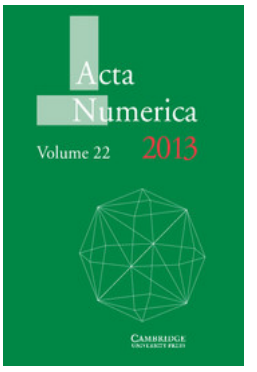

On first-order algorithms for l_1/nuclear norm minimization

Yurii Nesterov and Arkadi Nemirovski

Acta Numerica / Volume 22 / May 2013, pp 509 - 575
DOI: 10.1017/S096249291300007X, Published online: 02 April 2013

Link to this article: http://journals.cambridge.org/abstract_S096249291300007X

How to cite this article:
Yurii Nesterov and Arkadi Nemirovski (2013). On first-order algorithms for l_1/nuclear norm minimization. Acta Numerica, 22, pp 509-575 doi:10.1017/S096249291300007X

Acta Numerica (2013), pp. 509–575
doi:10.1017/S096249291300007X

On first-order algorithms for ℓ_1/nuclear norm minimization

Yurii Nesterov*
Center for Operations Research and Econometrics,
34 voie du Roman Pays, 1348, Louvain-la-Neuve, Belgium
E-mail: yurii.nesterov@uclivain.be

Arkadi Nemirovski†
School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, Georgia 30332, USA
E-mail: arkadi.nemirovski@isye.gatech.edu

In the past decade, problems related to l_1/nuclear norm minimization have attracted much attention in the signal processing, machine learning and optimization communities. In this paper, devoted to ℓ_1/nuclear norm minimization as 'optimization beasts', we give a detailed description of two attractive first-order optimization techniques for solving problems of this type. The first one, aimed primarily at lasso-type problems, comprises fast gradient methods applied to composite minimization formulations. The second approach, aimed at Dantzig-selector-type problems, utilizes saddle-point first-order algorithms and reformulation of the problem of interest as a *generalized bilinear saddle-point problem.* For both approaches, we give complete and detailed complexity analyses and discuss the application domains.

CONTENTS

* Partly supported by Direction de la Recherche Scientifique, Communauté Française de Belgique, via grant Action de Recherche Concerté ARC 04/09-315, and by the Laboratory of Structural Methods of Data Analysis in Predictive Modelling via the RF government grant 11.G34.31.0073c.
† Supported in part by NSF via grants DMS-0914785 and CMMI 1232623, by BSF via grant 2008302, and by ONR via grant N000140811104.

1. Introduction

1.1. Motivation

The goal of this paper is to review recent progress in applications of first-order optimization methods to problems of ℓ_1/nuclear norm minimization. For us, the problems of primary interest are the *lasso-type problems*,

$$\mathrm{Opt} = \min_{x \in E} \big\{ \lambda \|x\| + \|Ax - b\|_2^2 \big\}, \tag{1.1a}$$

and the *Dantzig-selector-type problems*,

$$\mathrm{Opt} = \min_{x \in E} \big\{ \|x\| : \|Ax - b\|_p \le \delta \big\}. \tag{1.1b}$$

Here E is a Euclidean space, $\|\cdot\|$ is a given norm on E (often different from the Euclidean norm), $x \mapsto Ax$ is a linear mapping from E to some $\mathbb{R}^m$, and $b \in \mathbb{R}^m$ is a given vector. In (1.1a), $\lambda > 0$ is a given *penalty parameter*. In (1.1b), $\delta \ge 0$, and p is either 2 ('ℓ_2-fit'), or ∞ ('uniform fit').[1] We will be mainly (but not exclusively) interested in the following cases:

- ℓ_1 *minimization*, where $E = \mathbb{R}^n$ and $\|\cdot\| = \|\cdot\|_1$,
- *nuclear norm minimization*, where $E = \mathbb{R}^{\mu \times \nu}$ is the space of $\mu \times \nu$ matrices and $\|\cdot\| = \|\cdot\|_{\mathrm{nuc}}$ is the *nuclear norm*, the sum of the singular values of a matrix.

The main source of ℓ_1 minimization problems of the form (1.1a,b) is *sparsity-oriented signal processing*, where the goal is to recover an unknown 'signal' $x \in \mathbb{R}^n$ from its noisy observations $y = Bx + \xi$, where ξ is the observation error. The signal is known to be *sparse*, *i.e.*, it has at most $s \ll n$ non-zero entries. The most interesting case here is *compressive sensing*, where the observations are 'deficient' ($m = \dim y < n = \dim x$). Hence the observations themselves, even noiseless ones, do not allow recovery of x unless some additional information on x is utilized, *e.g.*, sparsity.

The conceptually simplest way to recover a sparse x from noiseless observations is offered by ℓ_0 minimization, where we minimize the ℓ_0 'norm' of $x \in \mathbb{R}^n$ (*i.e.*, the number of non-zero entries in x) under the restriction $Bx = y$. It is immediately seen that if $y = Bx_*$ for some x_* with at most s non-zero entries, and if every $2s$ columns of B are linearly independent, then the ℓ_0 minimization does indeed recover x_*. The independence condition is true, for example, for any $m \times n$ matrix B in 'general position' provided that $m \ge 2s$.

[1] The restriction $p \in \{2, \infty\}$ is met in all applications of Dantzig selector models known to us. However, the following results can be straightforwardly extended to the case of $p \in [2, \infty]$ (but *not* to the case of $1 \le p < 2$).

This procedure, however, is heavily computationally intractable. To make it tractable, we can replace the ℓ_0 norm by its 'closest' convex approximation $\|\cdot\|_1$. Thus we come to the recovery problem $\widehat{x} \in \arg\min_x\{\|x\|_1 : Bx = y\}$, *i.e.*, problem (1.1b) with $A = B$, $b = y$ and $\delta = 0$.

A natural way to account for the observation noise ξ having $\|\cdot\|_2$ magnitude not exceeding a given δ is by passing from minimization of $\|\cdot\|_1$ on the subspace $Bx = y$ to minimization of this norm over the set $\{x : \|Bx - y\|_2 \le \delta\}$ of all signals compatible with our observations. Thus, we come to problem (1.1b) with $p = 2$. Note that the lasso problem (1.1a) with $\|\cdot\| = \|\cdot\|_1$ can be viewed as a re-parametrization of (1.1b) with $p = 2$. Indeed, geometrically, the curve of solutions to (1.1b) parametrized by δ is the same as the curve of solutions to (1.1a) parametrized by λ. In some situations (*e.g.*, when the magnitude of the observation error is not known), the latter parametrization is better than the former one.

For random observation error ξ (*e.g.*, white noise $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$), it would be too conservative to measure the noise in terms of its $\|\cdot\|_2$ magnitude, since this approach results in growth of δ with m as $\sigma\sqrt{m}$, and essentially ignores the random nature of the noise. From the statistical viewpoint, in this case it is much better to pass to 'artificial' observations $b = H^T y$, where H is a properly selected 'contrast matrix' with columns $h^1, \dots, h^K \in \mathbb{R}^m$. Then $b = (H^T B)x + \tilde{\xi}$ with the *uniform* norm of the new observation error $\tilde{\xi} = H^T\xi$ of order $\sigma \max_j \|h^j\|_2$. We can now apply the above recovery routine to the artificial observations, to obtain the estimator

$$\widehat{x} \in \arg\min_x \left\{ \|x\|_1 : \big\|\underbrace{H^T B}_{A} x - b\big\|_\infty \le \delta = \delta(\epsilon) := \sigma \max_i \|h_i\|_2 \mathrm{Erf\,Inv}\left(\frac{\epsilon}{K}\right) \right\},$$

where $\mathrm{Erf\,Inv}(\cdot)$ is the inverse error function, and $1 - \epsilon$ is the desired confidence. Note that our $\delta(\epsilon)$, with probability at least $1 - \epsilon$, is an upper bound on the uniform norm $\|H^T\xi\|_\infty$ of the new observation error. This bound is 'nearly independent' of ϵ and the dimensions of the problem. The resulting recovery routine is exactly (1.1b) with $p = \infty$.[2]

Nuclear norm minimization can be viewed as the 'matrix analogue' of sparsity-oriented signal processing, with the $\mu \times \nu$ matrix x playing the role of the n-dimensional signal vector, and the rank of the matrix (*i.e.*, the ℓ_0 norm of the vector of its singular values) replacing the number of non-zero entries in a vector. As in the case for the usual sparsity, the nuclear norm approach can be viewed as the 'best convex approximation' of the rank, which gives rise to problems (1.1a,b) with $\|\cdot\| = \|\cdot\|_{\mathrm{nuc}}$.

[2] The actual Dantzig selector as defined by Candès and Tao (2007) corresponds to the situation when all columns in B are normalized to have unit $\|\cdot\|_2$ norm, and $H = B$. For other choices of the contrast matrix, see Juditsky and Nemirovski (2011*c*) and Juditsky, Kılınç Karzan, Nemirovski and Polyak (2013*b*).

1.2. Justification of ℓ_1/nuclear norm recovery

We have already mentioned that ℓ_1/nuclear norm minimizations are just 'convex approximations' of ℓ_0/rank minimization. This gives rise to the crucial question of whether and when these approximations can be 'justified', that is, whether and when they admit meaningful error bounds in the case of noisy observations and nearly sparse/low-rank signals. Here, 'meaningful' means that the recovery error should tend to zero as the magnitude of noise and the deviation of the signal from sparse/low rank tend to zero. While the 'justification issues' go beyond the scope of this paper, which is devoted to (1.1a,b) as 'optimization beasts', we believe it makes sense to comment on them very briefly. The numerous related results in the literature (see Candès 2006, Candès and Tao 2006 and Candès, Romberg and Tao 2006, and references therein) show that by imposing appropriate conditions on the sensing matrix, the outlined procedures can indeed be justified. A typical result is as follows.

Theorem (Juditsky *et al.* 2013*b*). Let s be a positive integer, and let a sensing matrix $B \in \mathbb{R}^{m\times n}$ satisfy the *restricted isometry property* with parameters κ and k, meaning that for every k-sparse signal x we have

$$(1-\kappa)\|x\|_2^2 \le x^T B^T B x \le (1+\kappa)\|x\|_2^2.$$

Let $x_* \in \mathbb{R}^n$ be the 'true signal', and let $y = Bx_* + \xi$ be the observation. Assume that $k \ge 2s$ and $\kappa \le 0.1$.

(i) Let the observation error ξ satisfy $\|\xi\|_2 \le \delta$ with a given $\delta \ge 0$. Then the optimal solution $\widehat{x}$ to the Dantzig selector problem

$$\min_x \big\{ \|x\|_1 : \|Bx - y\|_2 \le \delta \big\}$$

satisfies

$$\|\widehat{x} - x_*\|_r \le O(1)\big[s^{\frac{1}{r}-\frac{1}{2}}\delta + s^{\frac{1}{r}-1}\|x_* - x_*^s\|_1\big], \quad 1 \le r \le 2,$$

where x_*^s is the best s-sparse approximation of x_* obtained by replacing all entries of x_* with zeros except the s largest in magnitude.[3]

(ii) Let the observation error ξ be Gaussian, $\xi \sim \mathcal{N}(0, \sigma^2 I)$, and let $\epsilon \in (0, 1/2)$. Then the optimal solution $\widehat{x}$ to the Dantzig selector problem

$$\min_x \big\{ \|x\|_1 : \|B^T(Bx - y)\|_\infty \le \delta := 2\sigma\sqrt{2\ln(m/\epsilon)} \big\}$$

satisfies

$$\|\widehat{x} - x_*\|_r \le O(1)\big[s^{\frac{1}{r}}\delta + s^{\frac{1}{r}-1}\|x_* - x_*^s\|_1\big], \quad 1 \le r \le 2,$$

with probability at least $1 - \epsilon$.

[3] Here and in what follows, $O(1)$ stands for positive absolute constants.

It is also known (see, *e.g.*, Candès 2006, Candès and Tao 2006 and Candès, Romberg and Tao 2006) that for a 'good' probability distribution $\mathbb{P}$ on $\mathbb{R}^{m\times n}$ (*e.g.*, the entries are independent $\mathcal{N}(0, 1/m)$ random variables, or they take the values $\pm 1/\sqrt{m}$ independently with probability 1/2), if m and n grow, then the matrix drawn from $\mathbb{P}$ satisfies, with probability approaching one, the restricted isometry property with parameters $\kappa = 0.1$, and k as large as $O(1)m/\ln(2n/m)$.

Similar results hold for nuclear norm recovery, but with rank in the role of the sparsity parameter s, the Frobenius norm replacing $\|\cdot\|_2$, and $\|x_* - x_*^s\|_{\rm nuc}$ replacing $\|x_* - x_*^s\|_1$.

In problems (1.1a,b), the most frequently used norms are of course ℓ_1 and the nuclear norm. However, there are other important norms (the list below is not exhaustive).

- *Block ℓ_1/ℓ_2 norm* $\|x\| = \sum_{j=1}^n \|x^j\|_2$, where $x^j \in \mathbb{R}^{k_j}$ are non-overlapping blocks in the $(k_1 + \cdots + k_n)$-dimensional signal x. This norm is used when there are reasons to believe that the true signal is *block-sparse*, *i.e.*, the number of non-zero blocks in an *a priori* known partition of the signal is relatively small.
- *Total variation* (the so-called TV norm) is defined on the space $\mathbb{R}^{\mu\times\nu}$ of $\mu \times \nu$ images x, normalized to have $x_{1,1} = 0$. Then

$$\mathrm{TV}(x) = \sum_{i=1}^{\mu-1}\sum_{j=1}^{\nu} |x_{i+1,j} - x_{i,j}| + \sum_{i=1}^{\mu}\sum_{j=1}^{\nu-1} |x_{i,j+1} - x_{i,j}|.$$

  This is a discrete analogue of the L_1 norm of the gradient field of a function. It is very popular for de-noising and de-blurring images.

The literature on design, justification, and applications of problems (1.1a) and (1.1b) with the outlined norms is really huge and rapidly growing. To gain some impression of it, we refer the reader to the papers by Bach, Mairal and Ponce (2008), Baraniuk, Cevher, Duarte and Hegde (2010), Bickel, Ritov and Tsybakov (2008), Cai, Candès and Shen (2008), Candès (2006), Candès and Recht (2008), Candès and Tao (2006, 2007, 2009), Candès, Romberg and Tao (2006), Chandrasekaran, Recht, Parrilo and Willsky (2012), Chandrasekaran, Sanghavi, Parrilo and Willsky (2011), Chesneau and Hebiri (2008), Cohen, Dahmen and DeVore (2009), Donoho and Tanner (2005), Donoho, Elad and Temlyakov (2006), Eldar, Kuppinger and Bölcskei (2010), Elhamifar and Vidal (2012), Huang and Zhang (2010), Juditsky and Nemirovski (2011*a*, 2011*c*), Juditsky, Kılınç Karzan and Nemirovski (2011*a*), Juditsky *et al.* (2013*b*), Juditsky, Kılınç Karzan, Nemirovski and Polyak (2011*b*), Liu and Zhang (2009), Liu, Shang, Jiang and Liu (2010), Lounici, Pontil, van de Geer and Tsybakov (2011), Meier, van de Geer and Bühlmann (2008), Meinshausen and Yu (2009), Nardi and Rinaldo (2008),

Recht, Fazel and Parrilo (2010), Recht, Xuy and Hassibi (2011*b*), Rudin, Osher and Fatemi (1992), Tibshirani (1996), Tropp (2006) and Yuan and Lin (2006) for design and justification of recovery routines, and Herman and Strohmer (2007), Lustig, Donoho and Pauly (2007), Men *et al.* (2011), Dai, Sheikh, Milenkovic and Baraniuk (2009), Parvaresh, Vikalo, Misra and Hassibi (2008), Santosa and Symes (1986), Studer *et al.* (2012), Taylor, Banks and McCoy (1979), Vasanawala *et al.* (2010) and Wagner, Schmieder, Stern and Hoch (1993), and references therein, for applications in imaging, including biological and hyperspectral imaging, radar control, magnetic resonance, bioinformatics, design of photonic crystals, and inverse problems in geophysics, to name just a few.[4]

1.3. Algorithms

Problems (1.1a) and (1.1b) are well-structured convex optimization problems. Hence, they can be solved in a theoretically efficient fashion. For example, all these problems are within the 'theoretical grasp' of polynomial-time interior-point methods (IPMs) for linear programming (ℓ_1/total variation minimization with uniform fit), conic quadratic programming (ℓ_1/total variation minimization with ℓ_2 fit, and ℓ_1/ℓ_2 minimization) and semidefinite programming (nuclear norm minimization). The 'interior-point option' should definitely be kept in mind when processing (1.1a,b) numerically. The difficulty, however, is that numerous applications lead to extremely large-scale problems, with sizes of A in the range of tens and hundreds of thousands, or even millions. In addition, it is necessary to operate with highly dense analytically given sensing matrices A. For example, those matrices typically arising in compressive sensing are randomly selected submatrices of discrete Fourier/Hadamard transform matrices, or matrices with independent entries drawn from Gaussian or binomial distributions. As a result, in most large-scale applications our problems of interest go beyond the 'practical grasp' of IPMs with their time-consuming iterations (in the dense case their complexity has cubic dependence on design dimension).

