

# Conditional gradient algorithms for norm-regularized smooth convex optimization

Zaid Harchaoui · Anatoli Juditsky ·  
Arkadi Nemirovski

Received: 28 March 2013 / Accepted: 31 March 2014 / Published online: 18 April 2014  
© Springer-Verlag Berlin Heidelberg and Mathematical Optimization Society 2014

**Abstract** Motivated by some applications in signal processing and machine learning, we consider two convex optimization problems where, given a cone  $K$ , a norm  $\|\cdot\|$  and a smooth convex function  $f$ , we want either (1) to minimize the norm over the intersection of the cone and a level set of  $f$ , or (2) to minimize over the cone the sum of  $f$  and a multiple of the norm. We focus on the case where (a) the dimension of the problem is too large to allow for interior point algorithms, (b)  $\|\cdot\|$  is “too complicated” to allow for computationally cheap Bregman projections required in the first-order proximal gradient algorithms. On the other hand, we assume that it is relatively easy to minimize linear forms over the intersection of  $K$  and the unit  $\|\cdot\|$ -ball. Motivating examples are given by the nuclear norm with  $K$  being the entire space of matrices, or the positive semidefinite cone in the space of symmetric matrices, and the Total Variation norm on the space of 2D images. We discuss versions of the Conditional Gradient algorithm capable to handle our problems of interest, provide the related theoretical efficiency estimates and outline some applications.

---

Research of the first and second authors was supported by the CNRS-Mastodons project GARGANTUA, and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025). Research of the third author was supported by the ONR Grant N000140811104 and NSF Grants DMS 0914785, CMMI 1232623.

---

Z. Harchaoui

LJK, Inria, 655 Avenue de l'Europe, Montbonnot, 38334 Saint-Ismier, France  
e-mail: zaid.harchaoui@inria.fr

A. Juditsky (✉)

LJK, Université Grenoble Alpes, B.P. 53, 38041 Grenoble Cedex 9, France  
e-mail: anatoli.juditsky@imag.fr

A. Nemirovski

Georgia Institute of Technology, Atlanta, GA 30332, USA  
e-mail: nemirovs@isye.gatech.edu

## 1 Introduction

We consider two *norm-regularized* convex optimization problems as follows:

$$[\text{norm minimization}] \min_{x \in K} \|x\|, \quad \text{subject to } f(x) \leq 0, \quad (1)$$

$$[\text{penalized minimization}] \min_{x \in K} f(x) + \kappa \|x\| \quad (2)$$

where  $f$  is a convex function with Lipschitz continuous gradient,  $K$  is a closed convex cone in a Euclidean space  $E$ ,  $\|\cdot\|$  is some norm, and  $\kappa$  is a positive parameter. Problems such as (1) and (2) are of definite interest for signal processing and machine learning. In these applications,  $f(x)$  quantifies the discrepancy between the observed noisy output of some parametric model and the output of the model with candidate vector  $x$  of parameters. Most notably,  $f$  is the quadratic penalty:  $f(x) = \frac{1}{2} \|\mathcal{A}x - y\|_2^2 - \delta$ , where  $\mathcal{A}x$  is the predicted output of the linear regression model  $x \mapsto \mathcal{A}x$ , and  $y = \mathcal{A}x_* + \xi$ , where  $x_*$  is the vector of true parameters,  $\xi$  is the observation error, and  $\delta$  is an a priori upper bound on  $\frac{1}{2} \|\xi\|_2^2$ . The cone  $K$  sums up a priori information on the parameter vectors (e.g.,  $K = E$ —no a priori information at all, or  $E = \mathbf{R}^p$ ,  $K = \mathbf{R}_+^p$ , or  $E = \mathbf{S}^p$ , the space of symmetric  $p \times p$  matrices, and  $K = \mathbf{S}_+^p$ , the cone of positive semidefinite matrices, as is the case of covariance matrices recovery). Finally,  $\|\cdot\|$  is a regularizing norm “promoting” a desired property of the recovery, e.g., the sparsity-promoting norm  $\ell_1$  on  $E = \mathbf{R}^n$ , or the low rank promoting nuclear norm on  $E = \mathbf{R}^{p \times q}$ , or the Total Variation (TV) norm, as in image reconstruction.

In the large-scale case, first-order algorithms of proximal-gradient type are popular to tackle such problems, see [34] for a recent overview. Among them, the celebrated Nesterov optimal gradient methods for smooth and composite minimization [26–28], and their stochastic approximation counterparts [21], are now state-of-the-art in compressive sensing and machine learning. These algorithms enjoy the best known so far theoretical estimates (and in some cases, these estimates are the best possible for the first-order algorithms). For instance, Nesterov’s algorithm for penalized minimization [27, 28] solves (2) to accuracy  $\epsilon$  in  $O(D_0 \sqrt{L/\epsilon})$  iterations, where  $L$  is the properly defined Lipschitz constant of the gradient of  $f$ , and  $D_0$  is the initial distance to the optimal set, measured in the norm  $\|\cdot\|$ . However, applicability and efficiency of proximal-gradient algorithms in the large-scale case require from the problem to possess “favorable geometry” (for details, see [28, Section A.6]). To be more specific, consider proximal-gradient algorithm for convex minimization problems of the form

$$\min_x \{f(x) : \|x\| \leq 1, x \in K\}. \quad (3)$$

The comments to follow, with slight modifications, are applicable to problems such as (1) and (2) as well. In this case, a proximal-gradient algorithm operates with a “distance generating function” (d.g.f.) defined on the domain of the problem and 1-strongly convex w.r.t. the norm  $\|\cdot\|$ . Each step of the algorithm requires minimizing the sum of the d.g.f. and a linear form. The efficiency estimate of the algorithm depends on

the variation of the d.g.f. on the domain and on regularity of  $f$  w.r.t.  $\|\cdot\|$ .<sup>1</sup> As a result, in order for a proximal-gradient algorithm to be practical in the large scale case, two “favorable geometry” conditions should be met: (a) the outlined sub-problems should be easy to solve, and (b) the variation of the d.g.f. on the domain of the problem should grow slowly (if at all) with problem’s dimension. Both these conditions indeed are met in many applications; see, e.g., [2,20] for examples. This explains the recent popularity of this family of algorithms.

However, sometimes conditions (a) and/or (b) are violated, and application of proximal algorithms becomes questionable. For example, for the case of  $K = E$ , (b) is violated for the usual  $\|\cdot\|_\infty$ -norm on  $\mathbf{R}^p$  or, more generally, for  $\|\cdot\|_{2 \rightarrow \infty}$  norm on the space of  $p \times q$  matrices given by

$$\|x\|_{2 \rightarrow \infty} := \max_{u \in \mathbf{R}^q: \|u\|_2 \leq 1} \|xu\|_\infty = \max_{1 \leq j \leq p} \|\text{Row}_j(x)\|_2,$$

where  $\text{Row}_j^T(x)$  denotes the  $j$ th row of  $x$ . Here the variation of (any) d.g.f. on problem’s domain is at least  $p$ . As a result, in the case in question the theoretical iteration complexity of a proximal algorithm grows rapidly with the dimension  $p$ . Furthermore, for some high-dimensional problems which do satisfy (b), solving the sub-problem can be computationally challenging. Examples of such problems include nuclear-norm-based matrix completion, Total Variation-based image reconstruction, and multi-task learning with a large number of tasks and features. This corresponds to  $\|\cdot\|$  in (1) or (2) being the nuclear norm [11, 14] or the TV-norm.

These limitations recently motivated alternative approaches, which do not rely upon favorable geometry of the problem domain and/or do not require to solve hard sub-problems at each iteration, and triggered a renewed interest in the Conditional Gradient (CndG) algorithm. This algorithm, also known as the Frank–Wolfe algorithm, which is historically the first method for smooth constrained convex optimization, originates from [9], and was extensively studied in the 70-s (see, e.g., [6, 8, 29] and references therein). CndG algorithms work by minimizing a linear form on the problem domain at each iteration; this auxiliary problem clearly is easier, and in many cases—significantly easier than the auxiliary problem arising in proximal-gradient algorithms. Conditional gradient algorithms for collaborative filtering were studied recently [17, 18], some variants and extensions were studied in [7, 11, 33]. Those works consider constrained formulations of machine learning or signal processing problems, i.e., minimizing the discrepancy  $f(x)$  under a constraint on the norm of the solution, as in (3). On the other hand, CndG algorithms for other learning formulations, such as norm minimization (1) or penalized minimization (2) remain open issues. An exception is the work of [7, 11], where a Conditional Gradient algorithm for penalized minimization was studied, although the efficiency estimates obtained in that paper were suboptimal. In this paper (for its conference version, see [12]) we present CndG-type algorithms aimed

<sup>1</sup> i.e., the Lipschitz constant of  $f$  w.r.t.  $\|\cdot\|$  in the nonsmooth case, or the Lipschitz constant of the gradient mapping  $x \mapsto f'(x)$  w.r.t. the norm  $\|\cdot\|$  on the argument and the conjugate of this norm on the image spaces in the smooth case.

at solving norm minimization and penalized minimization problems<sup>2</sup> and provide theoretical efficiency guarantees for these algorithms.

The main body of the paper is organized as follows. In Sect. 2, we present detailed setting of problems (1), (2) along with basic assumptions on the “computational environment” required by the CndG-based algorithms we are developing. These algorithms and their efficiency bounds are presented in Sect. 3 (problem (1)) and 5 (problem (2)). In Sect. 6 we outline some applications, and in Sect. 7 present preliminary numerical results. All proofs are relegated to the “Appendix”.

## 2 Problem statement

Throughout the paper, we shall assume that  $K \subset E$  is a closed convex cone in Euclidean space  $E$ ; we loose nothing by assuming that  $K$  linearly spans  $E$ . We assume, further, that  $\|\cdot\|$  is a norm on  $E$ , and  $f : K \rightarrow \mathbf{R}$  is a convex function with Lipschitz continuous gradient, so that

$$\|f'(x) - f'(y)\|_* \leq L_f \|x - y\| \quad \forall x, y \in K,$$

where  $\|\cdot\|_*$  denotes the norm dual to  $\|\cdot\|$ <sup>3</sup>

$$\forall x, y \in K : f(y) \leq f(x) + \langle f'(x), y - x \rangle + \frac{L_f}{2} \|y - x\|^2. \quad (4)$$

We consider two kinds of problems, detailed below.

*Norm-minimization.* Such problems correspond to

$$\rho_* = \min_x \{\|x\| : x \in K, f(x) \leq 0\}. \quad (5)$$

To tackle (5), we consider the following parametric family of problems

$$\text{Opt}(\rho) = \min\{f(x) : \|x\| \leq \rho, x \in K\}. \quad (6)$$

Note that whenever (5) is feasible, which we assume from now on, we have

$$\rho_* = \min\{\rho \geq 0 : \text{Opt}(\rho) \leq 0\}, \quad (7)$$

and both problems (5), (7) can be solved.

Given a tolerance  $\epsilon > 0$ , we want to find an  $\epsilon$ -solution to the problem, that is,  $x_\epsilon \in K$  such that

$$\|x_\epsilon\| \leq \rho_* \text{ and } f(x_\epsilon) \leq \epsilon. \quad (8)$$

<sup>2</sup> The penalized minimization problem, with a similar type of algorithm, was independently of and simultaneously with [12] considered in [38].

<sup>3</sup> Recall that the dual, a.k.a. conjugate, to a norm  $\|\cdot\|$  on a Euclidean space  $E$  is the norm on  $E$  defined as  $\|\xi\|_* = \max_{x \in E, \|x\| \leq 1} \langle \xi, x \rangle$ , where  $\langle \cdot, \cdot \rangle$  is the inner product on  $E$ .

Getting back to the problem of interest (5),  $x_\epsilon$  is then “super-optimal” and  $\epsilon$ -feasible.

*Penalized norm minimization.* These problems write as

$$\text{Opt} = \min_x \{f(x) + \kappa \|x\| : x \in K\} . \tag{9}$$

A equivalent formulation is

$$\text{Opt} = \min_{x,r} \{F([x; r]) = \kappa r + f(x) : x \in K, \|x\| \leq r\} . \tag{10}$$

We shall refer to (10) as the problem of *composite optimization* (CO). Given a tolerance  $\epsilon > 0$ , our goal is to find an  $\epsilon$ -solution to (10), defined as a feasible solution  $(x_\epsilon, r_\epsilon)$  to the problem satisfying  $F([x_\epsilon; r_\epsilon]) - \text{Opt} \leq \epsilon$ . Note that in this case  $x_\epsilon$  is an  $\epsilon$ -solution, in the similar sense, to (9).

*Special case.* In many applications where problems (5) and (9) arise, the function  $f$  enjoys a special structure:

$$f(x) = \phi(\mathcal{A}x - b),$$

where  $x \mapsto \mathcal{A}x - b$  is an affine mapping from  $E$  to  $\mathbf{R}^m$ , and  $\phi(\cdot) : \mathbf{R}^m \rightarrow \mathbf{R}$  is a convex function with Lipschitz continuous gradient; we shall refer to this situation as to *special case*. In such case, the quantity  $L_f$  can be bounded as follows. Let  $\pi(\cdot)$  be some norm on  $\mathbf{R}^m$ ,  $\pi_*(\cdot)$  be the conjugate norm, and  $\|\mathcal{A}\|_{\|\cdot\|, \pi}$  be the norm of the linear mapping  $x \mapsto \mathcal{A}x$  induced by the norms  $\|\cdot\|, \pi(\cdot)$  on the argument and the image spaces:

$$\|\mathcal{A}\|_{\|\cdot\|, \pi(\cdot)} = \max_{x \in E} \{\pi(\mathcal{A}x) : \|x\| \leq 1\}.$$

Let also  $L_{\pi(\cdot)}[\phi]$  be the Lipschitz constant of the gradient of  $\phi$  induced by the norm  $\pi(\cdot)$ , so that

$$\pi_*(\phi'(y) - \phi'(y')) \leq L_{\pi(\cdot)}[\phi] \pi(y - y') \quad \forall y, y' \in \mathbf{R}^m .$$

Then, one can take as  $L_f$  the quantity

$$L_f = L_{\pi(\cdot)}[\phi] \|\mathcal{A}\|_{\|\cdot\|, \pi(\cdot)}^2 . \tag{11}$$

*Example 1 Quadratic fit.* In many applications, we are interested in  $\|\cdot\|_2$ -discrepancy between  $\mathcal{A}x$  and  $b$ ; the related choice of  $\phi(\cdot)$  is  $\phi(y) = \frac{1}{2}y^T y$ . Specifying  $\pi(\cdot)$  as  $\|\cdot\|_2$ , we get  $L_{\|\cdot\|_2}[\phi] = 1$ .

*Example 2 Smoothed  $\ell_\infty$  fit.* When interested in  $\|\cdot\|_\infty$  discrepancy between  $\mathcal{A}x$  and  $b$ , we can use as  $\phi$  the function  $\phi(y) = \frac{1}{2}\|y\|_\beta^2$ , where  $\beta \in [2, \infty)$ . Taking  $\pi(\cdot)$  as  $\|\cdot\|_\infty$ , we get

$$L_{\|\cdot\|_\infty}[\phi] \leq (\beta - 1)m^{2/\beta}.$$

Note that

$$\frac{1}{2}\|y\|_\infty^2 \leq \phi(y) \leq \frac{m^{2/\beta}}{2}\|y\|_\infty^2,$$

so that for  $\beta = O(1) \ln(m)$  and  $m$  large enough (specifically, such that  $\beta \geq 2$ ),  $\phi(y)$  is within absolute constant factor of  $\frac{1}{2}\|y\|_\infty^2$ . The latter situation can be interpreted as  $\phi$  behaving as  $\frac{1}{2}\|\cdot\|_\infty^2$ . At the same time, with  $\beta = O(1) \ln(m)$ ,  $L_{\|\cdot\|_\infty}[\phi] \leq O(1) \ln(m)$  grows with  $m$  logarithmically.

Another widely used choice of  $\phi(\cdot)$  for this type of discrepancy is “logistic” function

$$\phi(y) = \frac{1}{\beta} \ln \left( \sum_{i=1}^m [e^{\beta y_i} + e^{-\beta y_i}] \right).$$

For  $\pi(\cdot) = \|\cdot\|_\infty$  we easily compute  $L_{\|\cdot\|_\infty}[\phi] \leq \beta$  and  $\|y\|_\infty \leq \phi(y) \leq \|y\|_\infty + \ln(2m)/\beta$ .

Note that in some applications we are interested in “one-sided” discrepancies quantifying the magnitude of the vector  $[Ax - b]_+ = [\max[0, (Ax - b)_1]; \dots; \max[0, (Ax - b)_m]]$  rather than the the magnitude of the vector  $Ax - b$  itself. Here, instead of using  $\phi(y) = \frac{1}{2}\|y\|_\beta^2$  in the context of examples 1 and 2, one can use the functions  $\phi_+(y) = \phi([y]_+)$ . In this case the bounds on  $L_{\pi(\cdot)}[\phi_+]$  are exactly the same as the above bounds on  $L_{\pi(\cdot)}[\phi]$ . The obvious substitute for the two-sided logistic function is its “one-sided version:”  $\phi_+(y) = \frac{1}{\beta} \ln \left( \sum_{i=1}^m [e^{\beta y_i} + 1] \right)$  which obeys the same bound for  $L_{\pi(\cdot)}[\phi_+]$  as its two-sided analogue.

