokin, D. Vetrov, and V. Kolmogorov, "Submodular decomposition framework for inference in associative markov networks with global constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, 21-23 June 2011, pp. 1889–1896.

[73] R. Yarkony, J. andMorshed, A. T. Ihler, and C. Fowlkes, "Tightening MRF relaxations with planar subproblems," in *Conference on Uncertainty in Artificial Intelligence*, Barcelona,Spain, 14-17 Jul. 2011, pp. 770–777.

[74] D. Sontag, T. Meltzer, A. Globerson, Y. Weiss, and T. Jaakkola, "Tightening LP relaxations for MAP using message passing," in *Conference on Uncertainty in Artificial Intelligence*, Helsinki, Finland, 9-12 Jul. 2008, pp. 656–664.

[75] N. Komodakis and N. Paragios, "Beyond loose LP-relaxations: Optimizing MRFs by repairing cycles," in *Proc. European Conference on Computer Vision*, Marseille, France, 12-18 Oct. 2008, pp. 806–820.

[76] E. Boros and P. L. Hammer, "Pseudo-Boolean optimization," *Discrete Appl. Math.*, vol. 123, no. 1-3, pp. 155–225, 2002.

[77] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[78] V. Jojic, S. Gould, and D. Koller, "Fast and smooth: Accelerated dual decomposition for MAP inference," in *International Conference on Machine Learning*, Haifa, Israel, 21-24 June 2010, pp. 503–510.

[79] B. Savchynskyy, J. H. Kappes, S. Schmidt, and C. Schnörr, "A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling," in *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, 21-23 June 2011, pp. 1817–1823.

[80] B. Savchynskyy, S. Schmidt, J. H. Kappes, and C. Schnörr, "Efficient MRF energy minimization via adaptive diminishing smoothing," in *Conference on Uncertainty in Artificial Intelligence*, Catalina Island, USA, 15-17 Aug. 2012, pp. 746–755.

[81] 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.

[82] X. Zhang, M. Burger, X. Bresson, and S. Osher, "Bregmanized nonlocal regularization for deconvolution and sparse reconstruction," *SIAM J. Imaging Sci.*, vol. 3, no. 3, pp. 253–276, 2010.

[83] A. Repetti, E. Chouzenoux, and J.-C. Pesquet, "A penalized weighted least squares approach for restoring data corrupted with signal-dependent noise," in *Proc. Eur. Sig. and Image Proc. Conference*, Bucharest, Romania, 27-31 Aug. 2012, pp. 1553–1557.

[84] S. Harizanov, J.-C. Pesquet, and G. Steidl, "Epigraphical projection for solving least squares Anscombe transformed constrained optimization problems," in *4th International Conference on Scale-Space and Variational Methods in Computer Vision*, A. Kuijper et al., Ed., Schloss Seggau, Leibnitz, Austria, 2-6 June 2013, vol. 7893 of *Lecture Notes in Computer Science*, pp. 125–136, Springer-Verlag, Berlin.

[85] T. Teuber, G. Steidl, and R.-H. Chan, "Minimization and parameter estimation for seminorm regularization models with i-divergence constraints," *Inverse Problems*, vol. 29, pp. 035007, Mar. 2013.

[86] C. Couprie, L. Grady, L. Najman, J.-C. Pesquet, and H. Talbot, "Dual constrained TV-based regularization on graphs," *SIAM J. Imaging Sci.*, vol. 6, pp. 1246–1273, 2013.

[87] L. Chaari, J.-C. Pesquet, A. Benazza-Benyahia, and Ph. Ciuciu, "A wavelet-based regularized reconstruction algorithm for SENSE parallel MRI with applications to neuroimaging," *Medical Image Analysis*, vol. 15, no. 2, pp. 185–201, Apr. 2011.

[88] A. Florescu, E. Chouzenoux, J.-C. Pesquet, Ph. Ciuciu, and S. Ciochina, "A Majorize-Minimize Memory Gradient method for complex-valued inverse problems," *Signal Process.*, vol. 103, pp. 285–295, Oct. 2014, Special issue on Image Restoration and Enhancement: Recent Advances and Applications.

[89] S. Ono, M. Yamagishi, and I. Yamada, "A sparse system identification by using adaptively-weighted total variation via a primal-dual splitting approach," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Vancouver, Canada, 26-31 May 2013, pp. 6029–6033.

[90] I. Bayram and O. D. Akyildiz, "Primal-dual algorithms for audio decomposition using mixed norms," *Sig., Image and Video Proc.*, vol. 8, no. 1, pp. 95–110, Jan. 2014.

[91] N. Papadakis, G. Peyré, and E. Oudet, "Optimal transport with proximal splitting," *SIAM J. Imaging Sci.*, vol. 7, no. 1, pp. 212–238, 2014.

[92] N. Pustelnik, P. Borgnat, and P. Flandrin, "Empirical Mode Decomposition revisited by multicomponent nonsmooth convex optimization," *Signal Process.*, vol. 102, pp. 313–331, Sept. 2014.