There seems to exist a consensus that the only methods in the modern 'convex optimization toolbox' that are applicable to the large-scale problems of ℓ_1/nuclear norm minimization are the *first-order methods* (FOMs). As far as problems (1.1a,b) are concerned, one can define FOMs to be algorithms based only on computing matrix–vector products involving matrices A and A^T, and avoiding the matrix–matrix multiplications/Cholesky decompositions with 'large-scale' operands needed by IPMs. One can think of an FOM as an algorithm which, instead of direct access to the data matrix A, is allowed to call a 'multiplication oracle' which, given on input a

[4] In addition to our (by far incomplete) list of references, we refer the reader to the depository *Compressive Sensing Resources* at http://dsp.rice.edu/cs.

vector of appropriate dimension, returns its products with A or A^T. In the large-scale case, iterations of typical FOMs are vastly cheaper than those of IPMs.

This is the main argument in favour of FOMs in large-scale applications. In the context of ℓ_1/nuclear norm minimization problems, an additional argument is that the geometry of these problems is especially well suited to first-order minimization. As applied to such a problem, properly designed FOMs exhibit dimension-independent rates of convergence (or nearly so). Unfortunately, this rate of convergence is only sublinear. In contrast to the linear convergence of IPMs, the inaccuracy tends to zero with the number of iterations t as $O(1/t^2)$ at best, or even $O(1/t)$ (see below). However, in the majority of applications of ℓ_1/nuclear norm minimization, we are not interested in high-accuracy solutions, only in medium-accuracy ones. Thus the relatively slow convergence of FOMs is more than compensated for by their insensitivity to problem size. FOMs are especially attractive in the case of ℓ_1 minimization, where the computational effort per iteration reduces essentially to $O(1)$ matrix–vector multiplication involving A and A^T. As a result, any acceleration of matrix–vector multiplications allowed by sparsity or the special structure of A directly translates into overall acceleration of the solution process.[5] In order to gain an impression of what can be achieved in this way, look at the data in Table 1.1.

1.4. Scope and organization of the paper

In the literature we find an extremely wide spectrum of FOMs aimed at our problems of interest. They range from adaptations of general-purpose convex optimization techniques (*e.g.*, a multitude of proximal-point algorithms, numerous methods based on augmented Lagrangians and alternating directions, active set methods, *etc.*) capable of solving (1.1a,b) independently of any assumptions on sparsity of their optimal solutions, to highly specialized algorithms based on thresholding and matching pursuit, and aimed at recovery of *sparse/low-rank* near-optimal solutions. Some impression of the related literature can be obtained from Beck and Teboulle (2009*a*, 2009*b*), Becker, Candès and Grant (2010), Chambolle (2004), Goldfarb and Ma (2011), Goldfarb, Scheinberg and Xi (2011), Goldfarb and Yin (2009), Huang, Ma and Goldfarb (2013), Jaggi and Sulovský (2010), Lee *et al.* (2010), Liu, Sun and Toh (2012), Ma, Goldfarb and Chen (2011), Nesterov (2007*b*), Recht, Ré, Wright and Niu (2011*a*), Osher, Mao, Dong and Yin (2010), Qin and Goldfarb (2012), Recht and Ré (2011), Scheinberg, Ma

[5] Recall that in many applications A is a submatrix of the DFT/Hadamard transform, or represents a discrete convolution, and as such allows for fast matrix–vector multiplications.

Table 1.1. IPM versus FOM on ℓ_1 minimization problem (1.1b) with ℓ_2 fit and $\delta = 0.006$. A is the $m \times n$ matrix comprising real and imaginary parts of randomly selected $m/2$ rows in the unitary $n \times n$ DFT matrix; $x_{\rm res}$ is the solution found by the method. Platform: Intel Core i7 CPU M620 @ 2.66 GHz 2.66 GHz 8 GB RAM under Windows 7 64-bit.

	IPM†		FOM‡	
$m \times n$	$\|x_{\rm res}\|_1$	CPU (sec)	$\|x_{\rm res}\|_1$	CPU (sec)
256×512	0.971	1.2	0.977	2.0
512×1024	0.963	9.1	0.967	2.3
1024×2048	0.951	71.3	0.954	10.7
2048×4096	0.946	644.9	0.957	20.1
4096×8192	0.933	5768.5	0.937	53.0
8192×16384	out of memory		0.931	53.5

† Commercial IPM solver mosekopt.
‡ FOM (Mirror Prox: see Section 3) using FFT.

and Goldfarb (2010), Shi, Yin, Osher and Sajda (2010), Wen, Yin, Goldfarb and Zhang (2012), Wen, Yin, Zhang and Goldfarb (2012), Yang and Yuan (2013) and Yin, Osher, Goldfarb and Darbon (2008), and references therein. Although our sample is far from complete, we hope it is sufficiently representative. A more detailed overview of the extremely wide spectrum of FOMs proposed for ℓ_1/nuclear norm minimization seems to be an impossible task, going far beyond the scope of this paper. Here we intend to consider just two approaches in depth:

- *composite minimization* for lasso problems (1.1a),
- *saddle-point $O(1/t)$-converging first-order algorithms* for Dantzig selector problems (1.1b).

The rationale behind this selection is twofold. One consideration is that this selection is the optimal fit to the professional expertise and research experience of the authors, presented in Lemarechal, Nemirovski and Nesterov (1995), Nemirovski (2004), Nesterov (2005*a*, 2007*a*, 2007*b*, 2009), Juditsky and Nemirovski (2011*b*) and Juditsky, Kılınç Karzan and Nemirovski (2013*a*). Another consideration is that the selected methods are highly instructive examples of first-order algorithms with the *currently best known theoretical complexity guarantees* in our context (large-scale ℓ_1/nuclear norm minimization). Specifically, assuming that the norm $\|\cdot\|$ in (1.1a,b) is one of those mentioned in Section 1.1, the outlined algorithms exhibit $O(1/t^2)$ (lasso) or $O(1/t)$ (Dantzig selector) rate of convergence with 'natural' and

nearly dimension-independent hidden factors in $O(\cdot)$. For example, we shall see the following.

- Applied to the ℓ_1/nuclear norm lasso problem, composite minimization ensures that inaccuracy in terms of the objective after $t = 1, 2, \ldots$ iterations does not exceed $O(1)\ln(\dim x)L^2\|x^*\|^2/t^2$, where x^* is an optimal solution to the problem and $L = \max\{\|Ax\|_2 : \|x\| \le 1\}$.
- Applied to the ℓ_1/nuclear norm Dantzig selector problem, saddle-point methods allow us to find a solution of accuracy ϵ in at most

$$O(1)\sqrt{\ln(\dim x)}[\ln(\dim Ax)]^{1/2-1/p}(L \cdot \mathrm{Opt}/\epsilon)\ln(L \cdot \mathrm{Opt}/\epsilon)$$

  iterations, where Opt is the optimal value of the problem and

$$L = \max\{\|Ax\|_p : \|x\| \le 1\}.$$

To put these results into proper perspective, note that there are strong reasons to believe that these convergence rates are the best achievable by FOMs in the large-scale case. Specifically, it is known (Nemirovski 1992) that when solving the least-squares problem

$$\min_{x \in \mathbb{R}^n}\{\|Ax - b\|_2 : \|x\|_2 \le R\}$$

by a multiplication-oracle-based algorithm, *with zero optimal value* and spectral norm of A not exceeding L, the number of oracle calls needed to achieve accuracy ϵ in terms of the objective in the large-scale case $n \ge LR/\epsilon$ cannot be smaller than $O(1)LR/\epsilon$. It immediately follows that the outlined $O(1/t^2)$ and $O(1/t)$ rates of convergence are indeed unimprovable when solving large-scale lasso/Dantzig selector problems with $\|\cdot\|_2$ in the role of $\|\cdot\|$. Rigorously speaking, this observation does not imply similar consequences for ℓ_1/nuclear norm minimization. However, common sense, in full accordance with what is known on the subject so far, strongly suggests that the latter problems can hardly be much easier than 'ℓ_2 minimization'. Thus, the algorithms we intend to focus on are indeed most probably optimal in the large-scale case.

The rest of the paper is organized as follows. The constructions and results related to composite minimization in general, and its application to lasso problems in particular, are presented in Section 2. This section also contains a description of *proximal set-ups* underlying all algorithms considered in the paper. Section 3 is devoted to saddle-point algorithms and their applications to Dantzig selector problems. To make the paper self-contained, we present all necessary proofs, independent of whether their reference substitutions are available. All proofs are relegated to the Appendix.

Henceforth we shall operate with elements from direct products of Euclidean spaces $E = E_1 \times \cdots \times E_K$. An element x of E will usually be denoted

by $x = [x^1; \dots; x^K]$, where $x^j \in E_j$ is the jth block of x. The inner product of two vectors x, y from a Euclidean space E is denoted by $\langle x, y\rangle$. It is always clear from the context which Euclidean space is meant. The inner product on $E = E_1 \times \cdots \times E_K$ is always inherited from the inner products on the E_j, that is,

$$\langle [x^1; \dots; x^K], [y^1; \dots; y^K]\rangle = \sum_j \langle x^j, y^j\rangle.$$

For any norm $\|\cdot\|_E$ on a Euclidean space E, its conjugate norm is denoted by $\|\cdot\|_{E,*}$:

$$\|x\|_{E,*} = \max_y \big\{\langle x, y\rangle : \|y\|_E \le 1\big\}. \tag{1.2}$$

For a linear map $x \mapsto Ax : E \to F$, where E and F are Euclidean spaces, the conjugate map $y \mapsto A^*y : F \to E$ is given by the identity

$$\langle y, Ax\rangle = \langle A^*y, x\rangle, \quad x \in E,\ y \in F.$$

Given norms $\|\cdot\|_E$, $\|\cdot\|_F$ on Euclidean spaces E and F, the induced norm of a linear map $x \mapsto Ax : E \to F$ is defined by

$$\|A\|_{\|\cdot\|_E, \|\cdot\|_F} = \max_x \big\{\|Ax\|_F : \|x\|_E \le 1\big\}. \tag{1.3}$$

Note the identity

$$\|A\|_{\|\cdot\|_E, \|\cdot\|_F} = \|A^*\|_{\|\cdot\|_{F,*}, \|\cdot\|_{E,*}}.$$

2. Composite minimization and lasso problems

2.1. Preliminaries: proximal set-up

The algorithms we are about to consider are of proximal point type. It makes sense to start by describing the *proximal set-up* underlying the methods of this family. Let Z be a closed convex domain in Euclidean space E. A *proximal set-up* for Z is given by a norm $\|\cdot\|_E$ on E (not necessarily the Euclidean one), and a *distance-generating function* (DGF) $\omega(z) : Z \to R$, with the following properties.

- $\omega(\cdot)$ is a continuous convex function on Z.
- $\omega(\cdot)$ admits a selection of subgradients which is continuous on $Z^o = \{z \in Z : \partial\omega(z) \ne \emptyset\}$: there exists a continuous vector field $\omega'(z) : Z^o \to E$ such that $\omega'(z) \in \partial\omega(z)$ for all $z \in Z^o$.
- $\omega(\cdot)$ is strongly convex with modulus 1, with respect to $\|\cdot\|_E$, that is,

$$\langle \omega'(z) - \omega'(z'), z - z'\rangle \ge \|z - z'\|_E^2, \quad \text{for all } z, z' \in Z^o.$$

The proximal set-up gives rise to several concepts heavily exploited in the associated proximal-point algorithms, as follows.

- The *ω-centre* $z_\omega = \arg\min_{z\in Z} \omega(z) \in Z^o$.
- The *Bregman distance* $V_z(w)$, defined for $z \in Z^o$, $w \in Z$ by the relation

$$V_z(w) = \omega(w) - \omega(z) - \langle \omega'(z), w - z\rangle \geq \frac{1}{2}\|w - z\|_E^2, \tag{2.1}$$

  where the final inequality is due to the strong convexity of $\omega(\cdot)$.
- The *prox-mapping* $\mathrm{Prox}_z(\xi) : E \to Z^o$, parametrized by the *prox-centre* $z \in Z^o$ and defined by

$$\mathrm{Prox}_z(\xi) = \arg\min_{w\in Z}\big[\langle \xi, w\rangle + V_z(w)\big] = \arg\min_{w\in Z}\big[\langle \xi - \omega'(z), w\rangle + \omega(w)\big].$$

  Perhaps the most important property of this mapping is the inequality

$$\langle \xi, z_+ - w\rangle \leq V_z(w) - V_{z_+}(w) - V_z(z_+), \tag{2.2}$$

  for all $w \in Z$, where $z \in Z^o$, $\xi \in E$ and $z_+ = \mathrm{Prox}_z(\xi)$.
- The *ω-radius*. Let Z' be a *bounded* subset of Z containing z_ω. The ω-radius of Z' is the quantity

$$\Omega[Z'] = \sqrt{2\Big[\max_{z\in Z'} \omega(z) - \min_{z\in Z} \omega(z)\Big]}. \tag{2.3}$$

  The name stems from the immediate observation that

$$\|z - z_\omega\|_E \leq \sqrt{2V_{z_\omega}(z)} \leq \Omega[Z'], \tag{2.4}$$

  for all $z \in Z'$. Indeed, we clearly have $\mathrm{Prox}_{z_\omega}(\omega'(z_\omega)) = z_\omega$, whence by (2.2), $\langle \omega'(z_\omega), z - z_\omega\rangle \geq 0$ for all $z \in Z$, and thus

$$\frac{1}{2}\|z - z_\omega\|_E^2 \leq V_{z_\omega}(z) \leq \omega(z) - \omega(z_\omega),$$

  for all $z \in Z$.

2.1.1. Some useful proximal set-ups

Our selection of proximal set-ups is primarily motivated by the desire to obtain the currently best known complexity estimates in the applications we are interested in (for more details on this subject, see Section A.6 in the Appendix). Note that this goal can be achieved not only with the set-ups we intend to present but with other set-ups too. For the sake of brevity we do not discuss these alternatives, referring the interested reader to Juditsky and Nemirovski (2011*b*) (where one should ignore a completely incorrect version of the ℓ_1/ℓ_2 set-up). The structure of our presentation is as follows. We consider a Euclidean space E and a norm $\|\cdot\|_E$ on E. For every such normed space we suggest a 'good' DGF for two 'canonical' sets Z, namely $Z_E(1) = \{z \in E : \|z\|_E \leq 1\}$, and $Z = E$ (we will see that these are exactly the cases we are primarily interested in).

It turns out that all our needs in this paper can be covered just by three proximal set-ups.

The ℓ_1/ℓ_2 norm. Let $E = \mathbb{R}^{k_1\times\cdots\times k_n}$, so that $z \in E$ is a block vector: $z = [z^1; \ldots; z^n]$ with blocks $z^j \in \mathbb{R}^{k_j}$. Let us equip E with the ℓ_1/ℓ_2 norm $\|z = [z^1; \ldots; z^n]\|_E = \sum_{j=1}^n \|z^j\|_2$. Note that this situation covers, as extreme cases,

- the plain ℓ_2 norm on $\mathbb{R}^k$, which occurs when $n = 1$,
- the plain ℓ_1 norm on $\mathbb{R}^n$, which occurs when $k_1 = \cdots = k_n = 1$.