*First-order and linear optimization oracles.* We assume that  $f$  is represented by a *first-order oracle*—a routine which, given on input a point  $x \in K$ , returns the value  $f(x)$  and the gradient  $f'(x)$  of  $f$  at  $x$ . As about  $K$  and  $\|\cdot\|$ , we assume that they are given by a *Linear Optimization (LO) oracle* which, given on input a linear form  $\langle \eta, \cdot \rangle$  on  $E$ , returns a minimizer  $x[\eta]$  of this linear form on the set  $\{x \in K : \|x\| \leq 1\}$ . We assume w.l.o.g. that for every  $\eta$ ,  $x[\eta]$  is either zero, or is a vector of the  $\|\cdot\|$ -norm equal to 1. To ensure this property, it suffices to compute  $\langle \eta, x[\eta] \rangle$  for  $x[\eta]$  given by the oracle; if this inner product is 0, we can reset  $x[\eta] = 0$ , otherwise  $\|x[\eta]\|$  is automatically equal to 1.

Note that an LO oracle for  $K$  and  $\|\cdot\|$  allows to find a minimizer of a linear form of  $z = [x; r] \in E^+ := E \times \mathbf{R}$  on a set of the form  $K^+[\rho] = \{[x; r] \in E^+ : x \in K, \|x\| \leq r \leq \rho\}$  due to the following observation:

**Lemma 1** *Let  $\rho \geq 0$  and  $\eta^+ = [\eta; \sigma] \in E^+$ . Consider the linear form  $\ell(z) = \langle \eta^+, z \rangle$  of  $z = [x; r] \in E^+$ , and let*

$$z^+ = \begin{cases} \rho[x[\eta]; 1], & \langle \eta^+, [x[\eta]; 1] \rangle \leq 0, \\ 0, & \text{otherwise} \end{cases}.$$

Then  $z^+$  is a minimizer of  $\ell(z)$  over  $z \in K^+[\rho]$ . When  $\sigma = 0$ , one has  $z^+ = \rho[x[\eta]; 1]$ .

Indeed, by the definition of  $x[\eta]$ , for all  $z = [x; r] \in K^+[\rho]$  we have  $\ell(r[x[\eta]; 1]) \leq \ell(z)$ . Now, let  $z_* = [x_*; r_*]$  be a minimizer of  $\ell(\cdot)$  over  $K^+[\rho]$ , then  $\ell(z_*) \leq \ell(z)$ , where  $z_* := r_*[x[\eta]; 1] \in K^+[\rho]$ . We conclude that any minimizer of  $\ell(\cdot)$  over the segment  $\{s[x[\eta]; 1] : 0 \leq s \leq \rho\}$  is also a minimizer of  $\ell(\cdot)$  over  $K^+[\rho]$ . It remains to note that the vector  $z^+$  in the lemma premise clearly is a minimizer of  $\ell(\cdot)$  on the above segment.

### 3 Conditional gradient algorithm

In this section, we present an overview of the properties of the standard Conditional Gradient algorithm, and highlight some memory-based extensions. These properties are not new. However, since they are key for the design of our proposed algorithms in the next sections, we present them for further reference.

#### 3.1 Conditional gradient algorithm

Let  $E$  be a Euclidean space and  $X$  be a closed and bounded convex set in  $E$  which linearly spans  $E$ . Assume that  $X$  is given by a LO oracle—a routine which, given on input  $\eta \in E$ , returns an optimal solution  $x_X[\eta]$  to the optimization problem

$$\min_{x \in X} \langle \eta, x \rangle$$

(cf. Sect. 2). Let  $f$  be a convex differentiable function on  $X$  with Lipschitz continuous gradient  $f'(x)$ , so that

$$\forall x, y \in X : f(y) \leq f(x) + \langle f'(x), y - x \rangle + \frac{1}{2}L\|y - x\|_X^2, \tag{12}$$

where  $\|\cdot\|_X$  is the norm on  $E$  with the unit ball  $X - X$ .<sup>4</sup> We intend to solve the problem

$$f_* = \min_{x \in X} f(x). \tag{13}$$

A generic CndG algorithm is a recurrence which builds iterates  $x_t \in X, t = 1, 2, \dots$ , in such a way that

$$f(x_{t+1}) \leq f(\tilde{x}_{t+1}), \tag{14}$$

where

$$\tilde{x}_{t+1} = x_t + \gamma_t[x_t^+ - x_t], \quad \text{with } x_t^+ = x_X[f'(x_t)] \quad \text{and } \gamma_t = \frac{2}{t+1}. \tag{15}$$

<sup>4</sup> Note being a special case of (4), (12) provides a description of regularity properties of  $f$  in terms which are “invariant with respect to the geometry of  $X$ ”.

As a byproduct of running generic CndG, after  $t$  steps we have at our disposal the quantities

$$\begin{aligned} f_{*,\tau} &= \min_{x \in X} [f(x_\tau) + \langle f'(x_\tau), x - x_\tau \rangle] \\ &= f(x_\tau) - \langle f'(x_\tau), x_\tau - x_X[f'(x_\tau)] \rangle, \quad 1 \leq \tau \leq t, \end{aligned} \quad (16)$$

which, by convexity of  $f$ , are lower bounds on  $f_*$ . Consequently, at the end of step  $t$  we have at our disposal a lower bound

$$f_*^t := \max_{1 \leq \tau \leq t} f_{*,\tau} \leq f_*, \quad t = 1, 2, \dots \quad (17)$$

on  $f_*$ .

Finally, we define the approximate solution  $\bar{x}_t$  found in course of  $t = 1, 2, \dots$  steps as the best—with the smallest value of  $f$ —of the points  $x_1, \dots, x_t$ . Note that  $\bar{x}_t \in X$ .

The following statement summarizes the well known properties of CndG [6, 8, 17, 29] (to make the presentation self-contained, we provide in “Appendix” the proof).

**Theorem 1** *For a generic CndG algorithm we have*

$$f(\bar{x}_t) - f_* \leq f(\bar{x}_t) - f_*^t \leq f(x_t) - f_*^t \leq \frac{2L}{t+1}, \quad t \geq 2. \quad (18)$$

Some remarks regarding the conditional algorithm are in order.

*Certifying quality of approximate solutions.* An attractive property of CndG is the presence of online lower bound  $f_*^t$  on  $f_*$  which certifies the theoretical rate of convergence of the algorithm. This accuracy certificate, first established in a different form in [6], see also [17], also provides a valuable stopping criterion when running the algorithm in practice.

*CndG algorithm with memory.* When computing the next search point  $x_{t+1}$ , the simplest CndG algorithm defined by the recursion  $x_{t+1} = x_t + \frac{2}{2+1}[x_t^+ - x_t]$  only uses the latest answer  $x_t^+ = x_X[f'(x_t)]$  of the LO oracle. Meanwhile, the CndG algorithm can be modified to make use of information supplied by previous oracle calls; we refer to this modification as CndG with memory (CndGM).<sup>5</sup> Assume that we have already carried out  $t - 1$  steps of the algorithm and have at our disposal current iterate  $x_t \in X$  (with  $x_1$  selected as an arbitrary point of  $X$ ) along with previous iterates  $x_\tau$ ,  $\tau < t$  and the vectors  $f'(x_\tau)$ ,  $x_\tau^+ = x_X[f'(x_\tau)]$ . At the step, we compute  $f'(x_t)$  and  $x_t^+ = x_X[f'(x_t)]$ . Thus, at this point in time we have at our disposal  $2t$  points  $x_\tau$ ,  $x_\tau^+$ ,  $1 \leq \tau \leq t$ , which belong to  $X$ . Let  $X_t$  be subset of these points, with the only restriction that the points  $x_t$ ,  $x_t^+$  are selected, and let us define the next iterate  $x_{t+1}$  as

<sup>5</sup> Note that in the context of “classical” Frank–Wolfe algorithm—minimization of a smooth function over a polyhedral set—such modification is referred to as *Restricted Simplicial Decomposition* [15, 16, 36].



$$x_{t+1} \in \underset{x \in \text{Conv}(X_t)}{\text{Argmin}} f(x), \tag{19}$$

that is,

$$x_{t+1} = \sum_{x \in X_t} \lambda_x^t x, \quad \lambda^t \in \underset{\lambda' = \{\lambda_x\}_{x \in X_t}}{\text{Argmin}} \left\{ f \left( \sum_{x \in X_t} \lambda_x x \right) : \lambda \geq 0, \sum_{x \in X_t} \lambda_x = 1 \right\}. \tag{20}$$

Clearly, it is again a generic CndG algorithm, so that conclusions in Theorem 1 are fully applicable to CndGM. Note that the “classical” CndG algorithm per se is nothing but CndGM with  $X_t = \{x_t, x_t^+\}$  and  $M = 2$  for all  $t$ .

In the above description of CndGM we assumed that the set  $X_t$  is a subset of points  $x_\tau, x_\tau^+, 1 \leq \tau \leq t$ , containing  $x_t$  and  $x_t^+$ . Note that we can try to improve the accuracy of the algorithm by adding to  $X_t$ , along with  $x_t, x_t^+$ , some other points of  $X$ . Recall, that after  $t$  iterations of the method we have at our disposal, along with  $t$  iterates  $x_\tau \in X$ , the corresponding values  $f(x_\tau)$  and gradients  $f'(x_\tau), 1 \leq \tau \leq t$ . Let  $X_t^b$  be a subset of  $\{x_1, \dots, x_t\}$  of cardinality  $m$ , and let us define  $x_t^b$  as a minimizer

$$x_t^b \in \underset{u \in X}{\text{Argmin}} \left\{ f_t^b(u) := \max_{x \in X_t^b} [f(x) + \langle f'(x), u - x \rangle] \right\},$$

of the current “bundle approximation”  $f_t^b$  of  $f$ ; note that the quantity  $f_t^b(x_t^b)$  bounds from below the optimal value  $f_*$  of (13). One can include the point  $x_t^b$  into  $X_t$  and replace the current lower bound  $f_t^*$  on  $f_*$  with  $\max\{f_t^*, f_t^b(x_t^b)\}$ .<sup>6</sup>

It remains to notice that if the cardinality  $m$  of the “bundle” is small the price of computing  $x_t^b$  is comparable to that of  $x_t^+$ . Indeed, when denoting  $x_{\tau_1}, \dots, x_{\tau_m}$  the elements of  $X_t^b$  and  $\Lambda = \{\lambda \in \mathbf{R}^m : \lambda_i \geq 0, \sum_{i=1}^m \lambda_i = 1\}$ , one has

---

<sup>6</sup> In fact,  $x_t^b$  may be substituted for  $x_t^+$  in the recurrence (15), and the resulting approximate solutions  $x_t$  along with the lower bounds  $f_t^* \geq \max_{1 \leq \tau \leq t} f_\tau^b(x_\tau^b)$  will still satisfy the bound (18) of Theorem 1. Indeed, the analysis of the proof of the theorem reveals that a point  $\xi \in X$  can be substituted for  $x_t^+ = x_X[f'(x_t)]$  as soon as the inequality

$$\langle f'(x_t), x_t - \xi \rangle \geq f(x_t) - f_t^*, \tag{21}$$

holds, where  $f_t^*$  is the best currently available lower bound for the optimal value  $f_*$  of (13). In other words, for the result of Theorem 1 to hold, one can substitute  $x_t^+ = x_X[f'(x_t)]$  with any vector  $\xi$  satisfying (21). Now assume that  $x_t \in X_t^b$ . By convexity of  $f_t^b$  (obviously,  $f'(x_t) \in \partial f_t^b(x_t)$ , where  $\partial f_t^b(x)$  is the subdifferential of  $f_t^b$  at  $x$ ), we have

$$\langle f'(x_t), x_t - x_t^b \rangle \geq f_t^b(x_t^b) - f(x_t) \geq f_t^* - f(x_t),$$

what is (21).

$$\begin{aligned} \min_{u \in X} f_t^b(u) &= \min_{u \in X} \left\{ \max_{\lambda \in \Lambda} \sum_{i=1}^m \lambda_i [f(x_{\tau_i}) + \langle f'(x_{\tau_i}), u - x_{\tau_i} \rangle] \right\} \\ &= \max_{\lambda \in \Lambda} \left\{ F(\lambda) := \min_{u \in X} \sum_{i=1}^m \lambda_i [f(x_{\tau_i}) + \langle f'(x_{\tau_i}), u - x_{\tau_i} \rangle] \right\}. \end{aligned}$$

When  $m$  is small, one can find high accuracy approximate solutions  $\widehat{\lambda}$  to the problem of maximizing  $F$  over the standard simplex using (linearly converging) cutting plane algorithms, for which LO oracle will readily supply first order information. Then corresponding “primal” approximate solutions  $\widehat{x}_t^b$  can be recovered using dual certificates, as described in [5, 24].

*CndGM: implementation issues.* Assume that the cardinalities of the sets  $X_t$  in CndGM are bounded by some  $M \geq 2$ . In this case, implementation of the method requires solving at every step an auxiliary problem (20) of minimizing over the standard simplex of dimension  $\leq M - 1$  a smooth convex function given by a first-order oracle induced by the first-oracle for  $f$ . When  $M$  is a once for ever fixed small integer, the arithmetic cost of solving this problem within machine accuracy by, say, the Ellipsoid algorithm is dominated by the arithmetic cost of just  $O(1)$  calls to the first-order oracle for  $f$ . Thus, CndGM with small  $M$  can be considered as implementable.<sup>7</sup>

Note that in the special case (Sect. 2), where  $f(x) = \phi(Ax - b)$ , assuming  $\phi(\cdot)$  and  $\phi'(\cdot)$  easy to compute, as is the case in most of the applications, the first-order oracle for the auxiliary problems arising in CndGM becomes cheap (cf. [39]). Indeed, in this case (20) reads

$$\min_{\lambda^t} \left\{ g_t(\lambda^t) := \phi \left( \sum_{x \in X_t} \lambda_x^t Ax - b \right) : \begin{array}{l} \lambda^t = \{\lambda_x^t\}_{x \in X_t} \geq 0, \\ \sum_{x \in X_t} \lambda_x^t = 1 \end{array} \right\}. \quad (22)$$

It follows that all we need to get a computationally cheap access to the first-order information on  $g_t(\lambda^t)$  for all values of  $\lambda^t$  is to have at our disposal the matrix-vector products  $Ax$ ,  $x \in X_t$ . With our construction of  $X_t$ , the only two “new” elements in  $X_t$  (those which were not available at preceding iterations) are  $x_t$  and  $x_t^+$ , so that the only two new matrix-vector products we need to compute at iteration  $t$  are  $Ax_t$  (which usually is a byproduct of computing  $f'(x_t)$ ) and  $Ax_t^+$ . Thus, we can say that the “computational overhead,” as compared to computing  $f'(x_t)$  and  $x_t^+ = x_X[f'(x_t)]$ , needed to get easy access to the first-order information on  $g_t(\cdot)$  reduces to computing the single matrix-vector product  $Ax_t^+$ .

<sup>7</sup> Assuming possibility to solve (20) exactly, while being idealization, is basically as “tolerable” as the standard in continuous optimization assumption that one can use exact real arithmetic or compute exactly eigenvalues/eigenvectors of symmetric matrices. The outlined “real life” considerations can be replaced with rigorous error analysis which shows that in order to maintain the efficiency estimates from Theorem 1, it suffices to solve  $t$ -th auxiliary problem within properly selected positive inaccuracy, and this can be achieved in  $O(\ln(t))$  computations of  $f$  and  $f'$ .

### 4 Conditional gradient algorithm for parametric optimization

In this section, we describe a multi-stage algorithm to solve the parametric optimization problem (6), (7), using the conditional algorithm to solve inner sub-problems (6), (7). The idea, originating from [22] (see also [19,26,28]), is to use a Newton-type method for approximating from below the positive root  $\rho_*$  of  $\text{Opt}(\rho)$ , with (inexact) first-order information on  $\text{Opt}(\cdot)$  yielded by approximate solving the optimization problems defining  $\text{Opt}(\cdot)$ ; the difference with the outlined references is that now we solve these problems with the CndG algorithm.

Our algorithm works stagewise. At the beginning of stage  $s = 1, 2, \dots$ , we have at hand a lower bound  $\rho_s$  on  $\rho_*$ , with  $\rho_1$  defined as follows:

We compute  $f(0)$ ,  $f'(0)$  and  $x[f'(0)]$ . If  $f(0) \leq \epsilon$  or  $x[f'(0)] = 0$ , we are done—the pair  $(\rho = 0, x = 0)$  is an  $\epsilon$ -solution to (7) in the first case, and is an optimal solution to the problem in the second case (since in the latter case 0 is a minimizer of  $f$  on  $K$ , and (7) is feasible). Assume from now on that the above options do not take place (“nontrivial case”), and let

$$d = -\langle f'(0), x[f'(0)] \rangle.$$

Due to the origin of  $x[\cdot]$ ,  $d$  is positive, and  $f(x) \geq f(0) + \langle f'(0), x \rangle \geq f(0) - d\|x\|$  for all  $x \in K$ , which implies that  $\rho_* \geq \rho_1 := \frac{f(0)}{d} > 0$ .