[93] M.-Q. Pham, C. Chaux, L. Duval, and J.-C. Pesquet, "Sparse template-based adaptive filtering with a primal-dual proximal algorithm: Application to seismic multiple removal," *IEEE Trans. Signal Process.*, 2014, accepted for publication, http://hal-upec-upem.archives-ouvertes.fr/hal-00914628.

[94] G. Moerkotte, M. Montag, A. Repetti, and G. Steidl, "Proximal operator of quotient functions with application to a feasibility problem in query optimization," 2014, http://hal.archives-ouvertes.fr/docs/00/94/24/53/PDF/Quotient_Functions.pdf.

[95] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer-Verlag, London, 3rd edition, 2009.

[96] C. Wang, N. Komodakis, and N. Paragios, "Markov random field modeling, inference & learning in computer vision & image understanding: A survey," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1610–1627, Nov. 2013.

[97] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, J. Lellmann, N. Komodakis, and C. Rother, "A comparative study of modern inference techniques for discrete energy minimization problems," in *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 25-27 June 2013, pp. 1328–1335.

[98] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 30, no. 6, pp. 1068–1080, June 2008.

[99] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[100] P. Strandmark, F. Kahl, and T. Schoenemann, "Parallel and distributed vision algorithms using dual decomposition," *Computer Vision and Image Understanding*, vol. 115, no. 12, pp. 1721–1732, 2011.

[101] S. Vicente, V. Kolmogorov, and C. Rother, "Joint optimization of segmentation and appearance models," in *Proc. IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, 29 Sep.-2 Oct. 2009, pp. 755–762.

[102] T. Pock, A. Chambolle, D. Cremers, and H. Bischof, "A convex relaxation approach for computing minimal partitions," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 20-25 June 2009, pp. 810–817.

[103] O. J. Woodford, C. Rother, and V. Kolmogorov, "A global perspective on MAP inference for low-level vision," in *Proc. IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, 27 Sept. - 4 Oct. 2009, pp. 2319–2326.

[104] D. Cremers, P. Thomas, K. Kolev, and A. Chambolle, "Convex relaxation techniques for segmentation, stereo and multiview reconstruction," in *Markov Random Fields for Vision and Image Processing*, A. Blake, P. Kohli, and C. Rother, Eds. The MIT Press, Boston, 2011.

[105] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys, "Joint 3D scene reconstruction and class segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 25-27 June 2013, pp. 97–104.

[106] L. Torresani, V. Kolmogorov, and C. Rother, "A dual decomposition approach to feature correspondence," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 35, no. 2, pp. 259–271, Feb. 2013.

[107] Y Zeng, C. Wang, Y. Wang, X. Gu, D. Samaras, and N. Paragios, "Intrinsic dense 3D surface tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, 21-23 June 2011, pp. 1225–1232.

[108] B. Glocker, N. Paragios, N. Komodakis, G. Tziritas, and N. Navab, "Optical flow estimation with uncertainties through dynamic MRFs," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 24-26 June 2008.

[109] M. P. Kumar and D. Koller, "Efficiently selecting regions for scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 13-18 June 2010, pp. 3217–3224.

[110] N. Komodakis and N. Paragios, "MRF-based blind image deconvolution," in *Proc. Asian Conference on Computer Vision*, Daejeon, Korea, 5-9 Nov. 2012, pp. 361–374.

[111] V. Kolmogorov and A. Shioura, "New algorithms for convex cost tension problem with application to computer vision," *Discrete Optim.*, vol. 6, no. 4, pp. 378–393, 2009.

[112] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich, "Diverse m-best solutions in Markov random fields," in *Proc. European Conference on Computer Vision*, Florence, Italy, 7-13 Oct. 2012, pp. 1–16.

[113] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comp. Vis.*, vol. 100, no. 2, pp. 190–202, Nov. 2012.

[114] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios, "Dense image registration through MRFs and efficient linear programming," *Medical Image Analysis*, vol. 12, no. 6, pp. 731–741, Dec. 2008.

[115] B. Glocker, A. Sotiras, N. Paragios, and N. Komodakis, "Deformable medical image registration: setting the state of the art with discrete methods," *Annual Reviews Biomedical Engineering*, vol. 13, pp. 219–244, Aug. 2011.

[116] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[117] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers, "A convex formulation of continuous multi-label problems," in *Proc. European Conference on Computer Vision*, Marseille, France, 12-18 Oct. 2008, pp. 792–805.

[118] N. Komodakis and N. Paragios, "Beyond pairwise energies: Efficient optimization for higher-order MRFs," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 20-25 June 2009, pp. 2985–2992.

[119] A. Fix, C. Wang, and R. Zabih, "A primal-dual method for higher-order multilabel Markov random fields," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 24-27 June 2014.

[120] C. Arora, S. Banerjee, P. Kalra, and S. N. Maheshwari, "Generic cuts: An efficient algorithm for optimal inference in higher order MRF-MAP," in *Proc. European Conference on Computer Vision*, Florence, Italy, 7-13 Oct. 2012, pp. 17–30.