Theorem 2.1. Given E and $\|\cdot\|_E$ as specified, let

$$Z = Z_E(1) := \{z \in E : \|z\|_E \le 1\}.$$

Then the function $\omega : E \to \mathbb{R}$ defined by

$$\omega(z) = \frac{1}{p\gamma}\sum_{j=1}^n \|z^j\|_2^p, \tag{2.5}$$

$$\text{where } p = \begin{cases} 2, & n \le 2, \\ 1 + \frac{1}{\ln n}, & n \ge 3, \end{cases} \quad \text{and } \gamma = \begin{cases} 1, & n = 1, \\ 1/2, & n = 2, \\ \frac{1}{e\ln(n)}, & n > 2, \end{cases}$$

is a DGF for Z compatible with $\|\cdot\|_E$, and

$$\Omega[Z] \le \sqrt{2e\ln(n+1)}. \tag{2.6}$$

Corollary 2.2. Let E and $\|\cdot\|_E$ be as in Theorem 2.1. Then the function $\widehat{\omega} : E \to \mathbb{R}$, given by

$$\widehat{\omega}(z) = \frac{n^{(p-1)(2-p)/p}}{2\gamma}\left[\sum_{j=1}^n \|z^j\|_2^p\right]^{\frac{2}{p}}, \tag{2.7}$$

with p, γ given by (2.5), is a DGF for $Z = E$ compatible with $\|\cdot\|_E$, and

$$\Omega[\{z \in E : \|z\|_E \le R\}] \le O(1)\sqrt{\ln(n+1)}R, \quad \text{for all } R > 0. \tag{2.8}$$

The Euclidean set-up. In the case of $n = 1$ the above construction yields $\|\cdot\|_E = \|\cdot\|_2$, $\omega(z) = \frac{1}{2}z^Tz$ for both $Z = Z_E(1)$ and $Z = E$. In fact, for every Euclidean space E, the Euclidean norm $\|z\|_E = \|z\|_2 := \sqrt{\langle z, z\rangle}$ and DGF $\omega(z) = \frac{1}{2}\langle z, z\rangle$ yield a proximal set-up for E. Restricting $\omega(\cdot)$ to a closed convex domain $Z \subset E$, we get a proximal Euclidean set-up for Z.

The nuclear norm. Let E be the Euclidean space $\mathbb{R}^{\mu\times\nu}$ of $\mu \times \nu$ matrices equipped with the Frobenius inner product. We assume without loss of generality that $\mu \le \nu$. Let $\sigma(z) = [\sigma_1(z); \ldots; \sigma_\mu(z)]$ be the vector of singular

values of $z \in E$, taken in non-ascending order. Let $\|\cdot\|_E$ be the nuclear norm $\|\cdot\|_{\rm nuc}(z) = \|\sigma(z)\|_1$ on E.

Theorem 2.3. With E and $\|\cdot\|_E$ as specified, let

$$Z = Z_E(1) = \{z \in E : \|z\|_E \le 1\}.$$

Then the function $\omega : Z \to \mathbb{R}$ defined by

$$\omega(z) = \frac{4\sqrt{\mathrm{e}}\ln(2\mu)}{2^q(1+q)} \sum_{i=1}^{\mu} \sigma_i^{1+q}(z), \quad \text{where } q = \frac{1}{2\ln(2\mu)}, \tag{2.9}$$

is a DGF for Z compatible with $\|\cdot\|_E$, and

$$\Omega[Z] \le 2\sqrt{2\sqrt{\mathrm{e}}\ln(2\mu)} \le 4\sqrt{\ln(2\mu)}. \tag{2.10}$$

Corollary 2.4. Let E, μ, ν, $\|\cdot\|_E$ be as in Theorem 2.3. Then the function $\hat{\omega} : E \to \mathbb{R}$, given by

$$\widehat{\omega}(z) = 2\mathrm{e}\ln(2\mu)\left[\sum_{j=1}^{\mu} \sigma_j^{1+q}(z)\right]^{\frac{2}{1+q}}, \quad \text{where } q = \frac{1}{2\ln(2\mu)}, \tag{2.11}$$

is a DGF for $Z = E$ compatible with $\|\cdot\|_E$, and

$$\Omega[\{z \in E : \|z\|_E \le R\}] \le O(1)\sqrt{\ln(2\mu)}R, \quad \text{for all } R > 0. \tag{2.12}$$

Computational issues. For all the following algorithms, an iteration requires $O(1)$ computations of the prox-mappings. Therefore, except for the complexity bounds of the associated algorithms (*i.e.*, bounds expressing the decay of the error of approximate solutions as a function of the number of iterations), a particular proximal set-up affects the complexity of an iteration. Let us now look at the computational cost of a single computation of this type for the outlined set-ups and our 'canonical' Z. Observe that computing the value of prox-mapping at a given input essentially reduces to finding, for a given $\xi \in E$, the point z_+ defined by

$$z_+ = \arg\min_{z \in Z}\big\{\omega(z) + \langle \xi, z\rangle\big\}. \tag{2.13}$$

(A) When $\|[z^1; \dots; z^n]\|_E = \sum_j \|z^j\|_2$ is the ℓ_1/ℓ_2 norm on $E = \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_n}$ and the DGF is either (2.5), or (2.7), then solving (2.13) is easy. Indeed, it is immediately seen that the blocks z_+^j in z_+ are multiples, with non-positive coefficients, of the blocks ξ^j in ξ. This reduces finding z_+ to solving either a problem of the form

$$\min_{t \in \mathbb{R}^n}\left\{\frac{1}{p}\sum_{j=1}^{n} t_j^p - \alpha_j t_j : t \ge 0, \sum_{j=1}^{n} t_j \le 1\right\} \tag{2.14}$$

($Z = Z_E(1)$ and DGF (2.5)), or a problem of the form

$$\min_{t\in\mathbb{R}^n}\left\{\frac{1}{2}\left(\sum_{j=1}^n t_j^p\right)^{2/p} - \sum_{j=1}^n \alpha_j t_j : t \geq 0\right\} \tag{2.15}$$

($Z = E$ and DGF (2.7)). In both cases, $p \in (1, 2]$ and $\alpha_j \geq 0$. Now, (2.14) is a simple convex problem with a separable objective function and one separable linear constraint. As such, it can be solved to machine accuracy in just $O(n)$ arithmetic operations.[6] Problem (2.15) is even easier. It admits the closed-form solution $t = \|\alpha\|_q^{(p-2)/(p-1)}\tau$, where

$$\tau = \left[\alpha_1^{1/(p-1)}; \ldots; \alpha_n^{1/(p-1)}\right] \quad \text{and} \quad q = \frac{p}{p-1}.$$

As a result, in the cases under consideration the complexity of computing the prox-mapping is linear in $\dim E$.

(B) When $\|\cdot\|_E$ is the nuclear norm on $E = \mathbb{R}^{\mu\times\nu}$ and the DGF $\omega(\cdot)$ is given by (2.9) or by (2.11), solving (2.13) essentially reduces to computing the singular value decomposition (SVD) $\xi = UDV^T$ of the matrix ξ. The main observation is that, with our DGFs, z_+ is of the form $z_+ = UGV^T$, where G is a diagonal matrix with non-positive diagonal entries $-t_j$, $1 \leq j \leq \mu$.[7] Computing the t_j clearly reduces to solving a problem of the form (2.14) when $Z = Z_E(1)$, or (2.15) when $Z = E$, with μ in the role of n and $\sigma_j(\xi)$ in the role of α_j. Thus, in the case of the nuclear norm, our set-up remains practical in the range of sizes μ, ν allowing for reasonably fast SVD.[8]

With the Euclidean proximal set-up, prox-mapping reduces to the metric projection onto Z. In situations (A) and (B), the computational cost of the Euclidean prox-mapping is the same as for the set-ups we have just considered.

[6] Indeed, for $\sum_j \alpha_j^{1/(p-1)}$, the optimal solution to (2.14) is given by $t_j = \alpha_j^{1/(p-1)}$. Otherwise it is $t_j = t_j(\lambda^*)$, where $t_j(\lambda) = \max[\alpha_j - \lambda, 0]^{1/(p-1)}$ and λ_* is the unique positive root of the equation $\sum_j t_j(\lambda) = 1$. This root can be rapidly approximated to high accuracy, *e.g.*, by bisection.

[7] Indeed, due to rotational invariance of $\omega(\cdot)$ ($\omega(PzQ^T) = \omega(z)$ for orthogonal P, Q), it suffices to verify that if ξ is diagonal then so is z_+. Assuming ξ to be diagonal, both the feasible set and the objective of (2.13) are invariant with respect to the transformations $z \mapsto \mathrm{Diag}\{\epsilon_1, \ldots, \epsilon_\mu\} z \mathrm{Diag}\{\epsilon_1, \ldots, \epsilon_\nu\}$ with $\epsilon_i = \pm 1$. Since problem (2.13) is convex with unique optimal solution z_+ (recall that $\omega(\cdot)$ is strongly convex), this solution is invariant with respect to the outlined transformations, implying that z_+ is indeed diagonal.

[8] With the existing hardware and software and $\mu \approx \nu$, μ, ν can be in the range of a few thousand.

Remark 2.5. It can be seen immediately that when a DGF $\omega(\cdot)$ for $Z \subset E$ is compatible with a given norm $\|\cdot\|_E$ on E, and $Z' \subset Z$ is a closed convex set intersecting the relative interior of Z, then the restriction of $\omega(\cdot)$ onto Z' is a DGF for the latter set, still compatible with $\|\cdot\|_E$. With this in mind, the particular DGFs we have presented give rise to DGFs compatible with $\|\cdot\|_E$, for *all* closed convex and non-empty sets $Z' \subset Z$, not only for our 'canonical' unit $\|\cdot\|_E$-ball and all of E. An immediate question here is: When does this conversion not increase the computational complexity of the prox-mapping too much? Here are some good examples of this type associated with our 'basic' proximal set-ups.

(1) Let $E = \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_n}$, $\|z\|_E = \sum_j \|z^j\|_2$. If $Z = Z_E(1)$ and the DGF for Z is (2.5), the computational complexity of prox-mapping remains basically intact when Z' is obtained from Z by imposing upper bounds $\|z^j\|_2 \le r_j$ on some of the blocks z^j of $z \in Z'$, and, further, imposing restrictions on signs of some of the entries in z. This also holds when $Z = E$ and the DGF for Z is (2.7).

(2) Let $E = \mathbb{R}^{\mu\times\nu}$, $\|\cdot\|_E = \|\cdot\|_{\rm nuc}$, and let $\widehat{E}$ be the subspace of E comprising all block-diagonal matrices $z \in E$ with a prescribed number n of diagonal blocks $z^j \in \mathbb{R}^{\mu_j\times\nu_j}$, $1 \le j \le n$, with some prescribed μ_j, ν_j. Restricting the DGF (2.11) to $\widehat{E}$ and specifying $\|\cdot\|_{\widehat{E}}$ to be the nuclear norm on $\widehat{E}$, we obtain a proximal set-up for $Z' = \widehat{E}$, and do not increase the complexity of the prox-mapping,[9] since both the SVD of $z \in \widehat{E}$ and computing this SVD are decomposable along the blocks z^j. We can further restrict the blocks z^j of $z \in \widehat{E}$ by (a) imposing upper bounds on the spectral norms of some of the blocks, or (b) requiring some of the blocks z^j to be symmetric (and perhaps also positive semidefinite).[10] All of this can be done without greatly complicating the prox-mapping. We can proceed in the same fashion with the DGF (2.9), getting the proximal set-ups for the nuclear norm ball in the space of block-diagonal matrices and for the sets obtained from this ball by operations (a) and (b). Again, this does not greatly complicate the prox-mapping. In particular, we can thereby obtain reasonably good proximal set-ups for the space of symmetric matrices equipped with the trace norm, and for the positive semidefinite (PSD) cone in this space, and for spectrahedra, *i.e.*, intersections of the PSD-cone with trace norm balls centred at the origin.

Remark 2.5 is fully applicable to Euclidean set-ups as well.

[9] In fact it decreases, and quite significantly so when $\mu_j \ll \mu$ or $\nu_j \ll \nu$.
[10] Of course, (b) relates only to square blocks z^j.

Remark 2.6. A discussion of proximal set-ups cannot be complete without mentioning the *entropy* DGF:

$$\eta(z) = \sum_{i=1}^{n} z_i \ln z_i, \quad z_i \geq 0,\ i = 1, \ldots, n.$$

Its absence from our discussion is explained by the unconstrained nature of the main problems (1.1a,b). Indeed, the natural feasible set for this function is the simplex

$$\Delta_n = \left\{ z \in R^n_+ : \sum_{i=1}^{n} z_i = 1 \right\}.$$

On this set, $\eta(\cdot)$ is a DGF for the $\|\cdot\|_1$ norm with $\Omega[\Delta_n] = \sqrt{2 \ln n}$. All the results of this paper can be easily adapted to the entropy DGF provided that we have additional simplex constraints in problems (1.1a,b).

2.2. Composite minimization via gradient methods

2.2.1. Problem formulation

The general problem of composite minimization is as follows. We want to solve the convex program

$$\min_{x \in Z} \big[\phi(x) \stackrel{\text{def}}{=} f(x) + \Psi(x) \big], \tag{2.16}$$

where Z is a closed convex set in a Euclidean space E equipped with a norm $\|\cdot\|_E$ (not necessarily the Euclidean one), $\Psi(z) : Z \to \mathbb{R} \cup \{+\infty\}$ is a lower-semicontinuous convex function which is finite on the relative interior of Z, and $f : Z \to \mathbb{R}$ is a convex function with Lipschitz-continuous gradient, that is,

$$\|\nabla f(x) - \nabla f(y)\|_{E,*} \leq L_f \|x - y\|_E, \quad x, y \in Z. \tag{2.17}$$

Recall that $\|\cdot\|_{E,*}$ is the norm conjugate to $\|\cdot\|_E$ (see (1.2)).

Condition (2.17) ensures that

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L_f}{2} \|x - y\|_E^2, \quad \text{for all } x, y \in Z,$$

whence by (2.1)

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq L_f V_x(y), \quad \text{for all } x \in Z^o,\ y \in Z, \tag{2.18a}$$

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq L_f V_y(x), \quad \text{for all } y \in Z^o,\ x \in Z. \tag{2.18b}$$

To avoid unnecessary complications, we assume the constant $L_f > 0$ is known. At the end of Section 2 (see Algorithm 2.15) we discuss an efficient strategy for using its updated approximations.

From now on, we assume that Z is equipped with a proximal set-up comprising our norm $\|\cdot\|_E$ on E and a DGF $\omega(z)$ for Z compatible with this norm. We associate with this function the *composite* prox-mapping

$$\mathrm{Prox}_{\Psi,z}(\xi) = \arg\min_{w\in Z}[\langle \xi, w\rangle + L_f V_z(w) + \Psi(w)],$$

where $z \in Z^o = \{z \in Z : \partial\omega(z) \neq \emptyset\}$ and $\xi \in E$. Under our assumptions this mapping is clearly well defined, since $Z \neq \emptyset$ is closed and convex, $\omega(\cdot)$ is continuous and strongly convex, with modulus 1, with respect to $\|\cdot\|_E$, on Z, while Ψ is convex lower-semicontinuous and finite on $\operatorname{int} Z$. One can easily check (see Lemma A.1) that it takes its values in Z^o. To be useful in practical methods, the composite prox-mapping should be easy to compute (either by an explicit formula, or by a cheap computational procedure); we assume this from now on. Finally, we assume that (2.16) is solvable.

2.2.2. Gradient algorithms for composite minimization

In this subsection we present several methods for minimizing composite functions. For the Euclidean DGF these methods were suggested in Nesterov (2007*b*). The fast gradient method with Bregman distances was proposed in Nesterov (2007*b*). The most general form of these methods with stochastic oracle was justified by Devolder (2011). In the Appendix we give simplified versions of the corresponding proofs.