At stage  $s$  we apply a generic CndG algorithm to the auxiliary problem

$$\text{Opt}(\rho_s) = \min_x \{f(x) : x \in K[\rho_s]\}, \quad K[\rho] = \{x \in K : \|x\| \leq \rho\}, \quad (23)$$

Note that the LO oracle for  $K, \|\cdot\|$  induces an LO oracle for  $K[\rho]$ ; specifically, for every  $\eta \in E$ , the point  $x_\rho[\eta] := \rho x[\eta]$  is a minimizer of the linear form  $\langle \eta, x \rangle$  over  $x \in K[\rho]$ , see Lemma 1.  $x_\rho[\cdot]$  is exactly the LO oracle utilized by CndG as applied to (23).

As explained above, after  $t$  steps of CndG as applied to (23), the iterates being  $x_\tau \in K[\rho_s], 1 \leq \tau \leq t$ ,<sup>8</sup> we have at our disposal current approximate solution  $\bar{x}_t \in \{x_1, \dots, x_t\}$  such that  $f(\bar{x}_t) = \min_{1 \leq \tau \leq t} f(x_\tau)$  along with a lower bound  $f_*^t$  on  $\text{Opt}(\rho_s)$ . Our policy is as follows.

1. When  $f(\bar{x}_t) \leq \epsilon$ , we terminate the solution process and output  $\bar{\rho} = \rho_s$  and  $\bar{x} = \bar{x}_t$ ;
2. When the above option is not met and  $f_*^t < \frac{3}{4} f(\bar{x}_t)$ , we specify  $x_{t+1}$  according to the description of CndG and pass to step  $t + 1$  of stage  $s$ ;
3. Finally, when neither one of the above options takes place, we terminate stage  $s$  and pass to stage  $s + 1$ , specifying  $\rho_{s+1}$  as follows:

<sup>8</sup> The iterates  $x_t$ , same as other indexed by  $t$  quantities participating in the description of the algorithm, in fact depend on both  $t$  and the stage number  $s$ . To avoid cumbersome notation when speaking about a particular stage, we suppress  $s$  in the notation.

We are in the situation  $f(\bar{x}_t) > \epsilon$  and  $f_*^t \geq \frac{3}{4}f(\bar{x}_t)$ . Now, for  $\tau \leq t$  the quantities  $f(x_\tau)$ ,  $f'(x_\tau)$  and  $x[f'(x_\tau)]$  define affine function of  $\rho \geq 0$

$$\ell_\tau(\rho) = f(x_\tau) + \langle f'(x_\tau), \rho x[f'(x_\tau)] - x_\tau \rangle.$$

By Lemma 1 we have for every  $\rho \geq 0$

$$\ell_\tau(\rho) = \min_{x \in K[\rho]} [f(x_\tau) + \langle f'(x_\tau), x - x_\tau \rangle] \leq \min_{x \in K[\rho]} f(x) = \text{Opt}(\rho),$$

where the inequality is due to the convexity of  $f$ . Thus,  $\ell_\tau(\rho)$  is an affine in  $\rho \geq 0$  lower bound on  $\text{Opt}(\rho)$ , and we lose nothing by assuming that all these univariate affine functions are memorized when running CndG on (23). Note that by construction of the lower bound  $f_*^t$  (see (16), (17) and take into account that we are in the case of  $X = K[\rho_s]$ ,  $x_X[\eta] = \rho_s x[\eta]$ ) we have

$$f_*^t = \ell^t(\rho_s), \quad \ell^t(\rho) = \max_{1 \leq \tau \leq t} \ell_\tau(\rho).$$

Note that  $\ell^t(\rho)$  is a lower bound on  $\text{Opt}(\rho)$ , so that  $\ell^t(\rho) \leq 0$  for  $\rho \geq \rho_*$ , while  $\ell^t(\rho_s) = f_*^t$  is positive. It follows that

$$r^t := \min \{ \rho : \ell^t(\rho) \leq 0 \}$$

is well defined and satisfies  $\rho_s < r^t \leq \rho_*$ . We compute  $r^t$  (which is easy) and pass to stage  $s + 1$ , setting  $\rho_{s+1} = r^t$  and selecting, as the first iterate of the new stage, any point known to belong to  $K[\rho]$  (e.g., the origin, or  $\bar{x}_t$ ). The first iterate of the first stage is 0.

The description of the algorithm is complete.

The complexity properties of the algorithm are given by the following statement (proved in Sect. 8.2)

**Theorem 2** *When solving a PO problem (6), (7) by the outlined algorithm,*

- (i) *the algorithm terminates with an  $\epsilon$ -solution, as defined in Sect. 2 [cf. (8)];*
- (ii) *The number  $N_s$  of steps at every stage  $s$  of the method admits the bound*

$$N_s \leq \max \left[ 6, \frac{72\rho_*^2 L_f}{\epsilon} + 3 \right].$$

- (iii) *The number of stages before termination does not exceed the quantity*

$$\max \left[ 1.2 \ln \left( \frac{f(0) + \frac{1}{2}L_f \rho_*^2}{\epsilon^2} \right) + 2.4, 3 \right].$$

*Comment.* A straightforward alternative to the proposed approach is obtained by using bisection instead of the Newton-type algorithm when localizing the smallest root  $\rho_*$  of  $\text{Opt}(\rho)$ . Though the theoretical complexity bound for bisection-based algorithm is, essentially, the same as that in Theorem 2, one would expect the Newton scheme to significantly outperform the bisection “in practice”. The reasons for this are twofold. First, we can hope for quadratic, rather than linear, convergence of the Newton root finding. Second, and perhaps more important, with Newton-type algorithm, we all the time are “super-optimal” in terms of the objective of the problem  $\rho_* = \min\{\rho : \text{Opt}(\rho) \leq 0\}$ , that is, the approximates  $\rho_s$  are  $\leq \rho_*$ , and all we need is to become “ $\epsilon$ -feasible”—to ensure  $\text{Opt}(\rho_s) \leq \epsilon$ . Numerical experience shows that when  $\text{Opt}(\rho_s) = \min_{\|x\| \leq \rho_s} f(x)$  is “large,” the algorithm rapidly recognizes this situation, and usually only last 2-3 stages of the algorithm are “long”—require evaluating  $\text{Opt}(\rho_s)$  within accuracy of order of  $\epsilon$ . In contrast to this, bisection “works on both sides of  $\rho_*$ ” and should recognize the fact that the current approximate  $\rho_s$  is “essentially larger” than  $\rho_*$  (this indeed will be the case at some of the bisection steps). When  $\text{Opt}(\rho_s)$  is only marginally negative, recognizing the situation in question requires to evaluate  $\text{Opt}(\rho_s)$  within accuracy  $\epsilon$ , so that we should expect more “long” iterations. This “common sense” explanation is fully supported by numerical experimentation: in dedicated experiments with relatively small matrix completion problems (problem (32) in Sect. 7.1) Newton-based algorithm consistently outperformed the bisection-based one, both in terms of the iteration count and in terms of the running time (e.g., by factor of 3 for matrices of size  $2,000 \times 2,000$ ).

### 5 Conditional gradient algorithm for composite optimization

In this section, we present a modification of the CndG algorithm capable of solving composite minimization problem (10). We assume in the sequel that  $\|\cdot\|, K$  are represented by an LO oracle for the set  $\{x \in K : \|x\| \leq 1\}$ , and  $f$  is given by a first order oracle. In order to apply CndG to the composite optimization problem (10), we make the assumption as follows:

**Assumption A:** There exists  $D < \infty$  such that  $\kappa r + f(x) \leq f(0)$  together with  $\|x\| \leq r, x \in K$ , imply that  $r \leq D$ .

We define  $D_*$  as the minimal value of  $D$  satisfying Assumption A, and assume that we have at our disposal a finite upper bound  $\bar{D}$  on  $D_*$ . An important property of the algorithm we are about to develop is that its efficiency estimate depends on the induced by problem’s data quantity  $D_*$ , and is independent of our a priori upper bound  $\bar{D}$  on this quantity, see Theorem 3 below.

*The algorithm.* We are about to present an algorithm for solving (10). Let  $E^+ = E \times \mathbf{R}$ , and  $K^+ = \{[x; r] : x \in K, \|x\| \leq r\}$ . From now on, for a point  $z = [x; r] \in E^+$  we set  $x(z) = x$  and  $r(z) = r$ . Given  $z = [x; r] \in K^+$ , let us consider the segment

$$\Delta(z) = \{\rho[x; f'(x)]; 1 : 0 \leq \rho \leq \bar{D}\}.$$

and the linear form

$$\zeta = [\xi; \tau] \rightarrow \langle f'(x), \xi \rangle + \kappa \tau = \langle F'(z), \zeta \rangle$$

(here  $F([x; r]) = \kappa r + f(x)$ , as defined in (10)). Observe that by Lemma 1, for every  $0 \leq \rho \leq \bar{D}$ , the minimum of this form on  $K^+[\rho] = \{[x; r] \in E^+, x \in K, \|x\| \leq r \leq \rho\}$  is attained at a point of  $\Delta(z)$  (either at  $[\rho x[f'(x)]; \rho]$  or at the origin). A generic Conditional Gradient algorithm for composite optimization (COCndG) is a recurrence which builds the points  $z_t = [x_t; r_t] \in K^+, t = 1, 2, \dots$ , in such a way that

$$z_1 = 0; F(z_{t+1}) \leq \min_z \{F(z) : z \in \text{Conv}(\Delta(z_t) \cup \{z_t\})\}, t = 1, 2, \dots \tag{24}$$

Let  $z_* = [x_*; r_*]$  be an optimal solution to (10) (which under Assumption A clearly exists), and let  $F_* = F(z_*)$  (i.e.,  $F_*$  is nothing but Opt, see (9)).

**Theorem 3** *A generic COCndG algorithm (24) maintains the inclusions  $z_t \in K^+$  and is a descent algorithm:  $F(z_{t+1}) \leq F(z_t)$  for all  $t$ . Besides this, we have*

$$F(z_t) - F_* \leq \frac{8L_f D_*^2}{t + 14}, t = 2, 3, \dots \tag{25}$$

*COCndG with memory.* The simplest implementation of a generic COCndG algorithm is given by the recurrence

$$z_1 = 0; z_{t+1} \equiv [x_{t+1}; r_{t+1}] \in \underset{z}{\text{Argmin}} \{F(z) : z \in \text{Conv}(\Delta(z_t) \cup \{z_t\})\}, t = 1, 2, \dots \tag{26}$$

Denoting  $\widehat{z}_t := \bar{D}[x[f'(x_t)]; 1]$ , the recurrence can be written

$$z_{t+1} = \lambda_t \widehat{z}_t + \mu_t z_t, \text{ where } (\lambda_t, \mu_t) \in \underset{\lambda, \mu}{\text{Argmin}} \left\{ F(\lambda \widehat{z}_t + \mu z_t) : \lambda + \mu \leq 1, \lambda \geq 0, \mu \geq 0 \right\}. \tag{27}$$

As for the CndG algorithm in Sect. 3, the recurrence (26) admits a version with memory COCndGM still obeying (24) and thus satisfying the conclusion of Theorem 3. Specifically, assume that we already have built  $t$  iterates  $z_\tau = [x_\tau; r_\tau] \in K^+, 1 \leq \tau \leq t$ , with  $z_1 = 0$ , along with the gradients  $f'(x_\tau)$  and the points  $x[f'(x_\tau)]$ . Then we have at our disposal a number of points from  $K^+$ , namely, the iterates  $z_\tau, \tau \leq t$ , and the points  $\widehat{z}_\tau = \bar{D}[x[f'(x_\tau)]; 1]$ . Let us select a subset  $Z_t$  of the set  $\{z_\tau, \widehat{z}_\tau, 1 \leq \tau \leq t\}$ , with the only restriction that  $Z_t$  contains the points  $z_t, \widehat{z}_t$ , and set

$$z_{t+1} \in \underset{z \in \mathcal{C}_t}{\text{Argmin}} F(z), \quad \mathcal{C}_t = \text{Conv}(\{0\} \cup Z_t). \tag{28}$$

Since  $z_t, \widehat{z}_t \in Z_t$ , we have  $\text{Conv}(\Delta(z_t) \cup \{z_t\}) \subset \mathcal{C}_t$ , whence the procedure we have outlined is an implementation of generic COCndG algorithm. Note that the basic COCndG algorithm is the particular case of the COCndGM corresponding to the case

where  $Z_t = \{z_t, \widehat{z}_t\}$  for all  $t$ . The discussion of implementability of CndGM in Sect. 3 fully applies to COCndGM.

Let us outline several options which can be implemented in COCndGM; while preserving the theoretical efficiency estimates stated in Theorem 3 they can improve the practical performance of the algorithm. For the sake of definiteness, let us focus on the case of quadratic  $f: f(x) = \|Ax - b\|_2^2$ , with  $\text{Ker } \mathcal{A} = \{0\}$ ; extensions to a more general case are straightforward.

**A.** We lose nothing (and potentially gain) when extending  $C_t$  in (28) to the conic hull

$$C_t^+ = \left\{ w = \sum_{\zeta \in Z_t} \lambda_\zeta \zeta : \lambda_\zeta \geq 0, \zeta \in Z_t \right\}$$

of  $Z_t$ .<sup>9</sup> When  $K = E$ , we can go further and replace (28) with

$$z_{t+1} \in \underset{z=[x;r], \lambda}{\text{Argmin}} \left\{ f(x) + \kappa r : x = \sum_{\zeta=[\eta;\rho] \in Z_t} \lambda_\zeta \eta, r \geq \sum_{\zeta=[\eta;\rho] \in Z_t} |\lambda_\zeta| \rho \right\} \quad (29)$$

(note that  $\sum_{\zeta=[\eta;\rho] \in Z_t} \rho |\lambda_\zeta|$  is an easy-to-compute upper bound on  $\|x\|$ ). Note that the preceding “conic case” is obtained from (29) by adding to the constraints of the right hand side problem the inequalities  $\lambda_\zeta \geq 0, \zeta \in Z_t$ . Finally, when  $\|\cdot\|$  is easy to compute, we can improve (29) to

$$z_{t+1} = \left[ \sum_{\zeta=[\eta;\rho] \in Z_t} \lambda_\zeta^* \eta; \left\| \sum_{\zeta=[\eta;\rho] \in Z_t} \lambda_\zeta^* \eta \right\| \right], \quad (30)$$

$$\lambda^* \in \underset{\{\lambda_\zeta, \zeta \in Z_t\}}{\text{Argmin}} \left\{ f \left( \sum_{\zeta=[\eta;\rho] \in Z_t} \lambda_\zeta \eta \right) + \kappa \sum_{\zeta=[\eta;\rho] \in Z_t} |\lambda_\zeta| \rho \right\}$$

(the definition of  $\lambda^*$  assumes that  $K = E$ , otherwise the constraints of the problem specifying  $\lambda^*$  should be augmented by the inequalities  $\lambda_\zeta \geq 0, \zeta \in Z_t$ ).

**B.** In the case of quadratic  $f$  and moderate cardinality of  $Z_t$ , optimization problems arising in (29) (with or without added constraints  $\lambda_\zeta \geq 0$ ) are explicitly given low-dimensional “nearly quadratic” convex problems which can be solved to high accuracy “in no time” by interior point solvers. With this in mind, we could solve these problems for the given value of the penalty parameter  $\kappa$  and also for several other values of the parameter. Thus, at every iteration we get feasible approximate solution to several instances of (9) for different values of the penalty parameter. Assume that we keep in memory, for every value of the penalty parameter in question, the best, in terms of the respective objective, of the related approximate solutions found so far. Then upon termination we will have at our disposal, along with the feasible approximate solution associated with the given value of

<sup>9</sup> We recover this way the Atom-Descent algorithm of [7, 11].

the penalty parameter, provably obeying the efficiency estimates of Theorem 3, a set of feasible approximate solutions to the instances of (9) corresponding to other values of the penalty.

- C. In the above description,  $Z_t$  was assumed to be a subset of the set  $Z^t = \{z_\tau = [x_\tau; r_\tau], \widehat{z}_\tau, 1 \leq \tau \leq t\}$  containing  $z_t$  and  $\widehat{z}_t$ . Under the latter restriction, we lose nothing when allowing for  $Z_t$  to contain points from  $K^+ \setminus Z^t$  as well. For instance, when  $K = E$  and  $\|\cdot\|$  is easy to compute, we can add to  $Z_t$  the point  $z'_t = [f'(x_t); \|f'(x_t)\|]$ . Assume, e.g., that we fix in advance the cardinality  $M \geq 3$  of  $Z_t$  and define  $Z_t$  as follows: to get  $Z_t$  from  $Z_{t-1}$ , we eliminate from the latter set several (the less, the better) points to get a set of cardinality  $\leq M-3$ , and then add to the resulting set the points  $z_t, \widehat{z}_t$  and  $z'_t$ . Eliminating the points according to the rule “first in—first out,” the projection of the feasible set of the optimization problem in (30) onto the space of  $x$ -variables will be a linear subspace of  $E$  containing, starting with step  $t = M$ , at least  $\lfloor M/3 \rfloor$  (here  $\lfloor a \rfloor$  stands for the largest integer not larger than  $a$ ) of gradients of  $f$  taken at the latest iterates, so that the method, modulo the influence of the penalty term, becomes a “truncated” version of the Conjugate Gradient algorithm for quadratic minimization. Due to nice convergence properties of Conjugate Gradient in the quadratic case, one can hope that a modification of this type will improve significantly the practical performance of COCndGM.

## 6 Application examples

In this section, we detail how the proposed conditional gradient algorithms apply to several examples. In particular, we detail the corresponding LO oracles, and how one could implement these oracles efficiently.

### 6.1 Regularization by nuclear/trace norm

The first example where the proposed algorithms seem to be more attractive than the proximal methods are large-scale problems (5), (9) on the space of  $p \times q$  matrices  $E = \mathbf{R}^{p \times q}$  associated with the nuclear norm  $\|\sigma(x)\|_1$  of a matrix  $x$ , where  $\sigma(x) = [\sigma_1(x); \dots; \sigma_{\min\{p,q\}}(x)]$  is the vector of singular values of a  $p \times q$  matrix  $x$ . Problems of this type with  $K = E$  arise in various versions of matrix completion, where the goal is to recover a matrix  $x$  from its noisy linear image  $y = \mathcal{A}x + \xi$ , so that  $f = \phi(\mathcal{A}x - y)$ , with some smooth and convex discrepancy measure  $\phi(\cdot)$ , most notably,  $\phi(z) = \frac{1}{2}\|z\|_2^2$ . In this case,  $\|\cdot\|$  minimization/penalization is aimed at getting a recovery of low rank ([3, 4, 10, 18, 23, 30, 31, 33, 35, 37] and references therein). Another series of applications relates to the case when  $E = \mathbf{S}^p$  is the space of symmetric  $p \times p$  matrices, and  $K = \mathbf{S}_+^p$  is the cone of positive semidefinite matrices, with  $f$  and  $\phi$  as above; this setup corresponds to the situation when one wants to recover a covariance (and thus positive semidefinite symmetric) matrix from experimental data. Restricted from  $\mathbf{R}^{p \times p}$  onto  $\mathbf{S}^p$ , the nuclear norm becomes the trace norm  $\|\lambda(x)\|_1$ , where  $\lambda(x) \in \mathbf{R}^p$  is the vector of eigenvalues of a symmetric  $p \times p$  matrix  $x$ , and regularization by this norm is, as above, aimed at building a low rank recovery.

With the nuclear (or trace) norm in the role of  $\|\cdot\|$ , all known proximal algorithms require, at least in theory, computing at every iteration the complete singular value



decomposition of  $p \times q$  matrix  $x$  (resp., complete eigenvalue decomposition of a symmetric  $p \times p$  matrix  $x$ ), which for large  $p, q$  may become prohibitively time consuming. In contrast to this, with  $K = E$  and  $\|\cdot\| = \|\sigma(\cdot)\|_1$ , LO oracle for  $(K, \|\cdot\| = \|\sigma(\cdot)\|_1)$  only requires computing the leading right singular vector  $e$  of a  $p \times q$  matrix  $\eta$  (i.e., the leading eigenvector of  $\eta^T \eta$ ):  $x[\eta] = -\bar{f}\bar{e}^T$ , where  $\bar{e} = e/\|e\|_2$  and  $\bar{f} = \eta e/\|\eta e\|_2$  for nonzero  $\eta$  and  $\bar{f} = 0, \bar{e} = 0$  when  $\eta = 0$ . Computing the leading singular vector of a large matrix is, in most cases, much cheaper than computing the complete eigenvalue decomposition of the matrix. Similarly, in the case of  $E = \mathbf{S}^p, K = \mathbf{S}_+^p$  and the trace norm in the role of  $\|\cdot\|$ , LO oracle requires computing the leading eigenvector  $e$  of a matrix  $\eta \in \mathbf{S}^p$ :  $x[-\eta] = \bar{e}\bar{e}^T$ , where  $\bar{e} = 0$  when  $e^T \eta e \geq 0$ , and  $\bar{e} = e/\|e\|_2$  otherwise. Here again, for a large symmetric  $p \times p$  matrix, the required computation usually is much easier than computing the complete eigenvalue decomposition of such a matrix. As a result, in the situations under consideration, algorithms based on the LO oracle remain “practically implementable” in an essentially larger range of problem sizes than proximal methods.

An additional attractive property of the CndG algorithms we have described stems from the fact that *since in the situations in question the matrices  $x[\eta]$  are of rank 1,  $t$ -th approximate solution  $x_t$  yielded by the CndG algorithms for composite minimization from Sect. 5 is of rank at most  $t$ . Similar statement holds true for  $t$ -th approximate solution  $x_t$  built at a stage of a CndG algorithm for parametric optimization from Sect. 3, provided that the first iterate at every stage is the zero matrix.*<sup>10</sup>

### 6.2 Regularization by total variation

Given integer  $n \geq 2$ , consider the linear space  $M^n := \mathbf{R}^{n \times n}$ . We interpret elements  $x$  of  $M^n$  as *images*—real-valued functions  $x(i, j)$  on the  $n \times n$  grid  $\Gamma_{n,n} = \{[i; j] \in \mathbf{Z}^2 : 0 \leq i, j < n\}$ . The (anisotropic) Total Variation (TV) of an image  $x$  is the  $\ell_1$ -norm of its (discrete) gradient field  $(\nabla_i x(\cdot), \nabla_j x(\cdot))$ :

$$\begin{aligned} \text{TV}(x) &= \|\nabla_i x\|_1 + \|\nabla_j x\|_1, \\ \nabla_i x(i, j) &= x(i + 1, j) - x(i, j) : \Gamma_{n-1,n} := \{[i; j] \in \mathbf{Z}^2 : 0 \leq i < n - 1, 0 \leq j < n\}, \\ \nabla_j x(i, j) &= x(i, j + 1) - x(i, j) : \Gamma_{n,n-1} := \{[i; j] \in \mathbf{Z}^2 : 0 \leq i < n, 0 \leq j < n - 1\} \end{aligned}$$

Note that  $\text{TV}(\cdot)$  is a norm on the subspace  $M_0^n$  of  $M^n$  comprising *zero mean images*  $x$  (those with  $\sum_{i,j} x(i, j) = 0$ ) and vanishes on the orthogonal complement to  $M_0^n$ , comprised of constant images.

Originating from the celebrated paper [32] and extremely popular *Total Variation-based image reconstruction* in its basic version recovers an image  $x$  from its noisy observation  $b = \mathcal{A}x + \xi$  by solving problems (5) or (9) with  $K = E = M^n, f(x) = \phi(\mathcal{A}x - b)$  and the seminorm  $\text{TV}(\cdot)$  in the role of  $\|\cdot\|$ . In the sequel, we focus on the versions of these problems where  $K = E = M^n$  is replaced with  $K = E = M_0^n$ , thus bringing the *TV*-regularized problems into our framework. This restriction is basically

<sup>10</sup> This property is an immediate corollary of the fact that in the situation in question, by description of the algorithms  $x_t$  is a convex combination of  $t$  points of the form  $x[\cdot]$ .

harmless; for example, in the most popular case of  $f(x) = \frac{1}{2}\|Ax - b\|_2^2$  reduction to the case of  $x \in M_0^n$  is immediate—it suffices to replace  $(\mathcal{A}, b)$  with  $(P\mathcal{A}, Pb)$ , where  $P$  is the orthoprojector onto the orthogonal complement to the one-dimensional subspace spanned by  $\mathcal{A}e$ , where  $e$  is the all-ones image.<sup>11</sup> Now, large scale problems (5), (9) with  $K = E = M_0^n$  and  $\text{TV}(\cdot)$  in the role of  $\|\cdot\|$  are difficult to solve by proximal algorithms. Indeed, in the situation in question a proximal algorithm would require at every iteration either minimizing function of the form  $\text{TV}(x) + \langle e, x \rangle + \omega(x)$  over the entire  $E$ , or minimizing function of the form  $\langle e, x \rangle + \omega(x)$  on a TV-ball,<sup>12</sup> where  $\omega(x)$  is albeit simple, but *nonlinear* convex function (e.g.,  $\|x\|_2^2$ , or  $\|\nabla_i x\|_2^2 + \|\nabla_j x\|_2^2$ ). Auxiliary problems of this type seem to be difficult in the large scale case, especially taking into account that when running a proximal algorithm we need to solve at least tens, and more realistically—hundreds of them.<sup>13</sup> In contrast to this, an LO oracle for the unit ball  $\mathcal{TV} = \{x \in M_0^n : \text{TV}(x) \leq 1\}$  of the TV norm is relatively cheap computationally—it reduces to solving a specific maximum flow problem. It should be mentioned here that the relation between flow problems and TV-based denoising (problem (9) with  $\mathcal{A} = I$ ) is well known and is utilized in many algorithms, see [10] and references therein. While we have no doubt that the simple fact stated Lemma 2 below is well-known, for reader convenience we present here in detail the reduction mechanism.

Consider the network (the oriented graph)  $G$  with  $n^2$  nodes  $[i; j] \in \Gamma_{n,n}$  and  $2n(n - 1)$  arcs as follows: the first  $n(n - 1)$  arcs are of the form  $([i + 1; j], [i; j])$ ,  $0 \leq i < n - 1, 0 \leq j < n$ , the next  $n(n - 1)$  arcs are  $([i; j + 1], [i; j])$ ,  $0 \leq i < n, 0 \leq j < n - 1$ , and the remaining  $2n(n - 1)$  arcs (let us call them *backward* arcs) are the inverses of the just defined  $2n(n - 1)$  *forward* arcs. Let  $\mathcal{E}$  be the set of arcs of our network, and let us equip all the arcs with unit capacities. Let us treat vectors from  $E = M_0^n$  as vectors of external supplies for our network; note that the entries of these vectors sum to zero, as required from external supply. Now, given a nonzero vector  $\eta \in M_0^n$ , let us consider the network flow problem where we seek for the largest multiple  $s\eta$  of  $\eta$  which, considered as the vector of external supplies in our network, results in a feasible capacitated network flow problem. The problem in question reads

$$s_* = \max_{s,r} \{s : Pr = s\eta, 0 \leq r \leq e\}, \tag{31}$$

<sup>11</sup> When  $f$  is more complicated, optimal adjustment of the mean  $t$  of the image reduces by bisection in  $t$  to solving small series of problems of the same structure as (5), (9) where the mean of the image  $x$  is fixed and, consequently, the problems reduce to those with  $x \in M_0^n$  by shifting  $b$ .

<sup>12</sup> Which one of these two options takes place depends on the type of the algorithm.

<sup>13</sup> On a closest inspection, “complex geometry” of the TV-norm stems from the fact that after parameterizing a zero mean image by its discrete gradient field and treating this field ( $g = \nabla_i x, h = \nabla_j x$ ) as our new design variable, the unit ball of the TV-norm becomes the intersection of a simple set in the space of pairs  $(g, h) \in F = \mathbf{R}^{(n-1) \times n} \times \mathbf{R}^{n \times (n-1)}$  (the  $\ell_1$  ball  $\Delta$  given by  $\|g\|_1 + \|h\|_1 \leq 1$ ) with a linear subspace  $P$  of  $F$  comprised of *potential* vector fields  $(f, g)$ —those which indeed are discrete gradient fields of images. Both dimension and codimension of  $P$  are of order of  $n^2$ , which makes it difficult to minimize over  $\Delta \cap P$  *nonlinear*, even simple, convex functions, which is exactly what is needed in proximal methods.

where  $P$  is the incidence matrix of our network<sup>14</sup> and  $\mathbf{e}$  is the all-ones vector. Now, problem (31) clearly is feasible, and its feasible set is bounded due to  $\eta \neq 0$ , so that the problem is solvable. Due to its network structure, this LP program can be solved reasonably fast even in the large scale case (say, when  $n = 512$  or  $n = 1024$ , which already is of interest for actual imaging). Further, an intelligent network flow solver as applied to (31) will return not only the optimal  $s = s_*$  and the corresponding flow, but also the dual information, in particular, the optimal vector  $z$  of Lagrange multipliers for the linear equality constraints  $Pr - s\eta = 0$ . Let  $\bar{z}$  be obtained by subtracting from the entries of  $z$  their mean; since the entries of  $z$  are indexed by the nodes,  $\bar{z}$  can be naturally interpreted as a zero mean image. It turns out that this image is nonzero, and the vector  $x[\eta] = -\bar{z}/\text{TV}(\bar{z})$  is nothing than a desired minimizer of  $\langle \eta, \cdot \rangle$  on  $\mathcal{TV}$ :

**Lemma 2** *Let  $\eta$  be a nonzero image with zero mean. Then (31) is solvable with positive optimal value, and the image  $x[\eta]$ , as defined above, is well defined and is a maximizer of  $\langle \eta, \cdot \rangle$  on  $\mathcal{TV}$ .*

*Bounding  $L_f$ .* When applying CndG algorithms to the TV-based problems (5), (9) with  $E = M_0^n$  and  $f(x) = \phi(\mathcal{A}x - b)$ , the efficiency estimates depend linearly on the associated quantity  $L_f$ , which, in turn, is readily given by the norm  $\|\mathcal{A}\|_{\text{TV}(\cdot), \pi(\cdot)}$  of the mapping  $x \mapsto \mathcal{A}x$ , see the end of Sect. 2. Observe that in typical applications  $\mathcal{A}$  is a simple operator (e.g., the discrete convolution), so that when restricting ourselves to the case when  $\pi(\cdot)$  is  $\|\cdot\|_2$  (quadratic fit), it is easy to find a tight upper bound on  $\|\mathcal{A}\|_{\|\cdot\|_2, \|\cdot\|_2}$ . To convert this bound into an upper bound on  $\|\mathcal{A}\|_{\text{TV}(\cdot), \|\cdot\|_2}$ , we need to estimate the quantity

$$Q_n = \max_x \{\|x\|_2 : x \in M_0^n, \text{TV}(x) \leq 1\}.$$

Bounding  $Q_n$  is not a completely trivial question, and the answer is as follows:

**Proposition 1**  *$Q_n$  is nearly constant, specifically,  $Q_n \leq O(\sqrt{\ln(n)})$ .*

Note that the result of Proposition 1 is in sharp contrast with one-dimensional case, where the natural analogy of  $Q_n$  grows with  $n$  as  $\sqrt{n}$ . We do not know whether it is possible to replace in Proposition 1  $O(1)\sqrt{\ln(n)}$  with  $O(1)$ , as suggested by Sobolev’s inequalities.<sup>15</sup> Note that on inspection of the proof, the proposition extends to the case of  $d$ -dimensional,  $d > 2$ , images with zero mean, in which case  $Q_n \leq C(d)$  with appropriately chosen  $C(d)$ .

<sup>14</sup> That is, the rows of  $P$  are indexed by the nodes, the columns are indexed by the arcs, and in the column indexed by an arc  $\gamma$  there are exactly two nonzero entries: entry 1 in the row indexed by the starting node of  $\gamma$ , and entry  $-1$  in the row indexed by the terminal node of  $\gamma$ .

<sup>15</sup> From the Sobolev embedding theorem it follows that for a smooth function  $f(x, y)$  on the unit square one has  $\|f\|_{L_2} \leq O(1)\|\nabla f\|_1$ ,  $\|\nabla f\|_1 := \|f'_x\|_1 + \|f'_y\|_1$ , provided that  $f$  has zero mean. Denoting by  $f^n$  the restriction of the function onto a  $n \times n$  regular grid in the square, we conclude that  $\|f^n\|_2/\text{TV}(f^n) \rightarrow \|f\|_{L_2}/\|\nabla f\|_1 \leq O(1)$  as  $n \rightarrow \infty$ . Note that the convergence in question takes place only in the 2-dimensional case.

## 7 Numerical examples

We present here some very preliminary simulation results.

### 7.1 CndG for parametric optimization: sparse matrix completion problem

The goal of the first series of our experiments is to illustrate how the performance and requirements of CndG algorithm for parametric optimization, when applied to the matrix completion problem [4], scale with problem size. Specifically, we apply the algorithm of Sect. 4 to the problem of nuclear norm minimization

$$\min \|\sigma(x)\|_1, \quad \text{subject to} \quad \sum_{(i,j) \in \Omega} (y_{ij} - x_{ij})^2 - \delta \leq 0, \quad (32)$$

where  $\sigma(x)$  is the singular spectrum of a  $p \times q$  matrix  $x$ . In our experiments, the set  $\Omega$  of observed entries  $(i, j) \in \{1, \dots, p\} \times \{1, \dots, q\}$  of cardinality  $m \ll pq$  was selected at random.