Preliminaries. Analysis of the subsequent composite minimization methods is based on the following simple observation.

Lemma 2.7. Let $f : Z \to \mathbb{R} \cup \{+\infty\}$ be a convex function such that the difference $f - L\omega$ for some $L > 0$ is convex lower-semicontinuous on Z and finite on the relative interior of Z. Let $\bar{w} = \arg\min_{w\in Z} f(w)$. Then $\bar{w} \in Z^o$, and for all $w \in Z$ we have

$$f(w) \geq f(\bar{w}) + LV_{\bar{w}}(w). \tag{2.19}$$

The primal gradient method. Inequality (2.19) justifies the following algorithm.

Algorithm 2.8 (primal gradient method).

Choose $x_0 \in Z^o$

For $t \geq 0$ **do**

$x_{t+1} = \mathrm{Prox}_{\Psi,x_t}(\nabla f(x_t))$ ◇

Let x^* denote an optimal solution to problem (2.16).

Theorem 2.9. Let the sequence $\{x_t\}_{t\geq 0}$ be constructed by Algorithm 2.8. Then $\phi(x_{t+1}) \leq \phi(x_t)$, and for every $T \geq 1$ we have

$$\sum_{t=1}^{T}(\phi(x_t) - \phi(x^*)) \leq L_f V_{x_0}(x^*). \tag{2.20}$$

As a result, for every $t \geq 1$, setting $x^t = \frac{1}{t}\sum_{\tau=1}^{t} x_\tau$, we have

$$\max\big[\phi(x_t), \phi(x^t)\big] - \phi(x^*) \leq \frac{L_f V_{x_0}(x^*)}{t} \tag{2.21}$$

and $x_t, x^t \in Z$.

The dual gradient method. In contrast to the primal gradient method (Algorithm 2.8), which ensures a sequence of monotonically decreasing function values $\{\phi(x_t)\}_{t\geq 0}$, our next method is not monotone.

Algorithm 2.10 (dual gradient method).

Set $s_0 = 0 \in E$

For $t \geq 0$ **do**

$x_t = \mathrm{Prox}_{t\Psi, z_\omega}(s_t)$

$y_t = \mathrm{Prox}_{\Psi, x_t}(\nabla f(x_t))$

$s_{t+1} = s_t + \nabla f(x_t)$ ◇

Theorem 2.11. Let $\{x_t\}_{t\geq 0}$ be constructed by Algorithm 2.10. Then $x_t \in Z^o$, so that the auxiliary points y_t, $t \geq 0$, are well defined and belong to Z^o. For every $T \geq 1$ we have

$$\sum_{t=0}^{T-1}(\phi(y_t) - \phi(x^*)) \leq L_f V_{z_\omega}(x^*). \tag{2.22}$$

As a result, letting $\bar{y}^t$ denote the best of the points $y_0, \ldots, y_{t-1}$, in terms of the values of ϕ, and letting $y^t = \frac{1}{t}\sum_{s=0}^{t-1} y_s$ be the average of these points, we have for all $t \geq 1$

$$\max\big[\phi(\bar{y}^t), \phi(y^t)\big] - \phi(x^*) \leq \frac{L_f V_{z_\omega}(x^*)}{t} \tag{2.23}$$

and $\bar{y}^t, y^t \in Z$.

Remark 2.12. It can be easily verified that, when redefining the points $\{y_t\}_{t\geq 0}$ in Algorithm 2.10 according to the usual composite gradient mapping, specifically,

$$y_t = \arg\min_{w\in Z}\{f(x_t) + \langle \nabla f(x_t), w - x_t\rangle + \tfrac{1}{2}L_f\|w - x_t\|_E^2 + \Psi(w)\}, \tag{2.24}$$

one preserves efficiency estimates (2.22) and (2.23).

It is interesting that for justification of both Algorithms 2.8 and 2.10 we use only inequality (2.18), which is weaker than the condition (2.17).

Fast composite gradient minimization. The methods we have described so far solve the composite minimization problem (2.16) with the $O(1/t)$ efficiency estimate. We are about to describe a 'fast' method for the same problem, with the $O(1/t^2)$ efficiency estimate.

The algorithm is as follows.

Algorithm 2.13 (fast gradient method).

Initialization $\psi_0(w) = L_f V_{z_\omega}(w),\ y_0 = z_\omega$

For $t \geq 0$ **do**

(a) Compute $z_t = \arg\min_{w \in Z} \psi_t(w)$. Set $\tau_t = \dfrac{2(t+2)}{(t+1)(t+4)}$.

(b) Choose $x_{t+1} = \tau_t z_t + (1-\tau_t) y_t$.

(c) Compute $\hat{x}_{t+1} = \arg\min_{w \in Z} \left[\langle \nabla f(x_{t+1}), w \rangle + \Psi(w) + \dfrac{2L_f}{t+2} V_{z_t}(w) \right]$.

(d) Update $y_{t+1} = \tau_t \hat{x}_{t+1} + (1-\tau_t) y_t$,

$$\psi_{t+1}(w) = \psi_t(w) + \frac{t+2}{2} \big[f(x_{t+1}) + \langle \nabla f(x_{t+1}), w - x_{t+1} \rangle + \Psi(w) \big].$$

◇

Note that $\psi_t(w) = \langle \xi_t, w \rangle + \gamma_t \Psi(w) + L_f V_{z_\omega}(w)$ with $\gamma_t > 0$, that is, operations (a) and (c) reduce to computing the values of composite prox-mappings with positive multiples of Ψ in the role of Ψ.

Theorem 2.14. Let the sequence $\{y_t\}_{t \geq 0}$ be generated by Algorithm 2.13. Then, for all $t \geq 1$ we have $y_t \in Z$ and

$$\phi(y_t) - \phi(x^*) \leq \frac{4 L_f V_{z_\omega}(x^*)}{t(t+3)}. \tag{2.25}$$

2.3. Composite minimization and lasso-type problems

The algorithms presented in Section 2.2.2 are well suited to solving lasso-type problems (1.1a) associated with the norms $\|\cdot\|$ of our primary interest. Indeed, (1.1a) is simply problem (2.16) with

$$\Psi(\cdot) \equiv \lambda \|\cdot\|, \quad f(x) \equiv \|Ax - b\|_2^2, \tag{2.26}$$

and $Z = E$. It is assumed from now on that the starting point x_0 in the primal gradient method is chosen as the ω-centre z_ω of $Z = E$, which makes

the efficiency estimates (2.21) and (2.23) of the primal and the dual gradient methods identical to each other.

Lasso with the ℓ_1/ℓ_2 norm. The first case we are interested in is that of $E = \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_n}$ and the ℓ_1/ℓ_2 norm

$$\|[z^1; \ldots; z^n]\| = \sum_{j=1}^{n} \|z^j\|_2.$$

Note that this covers the cases of plain ℓ_1 and plain ℓ_2 norms (see what happens when $k_j = 1$ for all j, respectively, when $n = 1$).

(A) When using $\|\cdot\|_E = \|\cdot\|$ and the DGF $\omega(\cdot)$ given by (2.7), the three gradient methods from Section 2.2.2 ensure the efficiency estimates[11]

$$[\lambda\|x^t\| + \|Ax^t - b\|_2^2] - \min_{x\in Z}[\lambda\|x\| + \|Ax - b\|_2^2] \le O(1)\frac{\ln(n+1)\|x^*\|^2\|A\|^2_{\|\cdot\|,\|\cdot\|_2}}{t^\kappa}, \tag{2.27}$$

$$\text{where } \kappa = \begin{cases} 1, & \text{primal and dual gradient methods,} \\ 2, & \text{fast gradient method,} \end{cases}$$

$$\text{and } \|A\|_{\|\cdot\|,\|\cdot\|_2} = \max_{x:\|x\|\le 1} \|Ax\|_2,$$

where x^t is an approximate solution built after the first t steps, and x^* is an optimal solution to problem (2.16).

(B) An alternative is to use the Euclidean proximal set-up,

$$\|x\|_E = \|x\|_2 := \sqrt{\langle x, x\rangle}, \quad \omega(x) = \frac{1}{2}\langle x, x\rangle,$$

which results in

$$[\lambda\|x^t\| + \|Ax^t - b\|_2^2] - \min_{x\in Z}[\lambda\|x\| + \|Ax - b\|_2^2] \le O(1)\frac{\|x^*\|_2^2\|A\|^2_{\|\cdot\|_2,\|\cdot\|_2}}{t^\kappa},$$
$$\text{where } \|A\|_{\|\cdot\|_2,\|\cdot\|_2} = \max_{x:\|x\|_2\le 1} \|Ax\|_2, \tag{2.28}$$

with the same κ as in (2.27).

Except for the efficiency estimates, the question of primary importance is the computational effort per iteration. From the description of the algorithms, in all three of them this effort is dominated by the need to compute $O(1)$ values of $\nabla f(\cdot)$, which requires $O(1)$ matrix–vector multiplications $x \mapsto Ax \mapsto A^*(Ax - b)$, and computation of the values of $O(1)$ composite prox-mappings. For both set-ups we have considered so far, the latter task is

[11] See (2.21), (2.23) and (2.25) and take into account that we have $L_f = 2\|A\|^2_{\|\cdot\|,\|\cdot\|_2}$.

easy. Let us show that it requires $O(\dim E)$ arithmetic operations. Indeed, in our present situation, and for both proximal set-ups in question, computing composite prox-mapping reduces to solving the following analogue of problem (2.13):

$$z_+ = \arg\min_{z\in Z}\left\{\omega(z) + \beta\sum_j \|z^j\|_2 + \sum_j \langle z^j, \xi^j\rangle\right\}, \quad \text{for } \beta \geq 0. \tag{2.29}$$

Now the two DGFs we are considering are of the form $\gamma(\sum_j \|z^j\|_2^p)^{2/p}$, where $\gamma > 0$ and $p \in (1, 2]$. With this in mind, it is clear that, at the minimum, z^j should be of the form $z^j = -t_j\xi^j/\|\xi^j\|_2$, $t_j \geq 0$. Thus (2.29) is reduced to the following problem:

$$\min_t\left\{\gamma\left(\sum_j t_j^p\right)^{2/p} + \sum_j \big[\beta - \|\xi^j\|_2\big]t_j : t \geq 0\right\}. \tag{2.30}$$

It can be seen immediately that, at the minimum, $t_j = 0$ for every j such that $\beta \geq \|\xi^j\|_2$. Eliminating these j, we end up with a problem of type (2.15) which, as we have seen in Section 2.1.1, admits an easily computable closed-form solution.

Lasso with the TV norm. The total variation on the space of $\mu \times \nu$ images x (normalized to have $x_{1,1} = 0$) can be viewed as follows. The space of discrete gradients of $\mu \times \nu$ images is a subspace E of the space

$$\mathbb{R}^N = \mathbb{R}^{(\mu-1)\times\nu} \times \mathbb{R}^{\mu\times(\nu-1)}$$

of first-order finite differences of an image $[x_{ij}]$ taken along the index $i \leq \mu$ ($\mathbb{R}^{(\mu-1)\times\nu}$) and along the index $j \leq \nu$ ($\mathbb{R}^{\mu\times(\nu-1)}$). The space E is cut from $\mathbb{R}^N$ by a system of homogeneous linear equations expressing the potentiality of the discrete gradient field of an image. Thus, the TV norm on E is a restriction of the standard ℓ_1 norm on $\mathbb{R}^N$ onto $E \subset \mathbb{R}^N$. This allows us to equip $Z = E$ with a DGF $\omega(\cdot)$ (namely, the restriction to E of the DGF (2.7) associated with the $\|\cdot\|_1$ on $\mathbb{R}^N$), which leads to a reasonably good iteration complexity bound. As an alternative, we could use the Euclidean proximal set-up for E.

The difficulty, however, is that we now need to take into account the linear constraints cutting E from $\mathbb{R}^N$. Hence both the plain and the composite prox-mappings associated with the outlined proximal set-ups are much more computationally demanding than in the case of lasso with the ℓ_1/ℓ_2 norm. As a result, the 'straightforward' TVversions of the methods we have presented are hardly advisable.

Lasso with the nuclear norm. In the second important case, we have $E = \mathbb{R}^{\mu\times\nu}$, $\|\cdot\| = \|\cdot\|_{\rm nuc}$, and $Z = E$.

(A) Using our three basic methods and the proximal set-up with $\|\cdot\|_E = \|\cdot\|$ and with $\omega(\cdot)$ given by (2.11), we come to the efficiency estimates

$$[\lambda\|x^t\|_{\rm nuc} + \|Ax^t - b\|_2^2] - \min_{x\in Z}[\lambda\|x\|_{\rm nuc} + \|Ax - b\|_2^2] \le O(1)\frac{\ln(\mu+1)\|x^*\|_{\rm nuc}^2\|A\|_{\|\cdot\|_{\rm nuc},\|\cdot\|_2}^2}{t^\kappa}, \tag{2.31}$$

$$\text{where } \kappa = \begin{cases} 1, & \text{primal and dual gradient methods,} \\ 2, & \text{fast gradient method,} \end{cases}$$

$$\text{and } \|A\|_{\|\cdot\|_{\rm nuc},\|\cdot\|_2} = \max_{x:\|x\|_{\rm nuc}\le 1}\|Ax\|_2,$$

where x^t is the approximate solution obtained after the first t iterations, and x^* is the optimal solution.

(B) An alternative is to use the Euclidean set-up,

$$\|x\|_E = \|x\|_{\rm Fro} := \sqrt{{\rm Tr}(x^Tx)}, \quad \omega(x) = \frac{1}{2}\langle x, x\rangle = \frac{1}{2}{\rm Tr}(x^Tx).$$

The corresponding efficiency estimate is

$$[\lambda\|x^t\|_{\rm nuc} + \|Ax^t - b\|_2^2] - \min_{x\in Z}[\lambda\|x\|_{\rm nuc} + \|Ax - b\|_2^2] \le O(1)\frac{\|x^*\|_{\rm Fro}^2\|A\|_{\|\cdot\|_{\rm Fro},\|\cdot\|_2}^2}{t^\kappa}, \tag{2.32}$$

$$\text{where } \|A\|_{\|\cdot\|_{\rm Fro},\|\cdot\|_2} = \max_{x:\|x\|_{\rm Fro}\le 1}\|Ax\|_2.$$

The computational effort per iteration is now dominated by the need to carry out $O(1)$ matrix–vector multiplications $x \mapsto Ax \mapsto A^*(Ax - b)$ along with $O(1)$ SVDs of $\mu \times \nu$ matrices. Arguments virtually identical to those used in (B) at the end of Section 2.1.1 show that computing composite prox-mapping, modulo a single SVD, for both our proximal set-ups reduces to solving a problem of the form (2.30). Thus it is easy.

It should be pointed out that the efficiency estimates (2.27), (2.28) and (2.31) are nearly independent of the sizes of corresponding lasso problems.

Finally, observe that Remark 2.5 is fully applicable to composite prox-mappings associated with the DGFs and corresponding norms $\|\cdot\|_E$. Thus the methods we have developed can be applied to lasso problems on more general domains than $Z \equiv E$, with the efficiency estimates and complexity of iterations being preserved.

Updating the Lipschitz constant for the gradient. In this subsection we assume that the Lipschitz constant L_f is known. However, it is easy to introduce into the methods a simple updating procedure, which allows us to adjust an estimate to L_f in accordance with the local curvature of the

objective function. We present below the corresponding modification of the primal gradient method (Algorithm 2.8).

Algorithm 2.15 (modified primal gradient method).

Choose $x_0 \in Z^o$, $L_0 \leq L_f$.