To meet the memory requirements of the CndGM, when running the algorithm, at iteration  $t$ , we store in memory rank-one updates  $x_\tau^+$ ,  $\tau = 1, \dots, t$ , the coefficients  $v_\tau^t$ ,  $\tau = 1, \dots, t$ , of the decomposition

$$x_t = \sum_{\tau=1}^t v_\tau^t x_\tau^+ \quad (33)$$

of the current approximate solution  $x_t$ , along with  $m$ -dimensional vectors  $\zeta_\tau = [x_\tau^+]_{ij}$ ,  $(i, j) \in \Omega$ , matrices  $Q^t = [x_r^+, x_s^+]_{r,s=1}^t$ , and vectors  $q^t = [\langle x_\tau^+, y \rangle_\Omega]_{\tau=1}^t$ , where  $\langle u, v \rangle_\Omega = \sum_{(i,j) \in \Omega} u_{ij} v_{ij}$ . When implementing, for instance, the “full memory” version of CndGM (19), (20), in which the set  $X_t$  contains  $x_t$  and all the points  $x_\tau^+$  for  $1 \leq \tau \leq t$ , at an iteration  $t$  we need to solve problem (22) which in our situation is a convex quadratic problem with  $M \leq t + 1$  variables  $\{\lambda_x^t\}_{x \in X_t}$ . Such a problem can be efficiently solved using existing software [1], provided we have at our disposal the coefficients of the quadratic objective of the problem. These coefficients are readily given by the reals  $v_\tau^t$ , see (33), and the entries of  $Q^t$ ,  $q^t$ , with updating  $(Q^{t-1}, q^{t-1}) \mapsto (Q^t, q^t)$  reducing to computing  $t + 1$  inner products of  $m$ -dimensional vectors, namely, the products  $\langle x_\tau^+, x_\tau^+ \rangle_\Omega$ ,  $1 \leq \tau \leq t$ , and  $\langle x_\tau^+, y \rangle_\Omega$ . As a consequence, assembling and solving auxiliary problems (22) represents only a small part of the total numerical effort, dominated by cost of computing  $x_\tau^+$  (that is, the leading singular vectors of a sparse  $p \times q$  matrix, which requires few tens of sparse matrix-vector multiplications), and extracting  $\zeta_t$  from  $x_t^+$ .

We compare the performance of CndGM algorithms and of a “memoryless” version of the CndG. To this end we have conducted the following experiment:

1. For matrix sizes  $p, q \in [1, 2, 4, 8, 16, 32] \times 10^3$  we generate  $n = 10$  sparse  $p \times q$  matrices  $y$  with density  $d = 0.1$  of non-vanishing entries as follows: we generate  $p \times r$  matrix  $U$  and  $q \times r$  matrix  $V$  with independent Gaussian

entries  $u_{ij} \sim \mathcal{N}(0, m^{-1})$ ,  $v_{ij} \sim \mathcal{N}(0, n^{-1})$ , and a  $r \times r$  diagonal matrix  $D = \text{diag}[d_1, \dots, d_r]$  with  $d_i$  drawn independently from a uniform distribution on  $[0, 1]$ . The non-vanishing entries of the sparse observation matrix  $y$  are obtained by sampling at random with probability  $d$  the entries of  $x^* = UDV^T$ , so that for every  $i, j$ ,  $y_{ij}$  is, independently over  $i, j$ , set to  $x_{ij}^*$  with probability  $d$  and to 0 with probability  $1 - d$ . Thus, the number of non-vanishing entries of  $y$  is approximately  $m = dpq$ . This procedure results in  $m \sim 10^5$  for the smallest matrices  $y$  ( $1,000 \times 1,000$ ), and in  $m \sim 10^8$  for the largest matrices ( $32,000 \times 32,000$ ).

2. We apply to parametric optimization problem (32) MATLAB implementations of the CndGM with memory parameter  $M = 2$  (“memoryless” CndG), CndGM with  $M = 6$  and full memory CndGM. The parameter  $\delta$  of (32) is chosen to be  $\delta = 0.001 \|y\|_f^2$  (here  $\|y\|_f = \left(\sum_{i,j} y_{ij}^2\right)^{1/2}$  stands for the Frobenius norm of  $y$ ). The optimization algorithm is tuned to the *relative accuracy*  $\varepsilon = 1/4$ , what means that it outputs an  $\varepsilon$ -solution  $\hat{x}$  to (32), in the sense of (8), with absolute accuracy  $\epsilon = \delta\varepsilon$ .

For each algorithm (memoryless CndG, CndGM with memory  $M = 6$  and full memory CndGM) we present in Table 1 the average, over algorithm’s runs on the (common for all algorithms) sample of  $n = 10$  matrices  $y$  we have generated, (1) total number of iterations  $N_{it}$  necessary to produce an  $\varepsilon$ -solution (it upper-bounds the rank of the resulting  $\varepsilon$ -solution), (2) CPU time in seconds  $T_{cpu}$  and (3) MATLAB memory usage in megabytes  $S_{mem}$ . This experiment was conducted on a Dell Latitude 6430 laptop equipped with Intel Core i7-3720QM CPU@2.60GHz and 16GB of RAM. Because of high memory requirements in our implementation of the full memory CndGM, this method was unable to complete the computation for the two largest matrix sizes.

We can make the following observation regarding the results summarized in Table 1: CndG algorithm with memory consistently outperforms the standard—memoryless—version of CndG. The full memory CndGM requires the smallest number of iteration to produce an  $\varepsilon$ -solution, which is of the smallest rank, as a result. On the other hand, the memory requirements of the full memory CndGM become prohibitive (at least, for the computer we used for this experiment and MATLAB implementation of the memory heap) for large matrices. On the other hand, a CndGM with memory  $M = 6$  appears to be a reasonable compromise in terms of numerical efficiency and memory demand.

### 7.2 CndG for composite optimization: multi-class classification with nuclear-norm regularization

We present here an empirical study of the CndG algorithm for composite optimization as applied to the machine learning problem of multi-class classification with nuclear-norm penalty. A brief description of the multi-class classification problem is as follows: we observe  $N$  “feature vectors”  $\xi_i \in \mathbf{R}^q$ , each belonging to exactly one of  $p$  classes  $C_1, \dots, C_p$ . Each  $\xi_i$  is augmented by its label  $y_i \in \{1, \dots, p\}$  indicating to which class  $\xi_i$  belongs. Our goal is to build a classifier capable to predict the class to which

**Table 1** Memoryless CndG versus CndGM with memory  $M = 6$  versus full memory CndGM

Matrix size $p \times q$	Memory-less CndG			CndGM with memory $M = 6$			Full memory CndG		
	$N_{it}$	$T_{cpu}$	$S_{mem}$	$N_{it}$	$T_{cpu}$	$S_{mem}$	$N_{it}$	$T_{cpu}$	$S_{mem}$
1,000 × 1,000	271.6	9.35	17.11	149.7	5.01	17.63	78.4	4.71	78.98
1,000 × 2,000	292.1	12.14	31.67	162.8	7.76	32.57	93.5	10.89	156.22
2,000 × 2,000	246.8	17.01	54.45	139.1	11.19	61.57	71.9	13.31	248.13
2,000 × 4,000	259.3	33.94	105.09	152.3	24.50	120.22	57.7	25.54	410.02
4,000 × 4,000	321.8	79.20	207.26	162.9	50.59	235.59	74.6	93.22	1014.7
4,000 × 8,000	360.1	169.8	399.16	147.3	88.81	464.68	63.3	135.6	1766.4
8,000 × 8,000	323.4	302.8	754.46	111.8	134.1	905.98	53.6	191.3	3061.5
8,000 × 16,000	324.1	614.3	1485.6	118.2	286.5	1800.7	50.5	329.4	5826.7
16,000 × 16,000	258.7	995.4	2898.5	99.7	495.5	3577.8	50.8	595.2	11,696
16,000 × 32,000	276.7	2,572	5721.7	70.3	859.2	7109.0	NA	NA	NA
32,000 × 32,000	305.4	5,028	11,352	57.6	2,541	14,186	NA	NA	NA

$N_{it}$ : total number of method iterations;  $T_{cpu}$ : CPU usage (sec), and  $S_{mem}$ : memory usage (MB) reported by MATLAB

a new feature vector  $\xi$  belongs. This classifier is given by a  $p \times q$  matrix  $x$  according to the following rule: given  $\xi$ , we compute the  $p$ -dimensional vector  $x\xi$  and take, as the guessed class of  $\xi$ , the index of the largest entry in this vector.

In some cases (see [7, 11]), when, for instance, one is dealing with a large number of classes, there are good reasons “to train the classifier”—to specify  $x$  given the training sample  $(\xi_i, y_i)$ ,  $1 \leq i \leq N$ —as the optimal solution to the nuclear norm penalized minimization problem

$$\text{Opt}(\kappa) = \min_{x \in \mathbf{R}^{p \times q}} F_\kappa(x) := \overbrace{\frac{1}{N} \sum_{i=1}^N \log \left\{ \sum_{\ell=1}^p \exp \left( (x_\ell^T - x_{y_i}^T) \xi_i \right) \right\}}^{f(x)} + \kappa \|\sigma(x)\|_1, \tag{34}$$

where  $x_\ell^T$  is the  $\ell$ -th row in  $x$ .

Below, we report on some experiments with this problem. Our goal was to compare two versions of CndG for composite minimization: the memoryless version defined in (24) and the version with memory defined in (28). To solve the corresponding sub-problems, we used the Center of Gravity method in the case of (24) and the Ellipsoid method in the case of (28) [25,26]. In the version with memory we set  $M = 6$ , as it appeared to be the best option from empirical evidence. We have considered the following datasets:

- 1. Simulated data:** for matrix of sizes  $p, q \in 10^3 \times \{2^s\}_{s=1}^4$ , we generate random matrices  $x_\star = USV$ , with  $p \times p$  factor  $U$ ,  $q \times q$  factor  $V$ , and diagonal  $p \times q$  factor  $S$  with random entries sampled, independently of each other, from  $\mathcal{N}(0, p^{-1})$  (for  $U$ ),  $\mathcal{N}(0, q^{-1})$  (for  $V$ ), and the uniform distribution on  $[0, 1]$  (for diagonal entries in  $S$ ). We use  $N = 20q$ , with the feature vectors  $\xi_1, \dots, \xi_N$  sampled,

independently of each other, from the distribution  $\mathcal{N}(0, I_q)$ , and their labels  $y_i$  being the indexes of the largest entries in the vectors  $x_\star \xi_i + \epsilon_i$ , where  $\epsilon_i \in \mathbf{R}^p$  were sampled, independently of each other and of  $\xi_1, \dots, \xi_N$ , from  $\mathcal{N}(0, \frac{1}{2}I_p)$ . The regularization parameter  $\kappa$  is set to  $10^{-3}\text{Tr}(x_\star x_\star^T)$ .

2. **Real-world data:** we follow a setting similar to [11]. We consider the Pascal ILSVRC2010 ImageNet dataset and focus on the ‘‘Vertebrate-craniate’’ subset, yielding 1,043 classes with 20 examples per class. The goal here is to train a multi-class classifier in order to be able to predict the class of each image (example) of the dataset. Each example is converted to a 65,536-dimensional feature vector of unit  $\ell_1$ -norm using state-of-the-art visual descriptors known as Fisher vector representation [11]. To summarize, we have  $p = 1,043, q = 65,536, N = 20,860$ . We set the regularization parameter to  $\kappa = 10^{-4}$ , which was found to result in the best predictive performance as estimated by cross-validation, a standard procedure to set the hyper parameters in machine learning [13].

In both sets of experiments, the computations are terminated when the ‘‘ $\epsilon$ -optimality conditions’’

$$\begin{aligned} \|\sigma(f'(x_t))\|_\infty &\leq \kappa + \epsilon \\ \langle f'(x_t), x_t \rangle + \kappa \|\sigma(x_t)\|_1 &\leq \epsilon \|\sigma(x_t)\|_1 \end{aligned} \tag{35}$$

were met, where  $\|\sigma(\cdot)\|_\infty$  denotes the usual operator norm (the largest singular value). These conditions admit transparent interpretation as follows. For every  $\bar{x}$ , the function

$$\phi_\kappa(x) = f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x} \rangle + \kappa \|\sigma(x)\|_1$$

underestimates  $F_\kappa(x)$ , see (34), whence  $\text{Opt}(\kappa') \geq f(\bar{x}) - \langle f'(\bar{x}), \bar{x} \rangle$  whenever  $\kappa' \geq \|\sigma(f'(\bar{x}))\|_\infty$ . Thus, whenever  $\bar{x} = x_t$  satisfies the first relation in (35), we have  $\text{Opt}(\kappa + \epsilon) \geq f(x_t) - \langle f'(x_t), x_t \rangle$ , whence

$$F_\kappa(x_t) - \text{Opt}(\kappa + \epsilon) \leq \langle f'(x_t), x_t \rangle + \kappa \|\sigma(x_t)\|_1.$$

We see that (35) ensures that  $F_\kappa(x_t) - \text{Opt}(\kappa + \epsilon) \leq \epsilon \|\sigma(x_t)\|_1$ , which, for small  $\epsilon$ , is a reasonable substitute for the actually desired termination when  $F_\kappa(x_t) - \text{Opt}(\kappa)$  becomes small. In our experiments, we use  $\epsilon = 0.001$ .

In Table 2 for each algorithm (memoryless CndG, CndGM with memory  $M = 6$ ) we present the average, over 20 collections of simulated data coming from 20 realizations of  $x_\star$ , of: (1) total number of iterations  $N_{\text{it}}$  necessary to produce an  $\epsilon$ -solution, (2) CPU time in seconds  $T_{\text{cpu}}$ . The last row of the table corresponds to the real-world data. Experiments were conducted on a Dell R905 server equipped with four six-core AMD Opteron 2.80 GHz CPUs and 64 GB of RAM. A maximum of 32 GB of RAM was used for the computations.

We draw the following conclusions from Table 1: CndG algorithm with memory routinely outperforms the standard—memoryless—version of CndG. However, there is a trade-off between the algorithm progress at each iteration and the computational

**Table 2** memoryless CndG versus CndGM with memory  $M = 6$

Matrix size $p \times q$	Memory-less CndG			CndGM with memory $M = 6$		
	$N_{it}$	$T_{cpu}$	$S_{mem}$	$N_{it}$	$T_{cpu}$	$S_{mem}$
2,000 × 2,000	174.3	355.8	135.3	102.7	226.1	177.5
4,000 × 4,000	157.1	1,042	549.7	88.3	726.9	714.5
8,000 × 8,000	195.5	2,793	2,210	123.6	2,201	2,885
16,000 × 16,000	236.6	6,775	9,082	138.0	4,963	11,798
32,000 × 32,000	277.3	26,803	30,660	143.8	18,990	31,327
1,043 × 65,536	185	2,133	2,110	111	1343	2,781

$N_{it}$ : total number of method iterations;  $T_{cpu}$ : CPU usage (sec) reported by MATLAB

load of each iteration. Note that, for large  $M$ , solving the sub-problem (28) can be challenging.

### 7.3 CndG for composite optimization: TV-regularized image reconstruction

Here we report on experiments with COCndGM as applied to TV-regularized image reconstruction. Our problem of interest is of the form (9) with quadratic  $f$ , namely, the problem

$$\min_{x \in M_0^n} \phi_\kappa(x) := \underbrace{\frac{1}{2} \|PAx - Pb\|_2^2}_{f(x)} + \kappa \text{TV}(x); \tag{36}$$

for notation, see Sect. 6.2.

*Test problems.* In our experiments, the mapping  $x \mapsto Ax$  is defined as follows: we zero-pad  $x$  to extend it from  $\Gamma_{n,n}$  to get a finitely supported function on  $\mathbf{Z}^2$ , then convolve this function with a finitely supported kernel  $\alpha(\cdot)$ , and restrict the result onto  $\Gamma_{n,n}$ . The observations  $b \in M^n$  were generated at random according to

$$b_{ij} = (Ax)_{ij} + \sigma \|x\|_\infty \xi_{ij}, \quad \xi_{ij} \sim \mathcal{N}(0, 1), \quad 1 \leq i, j \leq n, \tag{37}$$

with mutually independent  $\xi_{ij}$ . The relative noise intensity  $\sigma > 0$ , same as the convolution kernel  $\alpha(\cdot)$ , are parameters of the setup of an experiment.

*The algorithm.* We used the COCndG with memory, described in section 5; we implemented the options listed in A–C at the end of the section. Specifically,

1. We use the updating rule (30) with  $Z_t$  evolving in time exactly as explained in item C: the set  $Z_t$  is obtained from  $Z_{t-1}$  by adding the points  $z_t = [x_t; \text{TV}(x_t)]$ ,  $\widehat{z}_t = [x[\nabla f(x_t)]; 1]$  and  $z'_t = [\nabla f(x_t); \text{TV}(\nabla f(x_t))]$ , and deleting from the resulting set, if necessary, some “old” points, selected according to the rule “first in—first out,” to keep the cardinality of  $Z_t$  not to exceed a given  $M \geq 3$  (in our experiments we use  $M = 48$ ). This scheme is initialized with  $Z_0 = \emptyset$ ,  $z_1 = [0; 0]$ .



2. We use every run of the algorithm to obtain a set of approximate solutions to (36) associated with various values of the penalty parameter  $\kappa$ , as explained in **B** at the end of Sect. 5. Precisely, when solving (36) for a given value of  $\kappa$  (in the sequel, we refer to it as to the *working value*, denoted  $\kappa_w$ ), we also compute approximate solutions  $x_\kappa(\kappa')$  to the problems with the values  $\kappa'$  of the penalty, for  $\kappa' = \kappa\gamma$ , with  $\gamma$  running through a given finite subset  $G \ni 1$  of the positive ray. In our experiments, we used the 25-point grid  $G = \{\gamma = 2^{\ell/4}\}_{\ell=-12}^{12}$ .

The LO oracle for the TV norm on  $M_0^n$  utilized in COCndGM was the one described in Lemma 2; the associated flow problem (31) was solved by the commercial interior point LP solver `mosekopt` version 6 [1]. Surprisingly, in our application this “general purpose” interior point LP solver was by orders of magnitude faster than all dedicated network flow algorithms we have tried, including simplex-type network versions of `mosekopt` and `CPLEX`. With our solver, it becomes possible to replace in (31) every pair of opposite to each other arcs with a single arc, passing from the bounds  $0 \leq r \leq \mathbf{e}$  on the flows in the arcs to the bounds  $-\mathbf{e} \leq r \leq \mathbf{e}$ .

The *termination criterion* we use relies upon the fact that in COCndGM the (non-negative) objective decreases along the iterates: we terminate a run when the progress in terms of the objective becomes small, namely, when the condition

$$\phi_\kappa(x_{t-1}) - \phi_\kappa(x_t) \leq \epsilon \max[\phi_\kappa(x_{t-1}), \delta\phi_\kappa(0)]$$

is satisfied. Here  $\epsilon$  and  $\delta$  are small tolerances (we used  $\epsilon = 0.005$  and  $\delta = 0.01$ ).

*Organization of the experiments.* In each experiment we select a “true image”  $x^* \in M^n$ , a kernel  $\alpha(\cdot)$  and a (relative) noise intensity  $\sigma$ . Then we generate a related observation  $b$ , thus ending up with a particular instance of (36). This instance is solved by the outlined algorithm for working values  $\kappa_w$  of  $\kappa$  taken from the set  $G^+ = \{\gamma = 2^{\ell/4}\}_{\ell=-\infty}^{\infty}$ , with the initial working value, selected in pilot runs, of the penalty underestimating the best—resulting in the best recovery—penalty.

As explained above, a run of COCndGM, the working value of the penalty being  $\kappa_w$ , yields 25 approximate solutions to (36) corresponding to  $\kappa$  along the grid  $\kappa_w \cdot G$ . These sets are fragments of the grid  $G^+$ , with the ratio of the consecutive grid points  $2^{1/4} \approx 1.19$ . For every approximate solution  $x$  we compute its *combined relative error* defined as

$$v(x) = \left( \frac{\|\bar{x} - x^*\|_1 \|\bar{x} - x^*\|_2 \|\bar{x} - x^*\|_\infty}{\|x^*\|_1 \|x^*\|_2 \|x^*\|_\infty} \right)^{1/3};$$

here  $\bar{x}$  is the easily computable shift of  $x$  by a constant image satisfying  $\|\mathcal{A}\bar{x} - b\|_2 = \|P\mathcal{A}x - Pb\|_2$ . From run to run, we increase the working value of the penalty by the factor  $2^{1/4}$ , and terminate the experiment when in four consecutive runs there was no progress in the combined relative error of the best solution found so far. Our primary goals are (a) to quantify the performance of the COCndGM algorithm, and (b) to understand by which margin, in terms of  $\phi_\kappa(\cdot)$ , the “byproduct” approximate solutions yielded by the algorithm (those which were obtained when solving (36) with

**Table 3** Setups of the experiments

#	Image	$n$	$\alpha(\cdot)$	$\text{Cond}(\mathcal{A}^* \mathcal{A})$	$\sigma$
A	lenna <sup>†</sup>	256	$\text{fspecial}('gaussian', 7, 1)^1 (7 \times 7)$	$\approx 2.5e7$	0.05
B	cameraman <sup>‡</sup>	512	$\text{fspecial}('gaussian', 7, 1) (7 \times 7)$	$\approx 2.5e7$	0.05
C	lenna	256	$\text{fspecial}('unsharp') (3 \times 3)$	$\approx 40$	0.15
D	cameraman	512	$\text{fspecial}('unsharp') (3 \times 3)$	$\approx 40$	0.40

<sup>†</sup> <http://en.wikipedia.org/wiki/Lenna>

<sup>‡</sup> [http://en.wikipedia.org/wiki/Camera\\_operator](http://en.wikipedia.org/wiki/Camera_operator)

<sup>1</sup> <http://www.mathworks.com/help/images/ref/fspecial.html>

**Table 4** Performance of COCndGM; platform: T410 Lenovo laptop, Intel Core i7 M620 CPU@2.67 GHz, 8 GB RAM

#	Image size	Runs	Iterations per run			CPU per run, sec		CPU per iteration, sec
			Min	Mean	Max	Mean	Max	
A	$256 \times 256$	6	4	9.00	12	83.4	148.7	8.3
B	$512 \times 512$	9	4	7.89	11	212.9	318.2	25.9
C	$256 \times 256$	6	17	17.17	18	189.7	214.7	10.3
D	$512 \times 512$	6	16	16.00	16	615.9	768.3	36.0

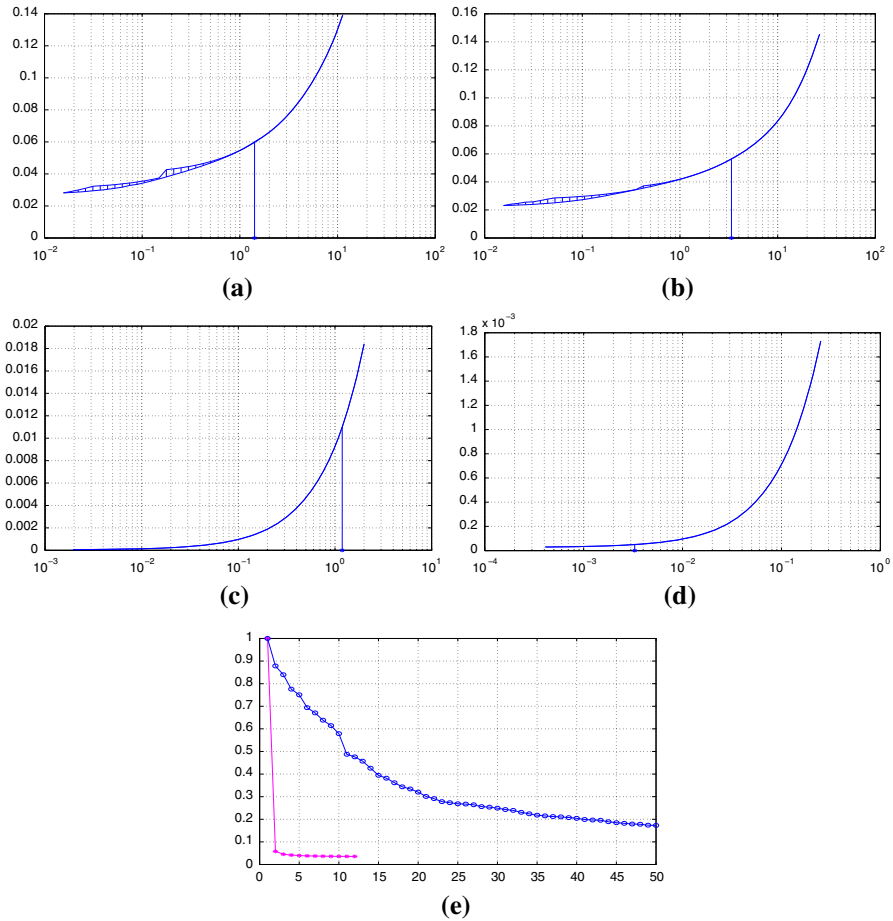
Flow solver: interior point method `mosekopt 6.0` [1]

the working value of penalty different from  $\kappa$ ) are worse than the “direct” approximate solution obtained for the working value  $\kappa$  of the penalty.

*Test instances and results.* We present below the results of four experiments with two popular images; these results are fully consistent with those of other experiments we have conducted so far. The corresponding setups are presented in Table 3. Table 4 summarizes the performance data. Our comments are as follows.

- In accordance to the above observations, using “large” memory (with the cardinality of  $Z_t$  allowed to be as large as 48) and processing “large” number (25) of penalty values at every step are basically costless: at an iteration, the *single* call to the LO oracle (which is a must for CndG) takes as much as 85 % of the iteration time.
- The COCndGM iteration count as presented in Table 4 is surprisingly low for an algorithm with sublinear  $O(1/t)$  convergence, and the running time of the algorithm appears quite tolerable<sup>16</sup> Seemingly, the *instrumental* factor here is that by reasons indicated in C, see the end of Sect. 5, we include into  $Z_t$  not only  $z_t = [x_t; \text{TV}(x_t)]$  and  $\hat{z}_t = [x[\nabla f(x_t)]; 1]$ , but also  $z'_t = [\nabla f(x_t); \text{TV}(\nabla f(x_t))]$ . To illustrate the difference, this is what happens in experiment A with the lowest

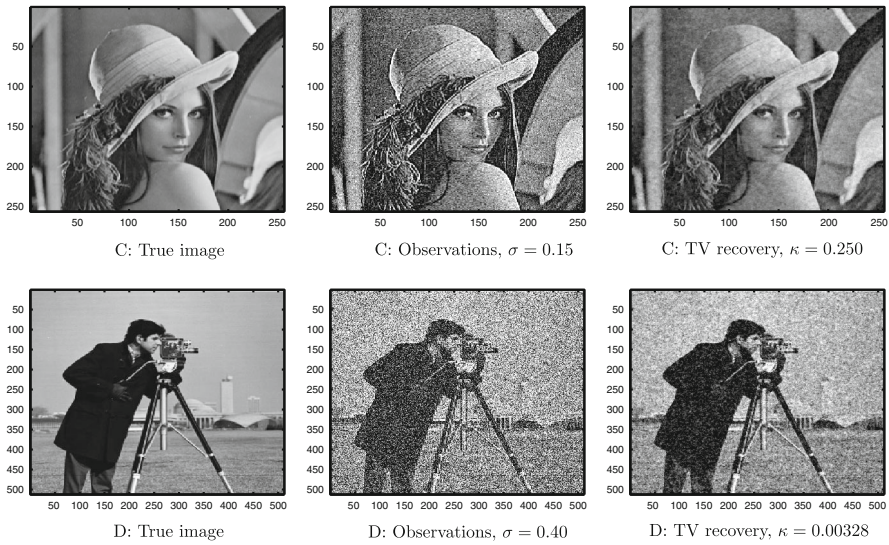
<sup>16</sup> For comparison: solving on the same platform problem (36) corresponding to Experiment A ( $256 \times 256$  image) by the state-of-the-art commercial interior point solver `mosekopt 6.0` took as much as 3,727 sec, and this—for a *single* value of the penalty (there is no clear way to get from a single run approximate solutions for a set of values of the penalty in this case).



**Fig. 1** a–d Lower and upper envelopes of  $\{\phi_\kappa(x_{\kappa_w}(\kappa)) : \kappa_w \in G\}$  versus  $\kappa$ , experiments A–D. Asterisks on the  $\kappa$ -axes: penalties resulting in the smallest combined relative recovery errors. e values of  $\phi_{1/8}(x_t)$  versus iteration number  $t$  with  $z'_t$  included (asterisks) and not included (circles) into  $Z_t$

(0.125) working value of penalty. With the outlined implementation, the run takes 12 iterations (111 sec), with the ratio  $\phi_{1/8}(x_t)/\phi_{1/8}(x_1)$  reduced from 1 ( $t = 1$ ) to 0.036 ( $t = 12$ ). When  $z'_t$  is not included into  $Z_t$ , the termination criterion is not met even in 50 iterations (452 s), the maximum iteration count we allow for a run, and in course of these 50 iterations the above ratio was reduced from 1 to 0.17, see plot e) on Fig. 1.

- An attractive feature of the proposed approach is the possibility to extract from a single run, the working value of the penalty being  $\kappa_w$ , suboptimal solutions  $x_{\kappa_w}(\kappa)$  for a bunch of instances of (9) differing from each other by the values of the penalty  $\kappa$ . The related question is, of course, how good, in terms of the objective  $\phi_\kappa(\cdot)$ , are the “byproduct” suboptimal solutions  $x_{\kappa_w}(\kappa)$  as compared to those obtained when  $\kappa$  is the working value of the penalty. In our experiments, the “byproduct”



**Fig. 2** Experiments C, D

solutions were pretty good, as can be seen from plots (a)—(c) on Fig. 1, where we see the upper and the lower envelopes of the values of  $\phi_\kappa$  at the approximate solutions  $x_{\kappa_w}(\kappa)$  obtained from different working values  $\kappa_w$  of the penalty. In spite of the fact that in our experiments the ratios  $\kappa/\kappa_w$  could be as small as  $1/8$  and as large as  $8$ , we see that these envelopes are pretty close to each other, and, as an additional bonus, are merely indistinguishable in a wide neighborhood of the best (resulting in the best recovery) value of the penalty (on the plots, this value is marked by asterisk).

Finally, we remark that in experiments A, B, where the mapping  $\mathcal{A}$  is heavily ill-conditioned (see Table 3), TV regularization yields moderate (just about 25%) improvement in the combined relative recovery error as compared to the one of the trivial recovery (“observations as they are”), in spite of the relatively low ( $\sigma = 0.05$ ) observation noise. In contrast to this, in the experiments C, D, where  $\mathcal{A}$  is well-conditioned, TV regularization reduces the error by 80% in experiment C ( $\sigma = 0.15$ ) and by 72% in experiment D ( $\sigma = 0.4$ ), see Fig. 2.

## 8 Appendix

### 8.1 Proof of Theorem 1

Define  $\epsilon_t = f(x_t) - f_*^t$ . When invoking convexity of  $f$  and the definition (16), (17) of  $f_*^t$ , we have

$$\langle f'(x_t), x_t^+ - x_t \rangle = f_{*,t} - f(x_t) \leq f_*^t - f(x_t) (\leq f_* - f(x_t)). \tag{38}$$

Observing that for a generic GC algorithm we have (cf. (14), (15))  $f(x_{t+1}) \leq f(x_t + \gamma_t(x_t^+ - x_t))$ ,  $\gamma_t = \frac{2}{t+1}$ , and invoking (12), we have

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \gamma_t \langle f'(x_t), x_t^+ - x_t \rangle + \frac{L}{2} \gamma_t^2 \|x_t^+ - x_t\|_X^2 \\ &\leq f(x_t) - \gamma_t (f(x_t) - f_*^t) + \frac{1}{2} L \gamma_t^2, \end{aligned} \tag{39}$$

where the concluding  $\leq$  is due to (38). Since  $f_*^{t+1} \geq f_*^t$ , it follows that

$$\epsilon_{t+1} \leq f(x_{t+1}) - f_*^t \leq (1 - \gamma_t) \epsilon_t + \frac{1}{2} L \gamma_t^2,$$

whence

$$\begin{aligned} \epsilon_{t+1} &\leq \epsilon_1 \prod_{i=1}^t (1 - \gamma_i) + \frac{1}{2} L \sum_{i=1}^t \gamma_i^2 \prod_{k=i+1}^t (1 - \gamma_k) \\ &= 2L \sum_{i=1}^t (i+1)^{-2} \prod_{k=i+1}^t \left(1 - \frac{2}{k+1}\right), \end{aligned}$$

where, by convention,  $\prod_{k=t+1}^t = 1$ . Noting that  $\prod_{k=i+1}^t (1 - \frac{2}{k+1}) = \prod_{k=i+1}^t \frac{k-1}{k+1} = \frac{i(i+1)}{t(t+1)}$ ,  $i = 1, \dots, t$ , we get

$$\epsilon_{t+1} \leq 2L \sum_{i=1}^t \frac{i(i+1)}{(i+1)^2 t(t+1)} \leq \frac{2Lt}{(t+1)^2} \leq 2L(t+2)^{-1}, \tag{40}$$

what is (18). □

### 8.2 Proof of Theorem 2

The proof, up to minor modifications, goes back to [22], see also [19,26]; we provide it here to make the paper self-contained. W.l.o.g. we can assume that we are in the nontrivial case (see description of the algorithm).

- 1<sup>0</sup>. As it was explained when describing the method, whenever stage  $s$  takes place, we have  $0 < \rho_1 \leq \rho_s \leq \rho_*$ , and  $\rho_{s-1} < \rho_s$ , provided  $s > 1$ . Therefore by the termination rule, the output  $\bar{\rho}, \bar{x}$  of the algorithm, if any, satisfies  $\bar{\rho} \leq \rho_*, f(\bar{x}) \leq \epsilon$ . Thus, (i) holds true, provided that the algorithm does terminate. Thus, all we need is to verify (ii) and (iii).
- 2<sup>0</sup>. Let us prove (ii). Let  $s \geq 1$  be such that stage  $s$  takes place. Setting  $X = K[\rho_s]$ , observe that  $X - X \subset \{x \in E : \|x\| \leq 2\rho_s\}$ , whence  $\|\cdot\| \leq 2\rho_s \|\cdot\|_X$ , and therefore the relation (4) implies the validity of (12) with  $L = 4\rho_s^2 L_f$ . Now, if stage  $s$  does not terminate in course of some number  $t$  steps, then, in the

notation from the description of the algorithm,  $f(\bar{x}_t) > \epsilon$  and  $f_*^t < \frac{3}{4}f(\bar{x}_t)$ , whence  $f(\bar{x}_t) - f_*^t > \epsilon/4$ . By Theorem 1.ii, the latter is possible only when  $4.5L/(t - 2) > \epsilon/4$ . Thus,  $t \leq \max \left[ 5, 2 + \frac{72\rho_*^2 L_f}{\epsilon} \right]$ . Taking into account that  $\rho_s \leq \rho_*$ , (ii) follows.