For $t \geq 0$ **do**

(a) Choose the smallest $i_t \geq 0$ such that, for $L = 2^{i_t} L_t$ and

$$w_t = \arg\min_{w \in Z} [\langle \nabla f(x_t), w \rangle + L V_{x_t}(w) + \Psi(w)],$$

we have

$$\phi(w_t) \leq f(x_t) + \langle \nabla f(x_t), w_t - x_t \rangle + L V_{x_t}(w_t) + \Psi(w_t).$$

(b) Set $x_{t+1} = w_t$, $L_{t+1} = 0.5 \cdot 2^{i_t} L_t$. ◇

It is easy to show that such a strategy ensures the rate of convergence (*cf.* (2.21))

$$\phi(x_t) - \phi(x^*) \leq \frac{2 L_f V_{x_0}(x^*)}{t}, \quad t \geq 1.$$

At the same time, it allows flexible dynamics for estimates L_t. It can be shown that the average number of 'trials' $i_t + 1$ per iteration in such a *backtracking strategy with recourse* is at most two (Nesterov and Polyak 2006). The dual gradient method (Algorithm 2.10) and the fast gradient method (Algorithm 2.13) can be modified in a similar way.

3. Saddle-point algorithms for Dantzig selector problems

The algorithmic scheme we intend to use for Dantzig-selector-type problems (1.1b) originates from Juditsky *et al.* (2013*a*). To allow for non-trivial extensions, we now remove the restriction that the 'model discrepancy' $Ax - b$ be measured in the $\|\cdot\|_p$ norm with $p = 2$ or $p = \infty$, and focus on a slightly more general problem,

$$\mathrm{Opt} = \min_{x \in E} \big\{ \|x\| : \pi(Ax - b) \leq \delta \big\}, \tag{3.1}$$

where $E = E_u$ is a Euclidean space,[12] $\|\cdot\|$ is a given norm on E, $x \mapsto Ax - b$ is a given affine mapping from E_x into another Euclidean space E_v, and $\pi(\cdot)$ is a given norm on E_v. Aside from the structural discrepancy between (3.1) and the associated lasso-type problem

$$\min_x \big\{ \lambda \|x\| + \pi^2(Ax - b) \big\} \tag{3.2}$$

[12] The reason for the notation $E = E_u$, instead of the seemingly more natural $E = E_x$, will become clear in a while.

(this discrepancy disappears when δ and λ are treated as parameters rather than given constants), the major difference (from the optimization viewpoint) between what we intend to do now and what was done in Sections 2.2 and 2.3 stems from the fact that we no longer impose any smoothness restrictions on the norm $\pi(\cdot)$, which makes it impossible to process (3.2) by the composite minimization technique from Section 2. On the other hand, allowing for non-smooth $\pi(\cdot)$ is instrumental when solving Dantzig-selector-type problems with uniform fit.

From now on we make the following assumption.

Standing Assumption. Optimization problem (3.1) is feasible (and therefore solvable) and Opt > 0 (equivalently $\pi(b) > \delta$).

3.1. Strategy

Our strategy when solving (3.1) is motivated by the desire to minimize, to the best of our abilities, the theoretical complexity of the resulting algorithm, and is as follows. Substituting $x = \rho^{-1}u$ with $\rho > 0$ and $u \in E_u$, $\|u\| \le 1$ and taking into account that Opt > 0, it is immediately seen that (3.1) can be equivalently rewritten as

$$\frac{1}{\text{Opt}} = \rho_* := \max\left\{\rho \ge 0 : \Phi(\rho) := \min_{u\in U} \overbrace{\max_{v\in V}\left[\langle v, Au - \rho b\rangle - \rho\delta\right]}^{=\pi(Au-\rho b)-\rho\delta} \le 0\right\},$$
$$\text{for} \quad U = \{u \in E_u : \|u\| \le 1\}, \quad V = \{v \in E_v : \pi_*(v) \le 1\}, \tag{3.3}$$

where $\pi_*(\cdot)$ is the norm on E_v conjugate to $\pi(\cdot)$.

Problem (3.3) (a specific *generalized bilinear saddle-point problem* in the terminology of Juditsky *et al.* 2013*a*) is a univariate problem. As is easily seen, under our Standing Assumption $\Phi(\rho)$ is a Lipschitz-continuous (with constant $2\pi(b)$) *convex* function of $\rho \ge 0$ with exactly two roots $\rho = 0$ and $\rho = \rho_* > 0$. Further, $\Phi(\rho) \to \infty$ as $\rho \to \infty$, and $\lim_{\rho\to\infty} \Phi'(\rho) = \pi(b) - \delta$.

Our goal is to find an ϵ-*solution* to (3.3), defined as a pair ρ_ϵ, u_ϵ satisfying

$$\rho_\epsilon \ge \rho_*, \quad u_\epsilon \in U \quad \text{and} \quad \pi(Au_\epsilon - \rho_\epsilon b) - \rho_\epsilon\delta \le \rho_\epsilon\epsilon, \tag{3.4}$$

where $\epsilon > 0$ is a given tolerance. As is seen immediately, such a solution gives rise to the solution

$$x_\epsilon = \rho_\epsilon^{-1}u_\epsilon$$

to the problem (3.1) such that

$$\|x_\epsilon\| \le \text{Opt} \;\&\; \pi(Ax_\epsilon - b) \le \delta + \epsilon. \tag{3.5}$$

In other words, x_ϵ is a 'super-optimal ϵ-feasible' solution to the problem of interest; from now on, we refer to a solution x_ϵ satisfying (3.5) as an ϵ-solution to (3.1).

Now, (3.3) is just a problem of finding the largest root of a univariate function $\Phi(\cdot)$; the only difficulty is that this function is given implicitly, as the saddle-point value of a parametric bilinear saddle-point problem. Our strategy for solving (3.3) originates from Lemarechal *et al.* (1995) and reduces to applying to (3.3) a Newton-type root-finding procedure, the (approximate) first-order information for the procedure being given by a first-order saddle-point algorithm applied to the auxiliary bilinear saddle-point problems

$$\Phi(\rho_s) = \min_{u\in U}\max_{v\in V}\big[\phi^{\rho_s}(u,v) := \langle v, Au - \rho_s b\rangle - \rho_s\delta\big], \tag{3.6}$$

the corresponding monotone operators being $F^s : Z := U \times V \to E := E_u \times E_v$, given by

$$F^s(u,v) = \tag{3.7}$$
$$[F^s_u(v) := \nabla_u\phi^{\rho_s}(u,v) = A^*v; F^s_v(u) := -\nabla_v\phi^{\rho_s}(u,v) = \rho^s b - Au].$$

Specifically, we work in *stages* $s = 1, 2, \ldots$. At stage s, we have at our disposal a current approximation $\rho_s \geq \rho_*$ to ρ_*, and apply to (3.6) a saddle-point algorithm $\mathcal{B}$, based only on the values of F^s, and thus only on a multiplication oracle computing matrix–vector products $(u,v) \mapsto (A^*v, Au)$. For the time being, the only requirement on the algorithm is that in $t = 1, 2, \ldots$ iterations, with $O(1)$ calls to the multiplication oracle per iteration, it produces approximate solutions $z^t = [u^t; v^t] \in Z$ to (3.6)[13] along with the values $F^s(z^t)$, such that the *duality gap* converges to zero as $t \to \infty$, where

$$\begin{aligned}\mathrm{DualityGap}^s(z^t) &:= \max_{v\in V}\phi^{\rho_s}(u^t,v) - \min_{u\in U}\phi^{\rho_s}(u,v^t) \qquad (3.8)\\ &= \underbrace{[\pi(F^s_v(u^t)) - \rho_s\delta]}_{\Phi^+_t} - \underbrace{[-\|F^s_u(v^t)\|_* - \rho_s\langle v^t, b\rangle - \rho_s\delta]}_{\Phi^-_t(\rho)\big|_{\rho=\rho_s}}\\ &= [\pi(Au^t - \rho_s b) - \rho_s\delta] - [-\|A^*v^t\|_* - \rho_s\langle v^t, b\rangle - \rho_s\delta],\end{aligned}$$

and $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$.

Observe the following.

(I) Given z^t and $F^s(z^t)$, it is easy to compute the quantity Φ^+_t and the affine function $\Phi^-_t(\cdot)$ (specifically, these computations reduce to single evaluation of $\pi(\cdot)$ and single evaluation of $\|\cdot\|_*$).

(II) The quantity $\Phi^+_t = \pi(Au^t - \rho_s b) - \rho_s\delta$ is an upper bound on $\Phi(\rho_s)$, while the affine function

$$\Phi^-_t(\rho) = [-\|A^*v^t\|_* - \rho\langle v^t, b\rangle - \rho\delta] = \min_{u\in U}\phi^{\rho}(u,v^t)$$

[13] Of course, z^t depends on s and not only on t. To simplify the notation we suppress the explicit dependence on s, which does no harm when speaking about a particular stage.

underestimates $\Phi(\rho)$ for all $\rho \geq 0$. As a result, the quantities

$$\Phi^{t,+} = \min_{1\leq\tau\leq t} \Phi_\tau^+, \quad \Phi^{t,-} = \max_{1\leq\tau\leq t} \Phi_\tau^-(\rho_s)$$

form upper bounds on $\Phi(\rho_s)$ which are non-increasing functions of t (resp. lower bounds non-decreasing with t). Further, because $\rho_s \geq \rho_*$, the affine function $\Phi_t^-(\cdot)$ underestimates $\Phi(\cdot)$, and $\Phi(\rho_*) = 0$, for every t such that $\Phi_t^-(\rho_s) > 0$, the slope of the affine function $\Phi_t^-(\cdot)$ is positive, and the only root r_t of this function belongs to $[\rho_*, \rho_s)$. Finally, we clearly have $\Phi^{t,+} - \Phi^{t,-} \leq \Phi_t^+ - \Phi_t^-(\rho_s) = \mathrm{DualityGap}^s(z^t)$, so the gap $\Phi^{t,+} - \Phi^{t,-}$ goes to zero as $t \to \infty$.

From (II) it follows that eventually one of the following two possibilities will be met.

(A) We arrive at $\Phi^{t,+} \leq \epsilon\rho_s$. Recalling that $\Phi^{t,+} = \pi(Au^\tau - \rho_s b) - \rho_s\delta$ with some $\tau \leq t$ and known $u^\tau \in U$, in the case in question we terminate the solution process and output $\rho_\epsilon = \rho_s$ and $u_\epsilon = u^\tau$. The resulting pair ρ_ϵ, u_ϵ clearly satisfies (3.4).

(B) We arrive at $\Phi^{t,-} > 0$ and $\Phi^{t,+}/\Phi^{t,-} \leq \theta$, where $\theta \in (1, 2)$ is a parameter of our construction. Since $\Phi^{t,-} > 0$, we have at our disposal a non-empty collection R_t of reals $r_\tau \in [\rho_*, \rho_s)$ coming from the iterations $\tau \leq t$ with $\Phi_\tau^-(\rho_s) > 0$. We define ρ_{s+1} as the minimal element of R_t, thus ensuring that $\rho_* \leq \rho_{s+1} < \rho_s$, terminate stage s and pass to stage $s + 1$.

Validity and complexity. We use the following result of Lemarechal *et al.* (1995).

Lemma 3.1. Let the root-finding routine described above, with parameter $\theta \in (1, 2)$, be initialized by some upper bound ρ_1 on ρ_*. Then the routine terminates in at most

$$S = \frac{1}{\ln(2/\theta)} \ln\left(1 + \frac{3\pi(b)\rho_1}{\epsilon\rho_*}\right) + 1 \tag{3.9}$$

stages. Further, suppose that the saddle-point method $\mathcal{B}$ used to solve problems (3.6) ensures that at every stage, for every $t = 1, 2, \ldots$, we have

$$\mathrm{DualityGap}^s(z^t) \leq \varepsilon_t \tag{3.10}$$

for some sequence ε_t (the same sequence for all stages) satisfying $\varepsilon_t \to 0$ as $t \to \infty$. Then the number of iterations at every stage does not exceed

$$t(\epsilon) = \min\left\{t : \varepsilon_t \leq \frac{\theta - 1}{\theta}\epsilon\rho_*\right\}. \tag{3.11}$$

Initialization. Our root finding should start with some upper bound ρ_1 on ρ_*. To get such a bound, note that every feasible solution x to (3.1) clearly satisfies $\pi(Ax) \geq \pi(b) - \delta > 0$, implying that

$$\rho_* = \frac{1}{\mathrm{Opt}} \leq \rho_1 := \frac{\|A\|_{\|\cdot\|,\pi(\cdot)}}{\pi(b) - \delta}, \quad \text{where } \|A\|_{\|\cdot\|,\pi(\cdot)} = \max_x \{\pi(Ax) : \|x\| \leq 1\}. \tag{3.12}$$

3.2. Workhorse: Mirror Prox algorithm

To implement the outlined strategy, we need an efficient first-order algorithm for solving bilinear saddle-point problems. We intend to use to this end the $O(1/t)$-converging *Mirror Prox* (MP) algorithm from Nemirovski (2004). This algorithm is applicable to a convex–concave saddle-point problem

$$\mathrm{SadVal} = \min_{u \in U} \max_{v \in V} \phi(u, v), \tag{3.13}$$

where $U \subset E_u$, $V \subset E_v$ are closed convex subsets of Euclidean spaces, and $\phi : Z := U \times V \to \mathbb{R}$ is a function convex in $u \in U$ and concave in $v \in V$ with Lipschitz-continuous gradient. We associate with (3.13) the vector field $F : Z := X \times Y \to E = E_u \times E_v$, given by

$$F(z = (u, v)) = [F_u(z) = \nabla_u \phi(u, v); F_v(z) = -\nabla_v \phi(u, v)]. \tag{3.14}$$

Recall that (3.13) gives rise to two convex optimization problems:

$$(P) \quad \mathrm{Opt}(P) = \min_{u \in U} \Big[\overline{\phi}(u) := \sup_{v \in V} \phi(u, v)\Big], \tag{3.15a}$$

$$(D) \quad \mathrm{Opt}(D) = \max_{v \in V} \Big[\underline{\phi}(v) := \inf_{u \in U} \phi(u, v)\Big], \tag{3.15b}$$

with $\mathrm{Opt}(P) \geq \mathrm{Opt}(D)$. Further, ϕ possesses a saddle point on $U \times V$ if and only if both (3.15a) and (3.15b) are solvable with equal optimal values, for which we have $\mathrm{Opt}(P) = \mathrm{Opt}(D) = \mathrm{SadVal}$. Moreover, saddle points of ϕ, when they exist, are exactly the pairs $[u_*; v_*]$, where u_*, v_* are optimal solutions to the respective problems (3.15a) and (3.15b). Finally, when U and V are bounded and convex–concave ϕ is continuous (as is the case with the problems (3.6)), ϕ definitely has a saddle point. Assuming that the latter is the case, a natural inaccuracy measure of a candidate solution $[u; v] \in Z$ to (3.13) is given by (*cf.* (3.8))

$$\mathrm{DualityGap}(u, v) = \overline{\phi}(u) - \underline{\phi}(v) = [\overline{\phi}(u) - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - \underline{\phi}(v)].$$

The Mirror Prox algorithm. The Mirror Prox (MP) algorithm is associated with a proximal set-up $(\|\cdot\|_E, \omega(\cdot))$ for the domain $Z = U \times V \subset E = E_u \times E_v$ (see Section 2.1), and is given by the following recurrence, where $\gamma_t > 0$ are the step sizes.

Algorithm 3.2 (MP algorithm).

Choose $z_1 = z_\omega = \arg\min_Z \omega(\cdot)$

For $0 \le \tau \le t-1$ **do**

$w_\tau = [u_\tau; v_\tau] := \mathrm{Prox}_{z_\tau}(\gamma_\tau F(z_\tau))$

$z_{\tau+1} = \mathrm{Prox}_{z_\tau}(\gamma_\tau F(w_\tau))$

$z^t = [u^t; v^t] := \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \sum_{\tau=1}^t \gamma_\tau z_\tau$ ◇

The efficiency estimate. The efficiency estimate of an MP is given by the following theorem.