3<sup>0</sup>. Let us prove (iii). This statement is trivially true when the number of stages is 1. Assuming that it is not the case, let  $S \geq 1$  be such that the stage  $S + 1$  takes place. For every  $s = 1, \dots, S$ , let  $t_s$  be the last step of stage  $s$ , and let  $u_s, \ell^s(\cdot)$  be what in the notation from the description of stage  $s$  was denoted  $f(\bar{x}_{t_s})$  and  $\ell^{t_s}(\rho)$ . Thus,  $u_s > \epsilon$  is an upper bound on  $\text{Opt}(\rho_s)$ ,  $\ell_s := \ell^s(\rho_s)$  is a lower bound on  $\text{Opt}(\rho_s)$  satisfying  $\ell_s \geq 3u_s/4$ , and  $\ell^s(\cdot)$  is a piecewise linear convex in  $\rho$  lower bound on  $\text{Opt}(\rho)$ ,  $\rho \geq 0$ , and  $\rho_{s+1} > \rho_s$  is the smallest positive root of  $\ell^s(\cdot)$ . Let also  $-g_s$  be a subgradient of  $\ell^s(\cdot)$  at  $\rho_s$ . Note that  $g_s > 0$  due to  $\rho_{s+1} > \rho_s$  combined with  $\ell^s(\rho_s) > 0$ ,  $\ell^s(\rho_{s+1}) = 0$ , and by the same reasons combined with convexity of  $\ell^s(\cdot)$  we have

$$\rho_{s+1} - \rho_s \geq \ell_s/g_s, \tag{41}$$

and, as we have seen,

$$1 \leq s \leq S \Rightarrow \begin{cases} (a) u_s > \epsilon, \\ (b) u_s \geq \text{Opt}(\rho_s) \geq \ell_s \geq \frac{3}{4}u_s, \\ (c) \ell_s - g_s(\rho - \rho_s) \leq \text{Opt}(\rho), \rho \geq 0. \end{cases} \tag{42}$$

Assuming  $1 < s \leq S$  and applying (41), we get  $\rho_s - \rho_{s-1} \geq \frac{3}{4}u_{s-1}/g_{s-1}$ , whence, invoking (42),

$$u_{s-1} \geq \text{Opt}(\rho_{s-1}) \geq \ell_s + g_s[\rho_{s-1} - \rho_s] \geq \frac{3}{4}u_s + \frac{3}{4}u_{s-1}\frac{g_s}{g_{s-1}}.$$

The resulting inequality implies that  $\frac{u_s}{u_{s-1}} + \frac{g_s}{g_{s-1}} \leq \frac{4}{3}$ , whence  $\frac{u_s g_s}{u_{s-1} g_{s-1}} \leq (1/4)(4/3)^2 = 4/9$ . It follows that

$$\sqrt{u_s g_s} \leq (2/3)^{s-1} \sqrt{u_1 g_1}, \quad 1 \leq s \leq S. \tag{43}$$

Now, since the first iterate of the first stage is 0, we have  $u_1 \leq f(0)$ , while (42) applied with  $s = 1$  implies that  $f(0) = \text{Opt}(0) \geq \ell_1 + \rho_1 g_1 \geq \rho_1 g_1$ , whence  $u_1 g_1 \leq f(0)/\rho_1 = d$ . Further, by (41)  $g_s \geq \ell_s/(\rho_{s+1} - \rho_s) \geq \ell_s/\rho_* \geq \frac{3}{4}u_s/\rho_*$ , where the concluding inequality is given by (42). We see that  $u_s g_s \geq \frac{3}{4}u_s^2/\rho_* \geq \frac{3}{4}\epsilon^2/\rho_*$ . This lower bound on  $u_s g_s$  combines with the bound  $u_1 g_1 \leq d$  and with (43) to imply that

$$\epsilon \leq \sqrt{4/3}(2/3)^{s-1} \sqrt{d\rho_*}, \quad 1 \leq s \leq S.$$

Finally observe that by the definition of  $\rho_*$  and due to the fact that  $\|x[f'(0)]\| = 1$  in the nontrivial case, we have

$$\begin{aligned} 0 \leq f(\rho_*x[f'(0)]) &\leq f(0) + \rho_*\langle f'(0), x[f'(0)] \rangle + \frac{1}{2}L_f\rho_*^2 \\ &= f(0) - \rho_*d + \frac{1}{2}L_f\rho_*^2 \end{aligned}$$

(we have used (4) and the definition of  $d$ ), whence  $\rho_*d \leq f(0) + \frac{1}{2}L_f\rho_*^2$  and therefore

$$\epsilon \leq \sqrt{4/3}(2/3)^{s-1} \sqrt{f(0) + \frac{1}{2}L_f\rho_*^2}, \quad 1 \leq s \leq S.$$

Since this relation holds true for every  $S \geq 1$  such that the stage  $S + 1$  takes place, (iii) follows. □

### 8.3 Proof of Theorem 3

By definition of  $z_t$  we have  $z_t \in K^+$  for all  $t$  and  $F(0) = F(z_1) \geq F(z_2) \geq \dots$ , whence  $r_t \leq D_*$  for all  $t$  by Assumption A. Besides this,  $r_* \leq D_*$  as well. Let now  $\epsilon_t = F(z_t) - F_*$ ,  $z_t = [x_t; r_t]$ , and let  $z_t^+ = [x_t^+, r_t^+]$  be a minimizer, as given by Lemma 1, of the linear form  $\langle F'(z_t), z \rangle$  of  $z \in E^+$  over the set  $K^+[r_*] = \{[x; r] : x \in K, \|x\| \leq r \leq r_*\}$ . Recalling that  $F'(z_t) = [f'(x_t); \kappa]$  and that  $r_t \leq D_* \leq \bar{D}$ , Lemma 1 implies that  $z_t^+ \in \Delta(z_t)$ . By definition of  $z_t^+$  and convexity of  $F$  we have

$$\begin{aligned} \langle [f'(x_t); \kappa], z_t - z_t^+ \rangle &= \langle f'(x_t), x_t - x_t^+ \rangle + \kappa(r_t - r_t^+) \\ &\geq \langle f'(x_t), x_t - x_* \rangle + \kappa(r_t - r_*) \\ &= \langle F'(z_t), z_t - z_* \rangle \geq F(z_t) - F(z_*) = \epsilon_t. \end{aligned}$$

Invoking (12), it follows that for  $0 \leq s \leq 1$  one has

$$\begin{aligned} F(z_t + s(z_t^+ - z_t)) &\leq F(z_t) + s\langle [f'(x_t); \kappa], z_t^+ - z_t \rangle + \frac{L_f s^2}{2} \|x(z_t^+) - x(z_t)\|^2 \\ &\leq F(z_t) - s\epsilon_t + \frac{1}{2}L_f s^2 (r_t + D_*)^2 \end{aligned}$$

using that  $\|x(z_t^+)\| \leq r_t^+$  and  $\|x(z_t)\| \leq r_t$  due to  $z_t^+, z_t \in K^+$ , and that  $r_t^+ \leq r_* \leq D_*$ . By (24) we have

$$F(z_{t+1}) \leq \min_{0 \leq s \leq 1} F(z_t + s(z_t^+ - z_t)) \leq F(z_t) + \min_{0 \leq s \leq 1} \left\{ -s\epsilon_t + \frac{1}{2}L_f s^2 (r_t + D_*)^2 \right\},$$

and we arrive at the recurrence

$$\epsilon_{t+1} \leq \epsilon_t - \begin{cases} \frac{\epsilon_t^2}{2L_f(r_t + D_*)^2}, & \epsilon_t \leq L_f(r_t + D_*)^2 \\ \epsilon_t - \frac{1}{2}L_f(r_t + D_*)^2, & \epsilon_t > L_f(r_t + D_*)^2 \end{cases}, \quad t = 1, 2, \dots \quad (44)$$

When  $t = 1$ , this recurrence, in view of  $z_1 = 0$ , implies that  $\epsilon_2 \leq \frac{1}{2}L_f D_*^2$ . Let us show by induction in  $t \geq 2$  that

$$\epsilon_t \leq \bar{\epsilon}_t := \frac{8L_f D_*^2}{t + 14}, \quad t = 2, 3, \dots \tag{45}$$

thus completing the proof. We have already seen that (45) is valid for  $t = 2$ . Assuming that (45) holds true for  $t = k \geq 2$ , we have  $\epsilon_k \leq \frac{1}{2}L_f D_*^2$  and therefore  $\epsilon_{k+1} \leq \epsilon_k - \frac{1}{8L_f D_*^2} \epsilon_k^2$  by (44) combined with  $0 \leq r_k \leq D_*$ . Now, the function  $s - \frac{1}{8L_f D_*^2} s^2$  is nondecreasing on the segment  $0 \leq s \leq 4L_f D_*^2$  which contains  $\bar{\epsilon}_k$  and  $\epsilon_k \leq \bar{\epsilon}_k$ , whence

$$\begin{aligned} \epsilon_{k+1} &\leq \epsilon_k - \frac{1}{8L_f D_*^2} \epsilon_k^2 \leq \bar{\epsilon}_k - \frac{1}{8L_f D_*^2} \bar{\epsilon}_k^2 = \left[ \frac{8L_f D_*^2}{k + 14} \right] - \frac{1}{8L_f D_*^2} \left[ \frac{8L_f D_*^2}{k + 14} \right]^2 \\ &= \frac{8L_f D_*^2 (k + 13)}{(k + 14)^2} \leq \frac{8L_f D_*^2}{(k + 1) + 14}, \end{aligned}$$

so that (45) holds true for  $t = k + 1$ . □

### 8.4 Proofs for Section 6

As we have already explained, (31) is solvable, so that  $z$  is well defined. Denoting by  $(s_*, r_*)$  an optimal solution to (31) produced, along with  $z$ , by our solver, note that the characteristic property of  $z$  is the relation

$$(s_*, r_*) \in \underset{s, r}{\text{Argmax}} \{s + \langle z, Pr - s\eta \rangle : 0 \leq r \leq \mathbf{e}\}.$$

Since the column sums in  $P$  are zeros and the sum of entries in  $\eta$  is zero, the above characteristic property of  $z$  is preserved when passing from  $z$  to  $\bar{z}$ , so that we may assume from the very beginning that  $z = \bar{z}$  is a zero mean image. Now,  $P = [Q, -Q]$ , where  $Q$  is the incidence matrix of the network obtained from  $G$  by eliminating backward arcs. Representing a flow  $r$  as  $[r_f; r_b]$ , where the blocks are comprised, respectively, of flows in the forward and backward arcs, and passing from  $r$  to  $\rho = r_f - r_b$ , our characteristic property of  $z$  clearly implies the relation

$$(s_*, \rho_* := r_{*f} - r_{*b}) \in \underset{s, \rho}{\text{Argmax}} \{s + \underbrace{\langle z, Q\rho - s\eta \rangle}_{\psi(s, \rho)} : \|\rho\|_\infty \leq 1\}. \tag{46}$$

By the optimality conditions in linear programming it follows that

- (a)  $\langle z, \eta \rangle = 1$ ,
  - (b)  $\|\rho_*\|_\infty \leq 1$ ,
  - (c)  $(Q^T z)_\gamma = \begin{cases} \leq 0, [\rho_*]_\gamma = -1, \\ = 0, [\rho_*]_\gamma \in (-1, 1), \\ \geq 0, [\rho_*]_\gamma = 1, \end{cases}$  for all forward arcs  $\gamma$ ,
  - (d)  $Q\rho_* = s_*\eta$ .
- (47)



Indeed, (a) stems from the fact that  $\psi(s, \rho)$ , which is affine in  $s$ , is above bounded, so that the coefficient at  $s$  in  $\psi$  should be zero; (b) is the constraint in the maximization problem in (46) to which  $(s_*, \rho_*)$  is an optimal solution; (c) is the optimality condition for the same problem w.r.t. the  $\rho$ -variable; and (d) expresses the fact that  $(s_*, r_*)$  is feasible for (31). (47.d) and (47.a) imply that  $\langle Q^T z, \rho_* \rangle = s_*$ , while (47.c) says that  $\langle Q^T z, \rho_* \rangle = \|Q^T z\|_1$ , and  $s_* = \|Q^T z\|_1$ . By (47.a)  $z \neq 0$ , and thus  $z$  is a nonzero image with zero mean; recalling what  $Q$  is, the first  $n(n-1)$  entries in  $Q^T z$  form  $\nabla_i z$ , and the last  $n(n-1)$  entries form  $\nabla_j z$ , so that  $\|Q^T z\|_1 = \text{TV}(z)$ . The gradient field of a nonzero image with zero mean cannot be identically zero, whence  $\text{TV}(z) = \|Q^T z\|_1 = s_* > 0$ . Thus  $x[\eta] = -z/\text{TV}(z) = -z/s_*$  is well defined and  $\text{TV}(x[\eta]) = 1$ , while by (47.a) we have  $\langle x[\eta], \eta \rangle = -1/s_*$ . Finally, let  $x \in \mathcal{TV}$ , implying that  $Q^T x$  is the concatenation of  $\nabla_i x$  and  $\nabla_j x$  and thus  $\|Q^T x\|_1 = \text{TV}(x) \leq 1$ . Invoking (47.b, d), we get  $-1 \leq \langle Q^T x, \rho_* \rangle = \langle x, Q\rho_* \rangle = s_* \langle x, \eta \rangle$ , whence  $\langle x, \eta \rangle \geq -1/s_* = \langle x[\eta], \eta \rangle$ , meaning that  $x[\eta] \in \mathcal{TV}$  is a minimizer of  $\langle \eta, x \rangle$  over  $x \in \mathcal{TV}$ .  $\square$

In the sequel, for a real-valued function  $x$  defined on a finite set (e.g., for an image),  $\|x\|_p$  stands for the  $L_p$  norm of the function corresponding to the counting measure on the set (the mass of every point from the set is 1). Let us fix  $n$  and  $x \in M_0^n$  with  $\text{TV}(x) \leq 1$ ; we want to prove that

$$\|x\|_2 \leq C\sqrt{\ln(n)} \tag{48}$$

with appropriately selected absolute constant  $C$ .

**1<sup>0</sup>.** Let  $\oplus$  stand for addition, and  $\ominus$  for subtraction of integers modulo  $n$ ;  $p \oplus q = (p+q) \bmod n \in \{0, 1, \dots, n-1\}$  and similarly for  $p \ominus q$ . Along with discrete partial derivatives  $\nabla_i x, \nabla_j x$ , let us define their periodic versions  $\widehat{\nabla}_i x, \widehat{\nabla}_j x$ :

$$\begin{aligned} \widehat{\nabla}_i x(i, j) &= x(i \oplus 1, j) - x(i, j) : \Gamma_{n,n} \rightarrow \mathbf{R}, \quad \widehat{\nabla}_j x(i, j) \\ &= x(i, j \oplus 1) - x(i, j) : \Gamma_{n,n} \rightarrow \mathbf{R}, \end{aligned}$$

same as periodic Laplacian  $\widehat{\Delta}x$ :

$$\widehat{\Delta}x = x(i, j) - \frac{1}{4} [x(i \ominus 1, j) + x(i \oplus 1, j) + x(i, j \ominus 1) + x(i, j \oplus 1)] : \Gamma_{n,n} \rightarrow \mathbf{R}.$$

For every  $j, 0 \leq j < n$ , we have  $\sum_{i=0}^{n-1} \widehat{\nabla}_i x(i, j) = 0$  and  $\nabla_i x(i, j) = \widehat{\nabla}_i x(i, j)$  for  $0 \leq i < n-1$ , whence  $\sum_{i=0}^{n-1} |\widehat{\nabla}_i(x)| \leq 2 \sum_{i=0}^{n-1} |\nabla_i x(i, j)|$  for every  $j$ , and thus  $\|\widehat{\nabla}_i x\|_1 \leq 2\|\nabla_i x\|_1$ . Similarly,  $\|\widehat{\nabla}_j x\|_1 \leq 2\|\nabla_j x\|_1$ , and we conclude that

$$\|\widehat{\nabla}_i x\|_1 + \|\widehat{\nabla}_j x\|_1 \leq 2. \tag{49}$$

**2<sup>0</sup>.** Now observe that for  $0 \leq i, j < n$  we have

$$\begin{aligned} x(i, j) &= x(i \ominus 1, j) + \widehat{\nabla}_i x(i \ominus 1, j) \\ x(i, j) &= x(i \oplus 1, j) - \widehat{\nabla}_i x(i, j) \\ x(i, j) &= x(i, j \ominus 1) + \widehat{\nabla}_j x(i, j \ominus 1) \\ x(i, j) &= x(i, j \oplus 1) - \widehat{\nabla}_j x(i, j) \end{aligned}$$

whence

$$\widehat{\Delta}x(i, j) = \frac{1}{4} [\widehat{\nabla}_i x(i \ominus 1, j) - \widehat{\nabla}_i x(i, j) + \widehat{\nabla}_j x(i, j \ominus 1) - \widehat{\nabla}_j x(i, j)] \quad (50)$$

Now consider the following linear mapping from  $M^n \times M^n$  into  $M^n$ :

$$B[g, h](i, j) = \frac{1}{4} [g(i \ominus 1, j) - g(i, j) + h(i, j \ominus 1) - h(i, j)], [i; j] \in \Gamma_{n,n}. \quad (51)$$