Theorem 3.3. Let the Mirror Prox algorithm (Algorithm 3.2) be applied to a convex–concave saddle-point problem (3.13) associated with the Lipschitz-continuous vector field F, *i.e.*,

$$\|F(z) - F(z')\|_{E,*} \le L\|z - z'\|_E, \quad \text{for all } z, z' \in Z, \tag{3.16}$$

where $\|\cdot\|_{E,*}$ is the norm conjugate to the norm $\|\cdot\|_E$. Further, let the step size policy ensure that

$$\gamma_\tau \ge L^{-1} \ \&\ \delta_\tau := \gamma_\tau \langle F(w_\tau), w_\tau - z_{\tau+1}\rangle - V_{z_\tau}(z_{\tau+1}) \le 0 \tag{3.17}$$

for all τ; this is definitely so when $\gamma_\tau = L^{-1}$ for all τ. Finally, let $Z' = U' \times V'$, where the sets $U' \subset U$ and $V' \subset V$ are closed and bounded and such that $z_\omega \in Z'$. Then, for every t we have $z^t = [u^t; v^t] \in Z$ and

$$\varepsilon_{Z'}(z^t) := \max_{v \in V'} \phi(u^t, v) - \min_{u \in U'} \phi(u, v^t) \le \frac{\Omega^2[Z']}{2\sum_{\tau=1}^t \gamma_\tau} \le \frac{\Omega^2[Z']L}{2t}, \tag{3.18}$$

with $\Omega[\cdot]$ defined by (2.3).

Discussion. Let the premises of Theorem 3.3 hold.

(A) Assume that U and V are bounded, as is the case with problems (3.6). Then we can set $Z' = Z$, which results in $\varepsilon_{Z'} \equiv \mathrm{DualityGap}$ and provides us with the bound

$$\mathrm{DualityGap}(u^t, v^t) \le \frac{\Omega^2[Z]L}{2t}, \quad \text{for } t = 1, 2, \ldots, \tag{3.19}$$

with L given by (3.16).

(B) Assume that V is bounded, and ϕ admits a saddle point $[u_*; v_*]$. Setting $V' = V$, $U' = [u_\omega, u_*]$, where u_ω is the u-component of z_ω, we extract from (3.18) that

$$\varepsilon_{Z'}(z^t) = \overline{\phi}(u^t) - \min_{u \in U'} \phi(u, v^t) \ge \overline{\phi}(u^t) - \phi(u_*, v^t) \ge \overline{\phi}(u^t) - \phi(u_*, v_*),$$

where the final inequality follows from the fact that $\phi(u_*, v) \le \phi(u_*, v_*)$ by the definition of a saddle point. As explained above, the existence of a saddle point $[u_*; v_*]$ implies that $\phi(u_*, v_*) = \mathrm{Opt}(P) = \mathrm{Opt}(D)$, and we obtain

$$\overline{\phi}(u^t) - \mathrm{Opt}(P) \le \varepsilon_{Z_*}(z^t) \le \frac{\Omega^2[Z_*]L}{2t}, \quad \text{where } Z_* = [u_\omega, u_*] \times V. \quad (3.20)$$

3.3. Assembling partial proximal set-ups

Our strategy for solving the Dantzig selector problem (3.1) reduces this task to using the Mirror Prox algorithm (Algorithm 3.2) to solve a 'small series' of bilinear saddle-point problems (3.6) on a common domain $Z = U \times V$ with *bounded* U, V, specifically the unit balls of the norms $\|\cdot\|$ and $\pi_*(\cdot)$, respectively. To implement this strategy, we need a good proximal set-up for Z, and we intend to 'assemble' this set-up from the basic set-ups listed in Section 2.1.1. The assembling to be implemented originates from Nemirovski (2004) and is as follows. Consider a convex–concave saddle-point problem (3.13) with smooth convex–concave cost function ϕ and *bounded* U, V, and let U and V be given as direct products of other closed and bounded convex domains:

$$U = Z^1 \times \cdots \times Z^p, \quad V = Z^{p+1} \times \cdots \times Z^K,$$

where $Z^k \subset E_k$ are convex, closed and bounded subsets of Euclidean spaces. We clearly have

$$Z := U \times V = Z^1 \times \cdots \times Z^K \subset E = E_1 \times \cdots \times E_K.$$

Now, let every pair Z^k, E_k be equipped with proximal set-up $\|\cdot\|_{E_k}, \omega_k(z^k)$. Consider an assembling of those set-ups into a proximal set-up for Z, E. Specifically, given positive α_k, $1 \le k \le K$ (the parameters of the construction), we set

$$\|z = [z^1; \ldots; z^K]\|_E = \sqrt{\sum_{k=1}^K \alpha_k^{-1} \|z^k\|_{E_k}^2}, \quad (3.21a)$$

$$\omega(z = [z^1; \ldots; z^K]) = \sum_{k=1}^K \alpha_k^{-1} \omega_k(z^k). \quad (3.21b)$$

It is immediately seen that this is indeed a proximal set-up for Z, E, and that computation of the associate proximal mapping reduces to computing proximal mappings associated with Z^k and our 'partial' DGFs $\omega_k(\cdot)$, $1 \le k \le K$. It is natural to choose the 'assembling parameters' in such a way as to obtain the best efficiency estimate of the MP algorithm allowed by our scheme as applied to (3.13). To this end, note that in view of the

direct product structure of E, the vector field (3.14) associated with (3.13) can be written as $F(z) = [F_1(z); \dots; F_K(z)]$ with $F_k(z) \in E_k$. Since this field is Lipschitz-continuous, we can find 'partial Lipschitz constants' $L_{k\ell}$, $1 \le k, \ell \le K$ such that, for all $z = [z^1; \dots; z^K] \in Z$, $w = [w^1; \dots; w^K] \in Z$, we have

$$\|F_k(z) - F_k(w)\|_{E_k,*} \le \sum_{\ell=1}^{K} L_{k\ell} \|z^\ell - w^\ell\|_{E_\ell}, \quad \text{for all } k \le K,$$

where $\|\cdot\|_{E_k,*}$ is the norm on E_k conjugate to $\|\cdot\|_{E_k}$. Note that we can always enforce $L_{k\ell} = L_{\ell k}$, which we assume from now on.

Example. When ϕ is bilinear, $\phi(u, v) = \langle a, u\rangle + \langle b, v\rangle + \langle v, Au\rangle + c$, the mapping F is affine,

$$F([z^1; \dots; z^K]) = [a; -b] + \mathcal{A} \cdot [z^1; \dots; z^K],$$

with linear mapping $\mathcal{A}$ given by

$$\big[\mathcal{A} \cdot [z^1; \dots; z^K]\big]^k = \sum_{\ell=1}^{K} \mathcal{A}^{k\ell} z^\ell, \quad 1 \le k \le K,$$

where the $\mathcal{A}^{k\ell}$ are linear mappings from E_ℓ to E_k. It is evident that a characteristic property of $\mathcal{A}$ is that it is *skew-symmetric*: the mapping conjugate to $\mathcal{A}^{k\ell}$ is $-\mathcal{A}^{\ell k}$. In addition, $\mathcal{A}^{k\ell}$ should be zero when both Z^k and Z^ℓ are either factors in U or factors in V. In this case, a natural choice of $L_{k\ell}$ is

$$L_{k\ell} = \|\mathcal{A}^{k\ell}\|_{\|\cdot\|_{E_\ell}, \|\cdot\|_{E_k,*}} := \max_{z^\ell \in E_\ell} \big\{ \|\mathcal{A}^{k\ell} z^\ell\|_{E_k,*} : \|z^\ell\|_{E_\ell} \le 1 \big\},$$

and we indeed have $L_{k\ell} = L_{\ell k}$.

Now, it is immediately seen that with the assembling (3.21), the ω-radius $\Omega[Z]$ of $Z = Z^1 \times \cdots \times Z^K$ is given by

$$\Omega^2[Z] = \sum_{k=1}^{K} \alpha_k^{-1} \Omega^2[Z^k],$$

where $\Omega[Z^k] = \sqrt{2[\max_{Z^k} \omega_k(\cdot) - \min_{Z_k} \omega_k(\cdot)]}$ is the ω_k-radius of Z^k. Further, the (natural upper bound on the) $(\|\cdot\|_E, \|\cdot\|_{E,*})$ Lipschitz constant of the vector field F is

$$\mathcal{L} = \sigma_{\max}\big(\big[L_{k\ell}\sqrt{\alpha_k \alpha_\ell}\big]_{1 \le k,\ell \le K}\big),$$

where $\sigma_{\max}(\cdot)$ is the maximal singular value of a matrix. The efficiency estimate of Mirror Prox associated with the resulting proximal set-up is $(\Omega^2[Z]\mathcal{L})/2t$ (see (3.19)), and the optimal choice of α_k should minimize $\Omega^2[Z]\mathcal{L}$ with the $\Omega[Z]$, $\mathcal{L}$ just defined. An optimal solution to the latter

problem is given by

$$\alpha_k = \frac{1}{2}\Omega^2[Z^k]\frac{\sum_{\kappa,\ell} L_{\kappa\ell}\Omega[Z^\kappa]\Omega[Z^\ell]}{\sum_\ell L_{k\ell}\Omega[Z^k]\Omega[Z^\ell]}, \quad \text{for } 1 \le k \le K, \tag{3.22}$$
$$\text{which results in} \quad \Omega^2[Z] = 2, \quad \mathcal{L} = \frac{1}{2}\sum_{k,\ell} L_{k\ell}\Omega[Z^k]\Omega[Z^\ell],$$

and the resulting Mirror Prox efficiency estimate is

$$\text{DualityGap}(u^t, v^t) \le \frac{\sum_{k,\ell} L_{k\ell}\Omega[Z^k]\Omega[Z^\ell]}{2t}. \tag{3.23}$$

3.4. Implementing the strategy

In summary of our developments we obtain the following algorithm for solving (3.1) to a given accuracy ϵ.

Algorithm 3.4 (solving (3.1) to given accuracy ϵ).

(1) The set-up for our algorithm is given by a proximal set-up $\|\cdot\|_{E_u}, \omega_U(\cdot)$ for $U = \{u \in E_u : \|u\| \le 1\}, E_u$ and by a proximal set-up $\|\cdot\|_{E_v}, \omega_V(\cdot)$ for $V = \{v \in E_v : \pi_*(v) \le 1\}, E_v$.

(2) Problem (3.1) is solved by the root-finding routine presented in Section 3.1 with ρ_1 chosen according to (3.12), and the auxiliary bilinear saddle-point problems (3.6) solved by the Mirror Prox algorithm (Algorithm 3.2).

(3) The proximal set-up for MP is obtained from those indicated in (1) by the construction presented in Section 3.3. In the notation from this subsection, we have $K = 2$, $Z^1 = U$, $Z^2 = V$. Further, the affine vector fields $F = F^s$ associated with problems (3.6), $s = 1, 2, \ldots$ differ from each other by constant terms (see (3.7)), implying that the linear mappings $\mathcal{A}$ associated with different stages are identical to each other:

$$\mathcal{A} = \begin{bmatrix} & A^* \\ -A & \end{bmatrix}.$$

As a result, the assembled proximal set-up is the same for all stages and is given by (*cf.* (3.22))

$$\|[u;v]\|_E = \sqrt{\frac{\|u\|^2_{E_u}}{\Omega^2[U]} + \frac{\|v\|^2_{E_v}}{\Omega^2[V]}}, \quad \omega(u,v) = \frac{\omega_U(u)}{\Omega^2[U]} + \frac{\omega_V(v)}{\Omega^2[V]}, \tag{3.24}$$

where $\Omega[U], \Omega[V]$ are the ω_U-radius and ω_V-radius of U, V, respectively. For every $t = 1, 2, \ldots$, let z^t_s denote the approximate solution obtained

in t iterations by the MP, Algorithm 3.2, as applied to (3.6). The complexity bound (3.23) then implies that, for every s,

$$\mathrm{DualityGap}^s(z_s^t) \le \varepsilon_t := \frac{\Omega[U]\Omega[V]\|A\|_{\|\cdot\|_{E_u},\|\cdot\|_{E_v,*}}}{t}, \tag{3.25}$$

with $\|A\|_{\|\cdot\|_{E_u},\|\cdot\|_{E_v,*}}$ given by (1.3). ◇

The computational effort per iteration in the 'stand-alone' MP algorithm (Algorithm 3.2) is dominated by the need to compute two values of the vector field F to which the algorithm is applied, together with two values of the prox-mapping. With the strategy from Section 3.1, this should be augmented by computing $F^s(z^t)$ and the bounds Φ_t^+ and $\Phi_t^-(\cdot)$. These additional computations keep the complexity of an iteration basically intact. Since F^s is affine, $F^s(z^t)$ is readily given by the vectors $F^s(w_\tau)$, $1 \le \tau \le t$, which are computed in any case. Given $F^s(z^t)$, computing the above bounds reduces to maximizing two given linear forms, one depending on u and the other on v, over U, V, respectively, which is certainly no more complicated than computing a single prox-mapping. Combining these observations, relation (3.25) and Lemma 3.1, we arrive at the following result.

Theorem 3.5. Under the Standing Assumption, given $\epsilon > 0$ and solving (3.1) by the outlined algorithm, an ϵ-solution to (3.1) is found in no more than

$$\frac{1}{\ln(2/\theta)} \ln\left(1 + \frac{3\pi(b)}{\pi(b) - \delta} \cdot \frac{1}{\nu}\right), \quad \nu = \frac{\epsilon}{\mathrm{Opt} \cdot \|A\|_{\|\cdot\|,\pi(\cdot)}} \tag{3.26}$$

stages, with at most

$$N = \mathrm{Ceil}\left(\frac{\varkappa\Omega[U]\Omega[V]}{\widetilde{\nu}}\right), \quad \varkappa = \frac{\theta}{\theta - 1}, \quad \widetilde{\nu} = \frac{\epsilon}{\mathrm{Opt} \cdot \|A\|_{\|\cdot\|_{E_u},\|\cdot\|_{E_v,*}}} \tag{3.27}$$

Mirror Prox iterations per stage. Here $\theta \in (1, 2)$ is the parameter of the root-finding procedure (see Section 3.1), and $\Omega[U], \Omega[V]$ are the $\omega_U(\cdot)$ radius of U and $\omega_V(\cdot)$ radius of V, respectively.

The computational effort per Mirror Prox iteration is dominated by the need to carry out two matrix–vector multiplications $(u, v) \mapsto (A^*v, Au)$, and to compute two values of the prox-mapping associated with the proximal set-up from step (3) of Algorithm 3.4.

3.5. Discussion

Let us look at what we get in the case of the Dantzig selector problems (3.1) of our primary interest. We shall assume that a suitable $\pi(\cdot)$ in (3.1) is the $\|\cdot\|_\infty/\|\cdot\|_2$ norm, that is, the image space E_v of the linear mapping A is $\mathbb{R}^{\ell_1} \times \cdots \times \mathbb{R}^{\ell_m}$, and for $v = [v^1; \ldots; v^m] \in E_v$ we have $\pi(v) = \max_{1 \le i \le m} \|v^i\|_2$. This choice allows us to treat in a unified fashion the cases of our primary

interest, as specified in the Introduction, specifically those of the uniform fit (where $\ell_1 = \cdots = \ell_m = 1$) and the ℓ_2 fit (where $m = 1$). Moreover, the ℓ_∞/ℓ_2 fit is important in its own right: it arises in some Dantzig-selector-type recoveries dealing with block-sparse signals (see Juditsky *et al.* 2013*b*). In what follows, we treat the parameter $\theta \in (1, 2)$ of the algorithm as fixed and thus as an absolute constant.

(A) We start with the case when the proximal set-ups underlying the algorithm use

$$\|\cdot\|_{E_u} = \|\cdot\|, \quad \|\cdot\|_{E_v} = \pi_*(\cdot). \tag{3.28}$$

Note that in this case

$$\|A\|_{\|\cdot\|_{E_u},\|\cdot\|_{E_v,*}} = \|A\|_{\|\cdot\|,\pi(\cdot)} = \max_u \big\{\pi(Au) : \|u\| \le 1\big\}, \quad \widetilde{\nu} = \nu. \tag{3.29}$$

With our $\pi(\cdot)$, the norm $\|\cdot\|_{E_v} \equiv \pi_*(\cdot)$ is the ℓ_1/ℓ_2 norm, and V is the unit ball of this norm. The latter allows us to specify $\omega_V(\cdot)$ as in Theorem 2.1, so that

$$\|[v^1; \ldots; v^m]\|_{E_v} = \sum_{i=1}^m \|v^i\|_2, \tag{3.30}$$

$$\omega_V(v^1, \ldots, v^m) = \frac{1}{p\gamma} \sum_{i=1}^m \|v^i\|_2^p, \quad \Omega[V] \le O(1)\sqrt{\ln(m+1)},$$

with γ and p given by (2.5) (where one should replace n with m).

The case of the ℓ_1/ℓ_2 norm $\|\cdot\|$. Here $E_u = \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^n$, and

$$\|[u^1; \ldots; u^n]\| \equiv \|u := [u^1; \ldots; u^n]\|_{E_u} := \sum_{j=1}^n \|u^j\|_2,$$

with $u^j \in \mathbb{R}^{k_j}$. Further, U is the unit ball of $\|\cdot\|_{E_u}$, and we can take as $\omega_U(\cdot)$ the DGF given by Theorem 2.1, thus arriving at

$$\|[u^1; \ldots; u^n]\|_{E_u} = \sum_{j=1}^n \|u^j\|_2, \tag{3.31}$$

$$\omega_U(u^1, \ldots, u^n) = \frac{1}{p\gamma} \sum_{j=1}^n \|u^j\|_2^p, \quad \Omega[U] \le O(1)\sqrt{\ln(n+1)},$$

with γ and p given by (2.5). With these choices, the quantity

$$\|A\|_{\|\cdot\|_{E_u}\|\cdot\|_{E_v,*}} = \|A\|_{\|\cdot\|,\pi(\cdot)}$$

is as follows. A is a linear mapping from $E_u = \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_n}$ to $E_v = \mathbb{R}^{\ell_1} \times \cdots \times \mathbb{R}^{\ell_m}$, and as such can be naturally identified with a block matrix

with $m \times n$ blocks $A^{ij} \in \mathbb{R}^{\ell_i \times k_j}$. It is immediately seen that

$$\|A\|_{\|\cdot\|,\pi(\cdot)} = \max_{i \le m, j \le n} \sigma_{\max}(A^{ij}). \tag{3.32}$$

Applying Theorem 3.5, finding an ϵ-solution to (3.1) requires at most S stages, where

$$S = O(1) \ln\left(1 + \frac{3 \max_i \|b^i\|_2}{\max_i \|b^i\|_2 - \delta} \cdot \frac{1}{\nu}\right), \tag{3.33}$$
$$\nu := \frac{\epsilon}{\mathrm{Opt} \cdot \|A\|_{\|\cdot\|,\pi(\cdot)}} = \frac{\epsilon}{\mathrm{Opt} \cdot \max_{i,j} \sigma_{\max}(A^{ij})},$$

with at most

$$O(1)\sqrt{\ln(m+1)\ln(n+1)}\mathrm{Ceil}(1/\nu) \tag{3.34}$$

iterations per stage. Note that ν can be naturally interpreted as relative accuracy.

The computational effort per iteration is dominated by the need to compute $O(1)$ matrix–vector products $(u, v) \mapsto (A^*v, Au)$, plus a 'computational overhead' of $O(1)(\dim u + \dim v)$ arithmetic operations.

In attempting to utilize the outlined approach for the Dantzig selector with the TV norm in the role of ℓ_1/ℓ_2, one runs into the same difficulties as in the case of lasso problems: see Section 2.3.

The case of $\|\cdot\| = \|\cdot\|_{\mathrm{nuc}}$. Now consider the case when $E_u = \mathbb{R}^{\mu \times \nu}$ and $\|\cdot\|$ is the nuclear norm on E_u. Without loss of generality we can assume $2 \le \mu \le \nu$. Taking into account that U is the unit ball of $\|\cdot\|_{E_u} \equiv \|\cdot\|$, we can choose $\omega_U(\cdot)$ according to Theorem 2.3, thus arriving at

$$\|u\|_{E_u} = \|u\|_{\mathrm{nuc}} = \sum_{i=1}^{\mu} \sigma_i(u), \tag{3.35}$$
$$\omega_U(u) = \alpha \sum_{i=1}^{\mu} \sigma_i^{1+q}(u), \quad \Omega[U] \le O(1)\sqrt{\ln(\mu+1)},$$

with α and q readily given by (2.9). The quantity

$$\|A\|_{\|\cdot\|_{E_u},\|\cdot\|_{E_v,*}} = \|A\|_{\|\cdot\|_{\mathrm{nuc}},\pi(\cdot)}$$

is as follows: the linear mapping $A : \mathbb{R}^{\mu \times \nu} \to E_v$ can be split into 'blocks' $A^i : \mathbb{R}^{\mu \times \nu} \to \mathbb{R}^{\ell_i}$, according to the direct product structure of E_v, and $\|A\|_{\|\cdot\|_{\mathrm{nuc}},\pi(\cdot)} = \max_i \|A^i\|_{\|\cdot\|_{\mathrm{nuc}},\|\cdot\|_2}$. For example, in the plain matrix completion problem, where the mapping $u \mapsto Au$ is the restriction of u to a given m-element pattern of cells and $\pi(\cdot)$ is the ℓ_∞/ℓ_2 norm, one can safely bound $\|A\|_{\|\cdot\|_{\mathrm{nuc}},\pi(\cdot)}$ from above by 1.

The algorithm we obtain finds an ϵ-solution to problem (3.1) in at most S stages, where

$$S = O(1)\ln\left(1 + \frac{3\max_i \|b^i\|_2}{\max_i \|b^i\|_2 - \delta} \cdot \frac{1}{\nu}\right), \tag{3.36}$$
$$\nu := \frac{\epsilon}{\mathrm{Opt} \cdot \|A\|_{\|\cdot\|_{\mathrm{nuc}},\pi}} = \frac{\epsilon}{\mathrm{Opt} \cdot \max_i \|A^i\|_{\|\cdot\|_{\mathrm{nuc}},\|\cdot\|_2}},$$

with at most

$$O(1)\sqrt{\ln(m+1)\ln(\mu+1)}\mathrm{Ceil}(1/\nu) \tag{3.37}$$

iterations per stage.

The computational effort per iteration is dominated by the need to compute $O(1)$ matrix–'vector' products $(u, v) \mapsto (A^*v, Au)$ and $O(1)$ SVDs of matrices from $\mathbb{R}^{\mu\times\nu}$, plus a computational overhead of $O(1)(\mu\nu + \dim E_v)$ arithmetic operations.

Observe that Remark 2.5 is fully applicable to our present situation, implying the possibility of extending the scope of the MP-based techniques beyond the case of the $\|\cdot\|$-unit ball in the role of U, as discussed in Section 2.3.

(B) Coming back to 'plain' Dantzig-selector-type problems (3.1) with ℓ_1/ℓ_2 or the nuclear norm in the role of $\|\cdot\|$ and the ℓ_∞/ℓ_2 norm in the role of $\pi(\cdot)$, note that so far we have been considering the case of (3.28) and proximal set-ups for U, E_u and V, E_v developed in Section 2.1.1. An alternative, just as in the composite minimization case, is to use the plain Euclidean set-ups for U, V, that is, to use $\|u\|_{E_u} = \sqrt{\langle u, u\rangle}$, $\omega_U(u) = \frac{1}{2}\langle u, u\rangle$, and similarly for V, E_v. With our U and V, this results in $\Omega[U] = \Omega[V] = 1$. The assembled set-up for $Z = U \times V, E = E_u \times E_v$ is in this case again the plain Euclidean set-up. When passing from our original set-ups to the Euclidean one, the complexity per iteration remains basically intact. Inequality (3.25) becomes

$$\mathrm{DualityGap}^s(z_s^t) \le \varepsilon_t := \frac{\|A\|_{2,2}}{t}, \quad \|A\|_{2,2} = \max_u\big\{\|Au\|_2 : \|u\|_2 \le 1\big\} \tag{3.38}$$

(for $\|\cdot\| = \|\cdot\|_{\mathrm{nuc}}$, $\|u\|_2$ is the Frobenius norm of u), the upper bounds (3.33) and (3.36) on the number of stages remain intact, and the upper bounds (3.34) and (3.37) on the number of iterations per stage become

$$O(1)\mathrm{Ceil}(1/\widetilde{\nu}), \quad \widetilde{\nu} = \frac{\epsilon}{\mathrm{Opt} \cdot \|A\|_{2,2}}. \tag{3.39}$$

Note that $\|A\|_{2,2}$ is never less than the norms $\|A\|_{\|\cdot\|,\pi}$ appearing in (3.33) and (3.36), and can be significantly larger. For example, when $\|\cdot\|$ is the plain $\|\cdot\|_1$ norm on $\mathbb{R}^n$, the ratio $\|A\|_{2,2}/\|A\|_{\|\cdot\|,\pi}$ can be as large as $\sqrt{n}$ when $\pi(\cdot)$ is the plain ℓ_2 norm, and as large as $\sqrt{mn}$ when π is the plain ℓ_∞ norm on $\mathbb{R}^m$. When $\|\cdot\|$ is the nuclear norm on $\mathbb{R}^{\mu\times\nu}$, the ratios in question

can be as large as $O(1)\sqrt{\mu}$ (π is the ℓ_2 norm) and $O(1)\sqrt{\mu m}$ (π is the ℓ_∞ norm). With this in mind, looking at (3.34), (3.37) and (3.39) we conclude that in our context, from the worst-case perspective, the Euclidean set-up is significantly outperformed by our original set-ups. Moreover, the original set-ups result in the best efficiency estimates achievable in the framework of our algorithmic scheme, up to logarithmic dependence on the problem size: see Section A.6. This being said, note that the worst-case-oriented analysis is not exactly the same as computational practice, where the Euclidean set-up behaves surprisingly well in many cases.

3.6. Comments

Some comments on the results of this section are in order.

3.6.1. On-line adjustable step sizes

(See the concluding remarks in Section 2.) Theorem 3.3 states that whenever the step sizes γ_τ ensure (3.17), the accuracy estimate (3.18) holds true; looking at this estimate, we see that the larger the step sizes the better. Theorem 3.3 also states that (3.17) is ensured by choosing $\gamma_\tau = L^{-1}$. The question is whether we could use step sizes which are larger than the 'safe' L^{-1}. The simplest way to achieve this goal is as follows. At the beginning of iteration τ of the MP algorithm, we have at our disposal a certain 'guessed step size' $\widehat{\gamma}_\tau \geq L^{-1}$, with $\widehat{\gamma}_1 = L^{-1}$. We start iteration τ with an attempt to use $\widehat{\gamma}_\tau$ as γ_τ. If this step size results in $\delta_\tau \leq 0$ (see (3.17)), we set $\widehat{\gamma}_{\tau+1} = \vartheta\widehat{\gamma}_\tau$ and pass to step $\tau + 1$; here $\vartheta > 1$ is the method's parameter. If $\gamma_\tau = \widehat{\gamma}_\tau$ results in $\delta_\tau > 0$, we reset γ_τ according to $\gamma_\tau := \min[\mu\gamma_\tau, 1/L]$, where $\mu \in (0, 1)$ is another method's parameter, and repeat the step with this reduced step size. If we get $\delta_\tau > 0$ with the reduced value of γ_τ also, we iterate, reducing the 'trial step size' once more, and perform a new trial step, and so on until $\delta_\tau \leq 0$ is obtained (this must eventually happen, since $\gamma_\tau = L^{-1}$ ensures $\delta_\tau \leq 0$). When $\delta_\tau \leq 0$ is obtained, we pass to step $\tau + 1$, using the latest value of γ_τ as $\widehat{\gamma}_{\tau+1}$. Extensive numerical experimentation (where we use $\vartheta = 1.2$ and $\mu = 0.8$) demonstrates that, with this on-line step size policy, the total number of values of F (or equivalently the total number of prox-mappings) we need to compute to arrive at a desired value of the duality gap is never significantly larger than the similar quantity for the off-line worst-case-oriented step sizes $\gamma_\tau \equiv L^{-1}$, and is typically orders of magnitude less. Note that this policy, with obvious modifications, can be utilized in the case when we know that F is Lipschitz-continuous but do not know the corresponding Lipschitz constant.

3.6.2. Nearly dimension-independent rate of convergence

It should be stressed that in the situations we have considered, the resulting efficiency estimates, *i.e.*, the total iteration count as a function of the desired

relative accuracy ν given by (3.33) or (3.36), up to a *logarithmic* factor in $1/\nu$ and the problem size, is just $1/\nu$. Logarithmic dependence of the efficiency estimate on the problem size reflects the 'nice geometry' of the norms $\|\cdot\|$, $\pi_*(\cdot)$ in question. Along with these and several other 'nice geometry' norms, there exist 'bad geometry' ones, *e.g.*, the norm $\|\cdot\|_\infty$ on $\mathbb{R}^n$ and its matrix analogue: the spectral norm $\sigma_{\max}(\cdot)$. For the two latter norms in the roles of $\|\cdot\|$ or $\pi_*(\cdot)$, the approach we have developed does *not* yield 'nearly dimension-independent' efficiency estimates, but this is equally true for any other approach known to us.

3.6.3. More general norms $\|\cdot\|$

The scope of the approach we have developed is not restricted to the norms $\|\cdot\|$, $\pi(\cdot)$ considered so far; all we need are 'good proximal set-ups' for the unit balls U, V of the norms $\|\cdot\|$, $\pi_*(\cdot)$, *i.e.*, moderate $\Omega[U]$ and $\Omega[V]$ and computationally tractable prox-mappings. As an instructive example, consider recovery of a matrix ('image') x from noisy observations of Px. In some cases it is natural to assume that the true image is the sum of sparse and low-rank images.[14] The Dantzig selector form of the problem is

$$\min_{y^1,y^2\in\mathbb{R}^{\mu\times\nu}}\{\mu_1\|y^1\|_1+\mu_2\|y^2\|_{\rm nuc}:\pi(P(y^1+y^2)-b)\le\delta\},$$

where $\mu_1>0$, $\mu_2>0$ are the penalty parameters and $\|y^1\|_1=\sum_{i,j}|y^1_{ij}|$. Equivalently, the recovery problem is

$$\min_{x=(x^1,x^2)\in E_u:=\mathbb{R}^{\mu\times\nu}\times\mathbb{R}^{\mu\times\nu}}\big\{\|x\|:\pi(Ax-b:=\mu_1^{-1}Px^1+\mu_2^{-1}Px^2-b)\le\delta\big\},$$
$$\|x:=(x^1,x^2)\|\equiv\|x^1\|_1+\|x^2\|_{\rm nuc}.$$

All we need to process the resulting problem is a good proximal set-up for the unit ball U of the norm $\|\cdot\|$, and such a set-up is readily given by the results of Section 2.1.1:

$$\|\cdot\|_{E_u}=\|\cdot\|,\quad \omega_U(u^1,u^2)=2\omega_1(\mathrm{Vec}(u^1))+2\omega_2(u^2),$$

where $\mathrm{Vec}(u^1)$ is the 'vectorization' of u^1, and the functions $\omega_1(\cdot)$, $\omega_2(\cdot)$ are given by (2.5) (where $n=\mu\nu$) and (2.9) respectively; note that the ω_U-radius of U does not exceed $O(1)\sqrt{\ln(\mu\nu)}$. When $\pi(\cdot)=\|\cdot\|_2$, we are also interested in processing the lasso version of the problem,

$$\min_{x=(x^1,x^2)}\{\|(x^1,x^2)\|+\|Ax-b\|_2^2\}.$$

[14] As a toy example, imagine the façade of a building, a black rectangle ('wall') with a regular grid of white rectangular holes ('windows'); this image has rank 2. Adding a few irregularities ('architectural elements') occupying a relatively small part of the façade, we get an image which is the sum of low-rank and sparse images.