From this definition and (50) it follows that

$$\widehat{\Delta}x = B[\widehat{\nabla}_i x, \widehat{\nabla}_j x]. \quad (52)$$

Let for  $u \in M^n$ ,  $DFT[u]$  stand for the 2D Discrete Fourier Transform of  $u$ :

$$DFT[u](p, q) = \sum_{0 \leq r, s < n} u(r, s) \exp\{-2\pi i(pr + qs)/n\}, [p; q] \in \Gamma_{n,n}.$$

Note that every image  $u$  with zero mean is the periodic Laplacian of another, uniquely, defined, image  $X[u]$  with zero mean, with  $X[u]$  given by its Fourier transform

$$DFT[X[u]](p, q) = Y[u](p, q) := \begin{cases} 0, & p = q = 0 \\ \frac{DFT[u](p,q)}{D(p,q)}, & 0 \neq [p; q] \in \Gamma_{n,n} \end{cases}, [p; q] \in \Gamma_{n,n},$$

$$D(p, q) = 1 - \frac{1}{2} [\cos(2\pi p/n) + \cos(2\pi q/n)], [p; q] \in \Gamma_{n,n}. \quad (53)$$

Indeed, representing an  $n \times n$  image  $x(\mu, \nu)$ ,  $0 \leq \mu, \nu < n$ , as a Fourier sum

$$x(\mu, \nu) = \sum_{0 \leq p, q < n} c_{p,q} \exp\{2\pi i[p\mu/n + q\nu/n]\},$$

we get

$$\begin{aligned} [\widehat{\Delta}x](\mu, \nu) &= \sum_{0 \leq p, q < n} c_{p,q} \left[ \exp\{2\pi i[\frac{p\mu}{n} + \frac{q\nu}{n}]\} \right. \\ &\quad - \frac{1}{4} \exp\{2\pi i[\frac{p(\mu \ominus 1)}{n} + \frac{q\nu}{n}]\} - \frac{1}{4} \exp\{2\pi i[\frac{p(\mu \oplus 1)}{n} + \frac{q\nu}{n}]\} \\ &\quad \left. - \frac{1}{4} \exp\{2\pi i[\frac{p\mu}{n} + \frac{(q \ominus 1)\nu}{n}]\} - \frac{1}{4} \exp\{2\pi i[\frac{p\mu}{n} + \frac{(q \oplus 1)\nu}{n}]\} \right] \\ &= \sum_{0 \leq p, q < n} c_{p,q} \exp\{2\pi i[\frac{p\mu}{n} + \frac{q\nu}{n}]\} \left[ 1 - \frac{1}{4} \exp\{-2\pi i\frac{p}{n}\} - \frac{1}{4} \exp\{2\pi i\frac{p}{n}\} \right. \\ &\quad \left. - \frac{1}{4} \exp\{-2\pi i\frac{q}{n}\} - \frac{1}{4} \exp\{2\pi i\frac{q}{n}\} \right] \\ &= \sum_{0 \leq p, q < n} [c_{p,q} D(p, q)] \exp\{2\pi i[\frac{p\mu}{n} + \frac{q\nu}{n}]\}, \end{aligned}$$

where

$$\begin{aligned}
 D(p, q) &= \left[ 1 - \frac{1}{4} \exp\{-2\pi i \frac{p}{n}\} - \frac{1}{4} \exp\{2\pi i \frac{p}{n}\} - \frac{1}{4} \exp\{-2\pi i \frac{q}{n}\} - \frac{1}{4} \exp\{2\pi i \frac{q}{n}\} \right] \\
 &= 1 - \frac{1}{2} \cos(2\pi \frac{p}{n}) - \frac{1}{2} \cos(2\pi \frac{q}{n}), \quad [p; q] \in \Gamma_{n,n}.
 \end{aligned}$$

In other words, in the Fourier domain passing from an image to its periodic Laplacian means multiplication of Fourier coefficient  $c_{p,q}$  by  $D(p, q)$ . From the expression for  $D(p, q)$  it is immediately seen that  $D(p, q)$  is nonzero whenever  $(p, q)$  with  $0 \leq p, q < n$  is nonzero. In other words, the periodic Laplacian of a whatever  $n \times n$  zero mean image is an image with zero mean (zero Fourier coefficient  $(0, 0)$ ), and every image of this type is a periodic Laplacian of another zero mean image described in (53).

In particular, invoking (52), we get

$$\text{DFT}[x] = Y[B[\widehat{\nabla}_i x, \widehat{\nabla}_j x]].$$

By Parseval identity,  $\|\text{DFT}[x]\|_2 = n\|x\|_2$ , whence

$$\|x\|_2 = n^{-1} \|Y[B[\widehat{\nabla}_i x, \widehat{\nabla}_j x]]\|_2.$$

Combining this observation with (49), we see that in order to prove (48), it suffices to check that

(!) Whenever  $g, h \in M^n$  are such that

$$(g, h) \in G := \{(g, h) \in M^n \times M^n : \|g\|_1 + \|h\|_1 \leq 2\},$$

we have

$$\|Y[B[g, h]]\|_2 \leq nC\sqrt{\ln(n)}. \tag{54}$$

4<sup>0</sup>. The good news about (!) is that  $Y[B[g, h]]$  is linear in  $(g, h)$ . Therefore, in order to justify (!), it suffices to prove that (54) holds true for the extreme point of  $G$ , i.e., (a) for pairs where  $h \equiv 0$  and  $g$  is an image which is equal to 2 at some point of  $\Gamma_{n,n}$  and vanishes outside of this point, and (b) for pairs where  $g \equiv 0$  and  $h$  is an image which is equal to 2 at some point of  $\Gamma_{n,n}$  and vanishes outside of this point. Task (b) clearly reduces to task (a) by swapping the coordinates  $i, j$  of points from  $\Gamma_{n,n}$ , so that we may focus solely on task (a). Thus, assume that  $g$  is a cyclic shift of the image  $2\delta$ :

$$g(i, j) \equiv 2\delta(i \ominus r, j \ominus s), \quad \delta(i, j) = \begin{cases} 1, & [i; j] = [0; 0] \\ 0, & [i; j] \neq [0; 0] \end{cases}, \quad [i; j] \in \Gamma_{n,n}.$$

From (51) it follows that then  $B[g, 0]$  is a cyclic shift of  $B[2\delta, 0]$ , whence  $|\text{DFT}[B[g, 0]](p, q)| = |\text{DFT}[B[2\delta, 0]](p, q)|$  for all  $[p; q] \in \Gamma_{n,n}$ , which, by (53),

implies that  $|Y[B[g, 0]](p, q)| = |Y[B[2\delta, 0]](p, q)|$  for all  $[p; q] \in \Gamma_{n,n}$ . The bottom line is that *all we need is to verify that (54) holds true for  $g = 2\delta, h = 0$ , or, which is the same, that with*

$$y(p, q) = \frac{(1 - \exp\{2\pi i p/n\})}{2[1 - \frac{1}{2}[\cos(2\pi p/n) + \cos(2\pi q/n)]]} \tag{55}$$

where the right hand side by definition is 0 at  $p = q = 0$ , it holds

$$C_n := \sum_{p,q=0}^{n-1} |y(p, q)|^2 \leq n^2 C^2 \ln(n).$$

Now, (55) makes sense for all  $[p; q] \in \mathbf{Z}^2$  (provided that we define the right hand side as zero at all points of  $\mathbf{Z}^2$  where the denominator in (55) vanishes, that is, at all point where  $p, q$  are integer multiples of  $n$ ) and defines  $y$  as a double-periodic, with periods  $n$  in  $p$  and in  $q$ , function of  $[p; q]$ . Therefore, setting  $m = \lfloor n/2 \rfloor \geq 1$  and  $W = \{[p; q] \in \mathbf{Z}^2 : -m \leq p, q < n - m\}$ , we have

$$C_n = \sum_{0 \neq [p; q] \in W} |y(p, q)|^2 = \sum_{[p; q] \in W} \frac{|1 - \exp\{2\pi i p/n\}|^2}{4|1 - \frac{1}{2}[\cos(2\pi p/n) + \cos(2\pi q/n)]|^2}.$$

Setting  $\rho(p, q) = \sqrt{p^2 + q^2}$ , observe that when  $0 \neq [p; q] \in W$ , we have  $|1 - \exp\{2\pi i p/n\}| \leq C_1 n^{-1} \rho(p, q)$  and  $2|1 - \frac{1}{2}[\cos(2\pi p/n) + \cos(2\pi q/n)]| \geq C_2 n^{-2} \rho^2(p, q)$  with positive absolute constants  $C_1, C_2$ , whence

$$C_n \leq (C_1/C_2)^2 \sum_{0 \neq [p; q] \in W} n^2 \rho^{-2}(p, q).$$

With appropriately selected absolute constant  $C_3$  we have

$$\sum_{0 \neq [p; q] \in W} \rho^{-2}(p, q) \leq C_3 \int_1^n r^{-2} r dr = C_3 \ln(n).$$

Thus,  $C_n \leq (C_1/C_2)^2 C_3 n^2 \ln(n)$ , meaning that (54), and thus (48), holds true with  $C = \sqrt{C_3} C_1/C_2$ . □

We are very grateful to Referees for their reading of the manuscript and thoughtful comments.

We have included after Theorem 2 a comment on comparison of the proposed parametric optimization algorithm with that based on bisection. However, we finally decided not to include the numerical comparison. Note that including such experiment would require explaining in some detail what the ‘‘bisection’’ algorithm is and how is it implemented, what would make the manuscript which is already too long with respect to the size requirements of the *Mathematical Programming* even longer.

We would like to thank Referee 2 for detailed and clarifying explanations of his comments on the initial manuscript. Some comments of Referee 2 concern notation we have not modified in our revision. Since his requests do not seem to be compulsory and the final decision is left to us, we have modified some (e.g., we replaced  $D^+$  with  $\bar{D}$ ) but preferred to keep other unchanged.

In what follows we provide our answers to other questions raised in his detailed report.

- 6.16\* *...I would suggest that the authors simply say, when introducing (12), that is it is implied by (4) but more general, to give some hint about why they are mentioning it*

We have added a small comment after display (12) to make clear the choice of the norm  $\|\cdot\|_X$  here

- 11.9\* *...some readers might be more familiar with a statement like “ $\mathcal{A}$  has full column rank,” or “ $\mathcal{A}$  has a trivial nullspace,” etc.*

We think that for a linear operator (which, in our case, may not be a matrix) a commonly adopted terminology is indeed “ $\mathcal{A}$  has a trivial nullspace,” or “ $\mathcal{A}$  has a trivial kernel”, what is the exact wording of  $\text{Ker}(\mathcal{A}) = \{0\}$ .

- 11.19\* *...but I think it would suffice to simply note in the text that the given value  $\sum |\lambda_\zeta| \rho$  is an upper bound on  $\|x\|$  for  $x$  in the given form.*

Thank you, we have added a comment in this sense after display (29).

- 30.5\* *...I'll say that the authors are practicing a false economy in not giving a more complete explanation, and that at the least, “we immediately see” should be “one can show”*

Thank you for insisting, you are probably right—we have reproduced in the text our answer to the corresponding comment on the previous version of the manuscript.

## References

1. Andersen, E.D., Andersen, K.D.: The MOSEK optimization tools manual. [http://www.mosek.com/fileadmin/products/6\\_0/tools/doc/pdf/tools.pdf](http://www.mosek.com/fileadmin/products/6_0/tools/doc/pdf/tools.pdf)
2. Bach, F., Jenatton, R., Mairal, J., Obozinski, G. et al.: Convex optimization with sparsity-inducing norms. In: Sra, S., Nowozin, S., Wright, S. J. (eds). Optimization for Machine Learning, pp. 19–53. MIT Press
3. Cai, J.-F., Candes, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20**(4), 1956–1982 (2008)
4. Candès, E., Recht, B.: Exact matrix completion via convex optimization. *Found. Comput. Math.* **9**(6), 717–772 (2009)
5. Cox, B., Juditsky, A., Nemirovski, A.: Dual subgradient algorithms for large-scale nonsmooth learning problems. *Math. Program.* 1–38 (2013). doi:[10.1007/s10107-013-0725-1](https://doi.org/10.1007/s10107-013-0725-1)
6. Demyanov, V., Rubinov, A.: Approximate Methods in Optimization Problems. American Elsevier, Amsterdam (1970)
7. Dudik, M., Harchaoui, Z., Mallick, J.: Lifted coordinate descent for learning with trace-norm regularization. In: AISTATS (2012)
8. Dunn, J.C., Harshbarger, S.: Conditional gradient algorithms with open loop step size rules. *J. Math. Anal. Appl.* **62**(2), 432–444 (1978)
9. Frank, M., Wolfe, P.: An algorithm for quadratic programming. *Naval Res. Logist. Q.* **3**, 95–110 (1956)
10. Goldfarb, D., Ma, S., Wen, Z.: Solving low-rank matrix completion problems efficiently. In: Proceedings of 47th Annual Allerton Conference on Communication, Control, and Computing (2009)

11. Harchaoui, Z., Douze, M., Paulin, M., Dudik, M., Malick, J.: Large-scale image classification with trace-norm regularization. In: CVPR (2012)
12. Harchaoui, Z., Juditsky, A., Nemirovski, A.: Conditional gradient algorithms for machine learning. In: NIPS Workshop on Optimization for Machine Learning. <http://opt.kyb.tuebingen.mpg.de/opt12/papers.html> (2012)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, Berlin (2008)
14. Hazan, E.: Sparse approximate solutions to semidefinite programs. In: Proceedings of the 8th Latin American Conference Theoretical Informatics, pp. 306–316 (2008)
15. Hearn, D., Lawphongpanich, S., Ventura, J.: Restricted simplicial decomposition: computation and extensions. *Math. Program. Stud.* **31**, 99–118 (1987)
16. Holloway, C.: An extension of the Frank-Wolfe method of feasible directions. *Math. Program.* **6**, 14–27 (1974)
17. Jaggi, M.: Revisiting Frank-Wolfe: projection-free sparse convex optimization. In: ICML (2013)
18. Jaggi, M., Sulovsky, M.: A simple algorithm for nuclear norm regularized problems. In: ICML (2010)
19. Juditsky, A., Karzan, F.K., Nemirovski, A.: Randomized first order algorithms with applications to  $\ell_1$ -minimization. *Math. Program.* **142**(1–2), 269–310 (2013)
20. Juditsky, A., Nemirovski, A.: First order methods for nonsmooth large-scale convex minimization, i: general purpose methods; ii: utilizing problem’s structure. In: Sra, S., Nowozin, S., Wright, S. (eds). *Optimization for Machine Learning*, pp. 121–184. The MIT Press, Cambridge (2012)
21. Lan, G.: An optimal method for stochastic composite optimization. *Math. Program.* **133**(1–2), 365–397 (2012)
22. Lemar chal, C., Nemirovskii, A., Nesterov, Y.: New variants of bundle methods. *Math. Program.* **69**(1–3), 111–147 (1995)
23. Ma, S., Goldfarb, D., Chen, L.: Fixed point and bregman iterative methods for matrix rank minimization. *Math. Program.* **128**, 321–353 (2011)
24. Nemirovski, A., Onn, S., Rothblum, U.G.: Accuracy certificates for computational problems with convex structure. *Math. Oper. Res.* **35**(1), 52–78 (2010)
25. Nemirovski, A.S., Yudin, D.B.: *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, New York (1983)
26. Nesterov, Y.: *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, Berlin (2003)
27. Nesterov, Y.: Gradient methods for minimizing composite functions. *Math. Program.* **140**(1), 125–161 (2013)
28. Nesterov, Y., Nemirovski, A.: On first-order algorithms for  $l_1$ /nuclear norm minimization. *Acta Numer.* **22**, 509–575 (2013)
29. Pshenichnyj, B., Danilin, Y.: *Numerical Methods in Extremal Problems*. Mir, Moscow (1978)
30. Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52**(3), 471–501 (2010)
31. Recht, B., R e, C.: Parallel stochastic gradient algorithms for large-scale matrix completion. *Math. Program. Comput.* **5**(2), 201–226 (2013)
32. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Phys. D* **60** (1992)
33. Shalev-Shwartz, S., Gonen, A., Shamir, O.: Large-scale convex minimization with a low-rank constraint. In: ICML (2011)
34. Sra, S., Nowozin, S., Wright, S.J.: *Optimization for Machine Learning*. MIT Press, Cambridge (2010)
35. Srebro, N., Shraibman, A.: Rank, trace-norm and max-norm. In: COLT (2005)
36. Ventura, J.A., Hearn, D.W.: Restricted simplicial decomposition for convex constrained problems. *Math. Program.* **59**, 71–85 (1993)
37. Yang, J., Yuan, X.: Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comput.* **82**(281), 301–329 (2013)
38. Zhang, X., Yu, Y., Schuurmans, D.: Accelerated training for matrix-norm regularization: a boosting approach. In: NIPS, pp. 2915–2923 (2012)
39. Zibulevski, M., Narkiss, G.: Sequential subspace optimization method for large-scale unconstrained problems. Technical Report CCIT No 559, Faculty of Electrical engineering, Technion (2005)