As above, the required proximal set-up is found to be

$$\|\cdot\|_E = \|\cdot\|, \quad \omega_E(u^1, u^2) = 2\widehat{\omega}_1(u^1) + 2\widehat{\omega}_2(u^2),$$

with $\widehat{\omega}_1(\cdot)$, $\widehat{\omega}_2(\cdot)$ given by (2.7) (where $n = \mu\nu$) and (2.11), respectively. This set-up ensures that the quantity $V_{z_\omega}(x^*)$ appearing in the bounds (2.21), (2.23) and (2.25) does not exceed $O(1)\ln(\mu\nu)\|x^*\|^2$. Finally, the computational cost of the prox-mapping associated with $\|\cdot\|$ is basically the same as the cost of computing the SVD of a matrix from $\mathbb{R}^{\mu\times\nu}$.

3.6.4. Plain saddle-point forms of ℓ_1/nuclear norm minimization
Along with lasso and Dantzig selector recovery routines, sparsity/low-rank oriented signal processing considers other recovery routines, *e.g.*, the *penalized recovery* (Juditsky and Nemirovski 2011*c*, Juditsky *et al.* 2013*b*, Juditsky *et al.* 2011*b*)

$$\widehat{x} \in \arg\min_{x\in E_x}\big\{\|x\| + \lambda\pi(Ax-b)\big\}, \quad \text{for } \lambda > 0, \tag{3.40}$$

possessing attractive statistical properties. Assuming for the sake of definiteness that $\|\cdot\|$ is either the ℓ_1/ℓ_2 norm on $E_x = \mathbb{R}^{k_1+\cdots+k_n}$ or the nuclear norm on $E_x = \mathbb{R}^{\mu\times\nu}$, and $\pi(\cdot)$ is the ℓ_∞/ℓ_2 norm on $E_v = \mathbb{R}^{\ell_1+\cdots+\ell_m}$ (this is indeed the case in typical applications), and rewriting the problem equivalently as

$$\mathrm{SadVal} = \min_{u=[x;\xi]\in U}\max_{v\in V}\big[\phi(u := [x;\xi], v) = \xi + \lambda\langle v, Ax-b\rangle\big], \tag{3.41}$$

$$U = \{[x;\xi] : \xi \geq \|x\|\} \subset E_u := E_x \times \mathbb{R}, \quad V = \{v \in E_v : \pi_*(v) \leq 1\},$$

we see that in this case we need to solve just one bilinear saddle-point problem rather than a 'small series' of them, so that a single application of MP is enough. Equipping V, E_v with the same ℓ_1/ℓ_2 proximal set-up $\|\cdot\|_{E_v} = \pi_*(v)$, $\omega_V(\cdot)$ as above, and U, E_u with the proximal set-up

$$\|u := [x;\xi]\|_{E_u} = \|x\| + |\xi|, \quad \omega_U(u := [x;\xi]) = 2\omega(u) + \xi^2,$$

where $\omega(u)$ is given by (2.7) or (2.11), depending on whether $\|\cdot\|$ is the ℓ_1/ℓ_2 or the nuclear norm, we can assemble these 'partial' proximal set-ups for U, E_u and V, E_v into a proximal set-up $\|\cdot\|_E$, $\omega(\cdot)$ for $Z = U \times V$, $E = E_u \times E_v$ according to

$$\|[[x;\xi], v]\|_E = \sqrt{\alpha^{-2}(\|x\| + |\xi|)^2 + \pi_*^2(v)},$$
$$\omega([x;\xi], v) = \alpha^{-2}[2\omega(x) + \xi^2] + \omega_V(v),$$

where $\alpha > 0$ is a positive parameter. It is immediately seen that this

assembling indeed results in a proximal set-up. Applying (2.6), (2.8), (2.12) and (3.20), we obtain the efficiency estimate

$$\overline{\phi}(x^t) - \min_x \overline{\phi}(x) \le O(1)\frac{\alpha^{-1}(\alpha\kappa_1 R_* + \kappa_2)^2 \lambda \|A\|_{\|\cdot\|,\pi(\cdot)}}{t}, \quad t = 1, 2, \ldots,$$
$$\text{where} \quad \overline{\phi}(x) = \|x\| + \lambda\pi(Ax - b), \quad R_* = \min_{x_* \in \arg\min_{x \in E_x} \overline{\phi}(x)} \|x\|, \tag{3.42}$$
$$\kappa_1 = \begin{cases} \sqrt{\ln(n+1)}, & E_x = \mathbb{R}^{k_1+\cdots+k_n}, \|x := [x^1; \ldots; x^n]\| = \sum_j \|x^j\|_2, \\ \sqrt{\ln(\mu+1)}, & E_x = \mathbb{R}^{\mu\times\nu}(\mu \le \nu), \|x\| = \|x\|_{\mathrm{nuc}}, \end{cases}$$
$$\kappa_2 = \sqrt{\ln(m+1)}.$$

Here x^t is the x-component of the approximate solution obtained by t MP iterations. Assuming $R_* > 0$ and optimizing the efficiency estimate over the 'assembling parameter' α, we obtain $\alpha = \alpha_* = \kappa_2/(\kappa_1 R_*)$ and the efficiency estimate

$$\overline{\phi}(x^t) - \min_x \overline{\phi}(x) \le O(1)\frac{\kappa_1\kappa_2\lambda\|A\|_{\|\cdot\|,\pi(\cdot)} R_*}{t}.$$

The outlined choice of α is usually unrealistic; a realistic choice is $\alpha = \kappa_2/(\kappa_1 R)$, where R is our guess for R_*. The efficiency estimate associated with this α is within the factor $O(1)\max[R/R_*, R_*/R]$ of the 'optimal' efficiency. Finally, note that the computational price of the prox mapping associated with our set-up is basically the same as in the case when U is the $\|\cdot\|$-unit ball of E_u: see Section 2.1.1.

3.6.5. Accelerating the MP algorithm

Under favourable structural assumptions on the cost function ϕ of (3.13), the MP algorithm can be accelerated (Juditsky and Nemirovski 2011*b*, Section 6.4); in particular, in the case of the ℓ_1/ℓ_2 or nuclear norm $\|\cdot\|$, a properly accelerated MP algorithm allows us to solve lasso-type problems (1.1a) with the same $O(1/t^2)$ efficiency estimate as the fast gradient method from Section 2.

3.6.6. Randomization

When solving ℓ_1/nuclear norm minimization problems (1.1a,b) to medium relative accuracy, the 'practical grasp' of first-order algorithms is limited by our abilities to carry out, in reasonable time, several hundred (or a few thousand) matrix–vector multiplications involving A and A^* and, in the case of nuclear norm minimization and $x \in \mathbb{R}^{\mu\times\nu}$, $\mu \le \nu$, the ability to compute the same number of SVDs of $\mu \times \nu$ matrices. While these limitations are essentially less restrictive than those for IPMs, where one needs to assemble

and solve a few tens of linear systems with $\dim x$ unknowns,[15] they cannot be neglected. Let us discuss the related difficulties and options.

In ℓ_1 or ℓ_1/ℓ_2 minimization, with their inexpensive prox-mappings, difficulties arise when the required matrix–vector multiplications become prohibitively time-consuming. There is, however, a way to overcome these difficulties to some extent, by *randomizing* matrix–vector multiplications.

The simplest way to randomize the matrix–vector multiplication $u \mapsto Bu : \mathbb{R}^N \to \mathbb{R}^M$ is as follows. Let us associate with $u \in \mathbb{R}^N$ a stochastic vector with the entries $|u_j|/\|u\|_1$. Treat this vector as a probability distribution on the set $\{B_1, \ldots, B_N\}$ of columns of B, draw from this distribution a random column $B_\jmath$, and return the vector $\xi = \|u\|_1 \mathrm{sign}(u_\jmath) B_\jmath$. It is immediately seen that the resulting random estimate ξ of Bu is unbiased: $\mathbb{E}\{\xi\} = Bu$. Generating this estimate costs $O(M + N)$ arithmetic operations,[16] which compares favourably with the cost of $O(1)MN$ arithmetic operations of the deterministic computation of the precise value of Bu.[17] Of course, the unbiased estimate ξ of Bu is noisy, with the magnitude of the noise, measured in some norm $\|\cdot\|$, bounded by $\|u\|_1 \|B\|_{\|\cdot\|_1, \|\cdot\|} = \|u\|_1 \max_j \|B_j\|$.

It turns out that the strategy we have developed for (3.1), and its MP-based implementation, 'survive' replacing precise matrix–vector products $(u, v) \mapsto (A^*v, Au)$ by their unbiased random estimates, albeit at the price of a certain growth in the iteration count and sensitivity of the approach to the geometry of the domains U, V underlying (3.6).[18] The potential of the outlined randomization in the context of ℓ_1 minimization problems (1.1b) with the uniform or the ℓ_2 fit was investigated by Juditsky *et al.* (2013*a*) (see also Juditsky and Nemirovski 2011*b*), whose results can be summarized as follows.

- For ℓ_1 minimization with the uniform and ℓ_2 fit, and with a dense 'general' $m \times n$ matrix A, building an ϵ-solution to (3.1) with the *deterministic* algorithm we have developed takes ν^{-1} iterations,[19] with

[15] Indeed, assuming that the image and the argument dimensions of the mapping $x \to Ax$ are of the order of $n := \dim x$, a single matrix–vector multiplication costs $O(1)n^2$ arithmetic operations (or less when A is 'nicely structured', *e.g.*, sparse, or discrete-convolution-like or discrete Fourier/wavelet transform), and a single SVD costs $O(1)\mu^2\nu \le O(1)n^{3/2}$, while the cost of a single IPM iteration is of the order of n^3 in at least some of the applications we are interested in.

[16] Assuming, as is usually the case, an $O(1)M$-arithmetic-operation cost of extracting a column from B given its index.

[17] In this comparison, we assume that B has no 'nice structure' allowing for fast matrix–vector multiplications

[18] This sensitivity primarily stems from the fact that the noisiness of the unbiased estimate of Bu is in proportion to $\|u\|_1$ and can thus be large when U or V contains vectors of large $\|\cdot\|_1$ norm.

[19] Here and in the discussion to follow, we skip log-factors.

mn arithmetic operations per iteration, where the A_j are the columns of A, $\nu = \epsilon(\mathrm{Opt} \cdot \max_{1\le j\le n} \|A_j\|_p)^{-1}$ is the relative accuracy, and $p = \infty$ (uniform fit) or $p = 2$ (ℓ_2 fit).[20] The overall arithmetic complexity is therefore $mn\nu^{-1}$, which, as far as we know, is the best possible complexity bound achievable with existing *deterministic* algorithms for the large-scale problems under consideration.

- For ℓ_1 minimization with uniform fit, randomization allows us to reduce the cost of an iteration to $m + n$ at the price of increasing the iteration count to ν^{-2}. Thus the overall complexity becomes $(m + n)\nu^{-2}$. We see that, with ν fixed and $n = O(m)$ growing, the randomized algorithm eventually progressively outperforms the deterministic ones. Numerical experiments demonstrate that randomization results in fairly significant acceleration (by a factor of around 5.4 in terms of CPU time) for 'not too large' problems (specifically, those with $8\,000 \times 16\,000$ randomly generated matrices). It should be added that when $\ln(n) = O(\ln(m))$ and m, n are large, the randomized algorithm generates an ϵ-solution with high confidence by inspecting a negligible fraction (tending to zero as $m, n \to \infty$) of the data, thus exhibiting *sublinear time* behaviour.[21]

- For ℓ_1 minimization with ℓ_2 fit, the situation is the same as that of the uniform fit, modulo two points: first, δ in (1.1b) should be 'small enough' (specifically, $\le O(1)\|b\|_2/\sqrt{m}$);[22] second, we need $|A_{ij}| \le O(1)\max_j \|A_j\|_2\sqrt{\ln(m)/m}$ (roughly speaking, the entries in A should be of nearly the same magnitude). Juditsky *et al.* (2013*a*) showed that the latter assumption can be ensured by an appropriate randomized preprocessing which costs $mn\ln(m)$ arithmetic operations; note that when $\nu \ll 1$ is fixed and m, n are large, the cost of this preprocessing is negligibly small compared to the cost of the deterministic algorithm. Finally, for the ℓ_2 fit, the experimentally observed acceleration due to randomization in Juditsky *et al.* (2013*a*), although smaller than for the uniform fit, is still quite significant (by a factor of 2.4 on $8\,000 \times 16\,000$ problems).

[20] See (3.34) and (3.26), and take into account that we are in a situation where either for every j the blocks A^{ij} are just entries in A (uniform fit), or for every j there exists only one block A^{ij}, specifically the entire column A_j (ℓ_2 fit).

[21] For ℓ_1 minimization with uniform fit, the saddle-point problems (3.6) can be reduced to matrix games. The possibility of solving matrix games in sublinear time was discovered in Grigoriadis and Khachiyan (1995). The *ad hoc* randomized matrix game algorithm proposed in this reference is in hindsight pretty similar – though not identical – to the 'uniform fit' version of the algorithm from Juditsky *et al.* (2013*a*). In particular, both algorithms share the same complexity bound.

[22] This restriction can be lifted: see Juditsky and Nemirovski (2011*b*).

Table 3.1. Randomization making a difference: $32\,000 \times 64\,000$ matrix A, $\delta = 0.005$, termination on either achieving accuracy $\epsilon = 0.0025$ or reaching the CPU limit of 7 200 sec. The computational platform is the same as in Table 1.1.

	Method	Steps	CPU (sec)	$\|Ax_\epsilon - b\|_p$
$p = \infty$	DMP†	31	7 363	0.1598
	SMP‡	7 501	5 352	0.0074
$p = 2$	DMP	30	7 536	0.0248
	SMP	2 602	2 350	0.0072

†Deterministic MP with on-line adjustable step size policy.
‡Stochastic, or randomized, version of MP (Juditsky *et al.* 2013*a*).

For an impression of the potential of randomization, see Table 3.1. The data are taken from Juditsky *et al.* (2013*a*).

Now, in the case of nuclear norm minimization, the difficulties can come from the impossibility of carrying out in a reasonable time the required number of either (a) SVDs, or (b) matrix–'vector' multiplications $(u, v) \mapsto (A^*v, Au)$. While the difficulties arising in case (a) can sometimes be circumvented (Arora, Hazan and Kale 2005, Arora and Kale 2007, d'Aspremont 2011, Baes, Bürgisser and Nemirovski 2013), we would say that typically case (a) rules out proximal-point algorithms as we know them today. In contrast to this, in case (b) one can try to save the day by randomization. For the sake of definiteness, assume that we are speaking about nuclear norm minimization with uniform fit, so that $u \in \mathbb{R}^{\mu\times\nu}$ and $Au \in \mathbb{R}^m$. Assume also that for every i, the ith entry in Au is given by

$$(Au)_i = \sum_{\ell=1}^{k_i} p_{i\ell}^T u q_{i\ell},$$

where $p_{i\ell} \in \mathbb{R}^\mu$ and $q_{i\ell} \in \mathbb{R}^\nu$. Letting $\#(r)$ denote the number of non-zero entries in the vector r, the straightforward computation of

$$\left(A^*v = \sum_{i=1}^m v_i \sum_{\ell=1}^{k_i} p_{i\ell} q_{i\ell}^T, Au \right)$$

for given u, v costs

$$\mathcal{C}_d = O(1) \sum_{i=1}^m \sum_{\ell=1}^{k_i} \#(p_{i\ell})\#(q_{i\ell}) \text{ arithmetic operations.}$$

Similarly, an unbiased random estimate η of A^*v can be constructed using $\|v\|_1 \mathrm{sign}(v_\imath) \sum_{\ell=1}^{k_\imath} p_{\imath\ell} q_{\imath\ell}^T$, where $\imath \in \{1, \ldots, m\}$ is drawn at random according to $\mathrm{Prob}\{\imath = i\} = |v_i|/\|v\|_1$. The cost of generating a realization of η is

$$\mathcal{C}_\eta = O(1)\left[m + \max_i \sum_{\ell=1}^{k_i} \#(p_{i\ell})\#(q_{i\ell})\right] \text{ arithmetic operations.}$$

To randomize computing Au, one can use a 'matrix analogue' of the randomization used earlier. Specifically, given the SVD $u = \sum_{j=1}^{\mu} \sigma_j(u) e_j f_j^T$ of u,[23] we draw $\jmath \in \{1, \ldots, \mu\}$ at random according to

$$\mathrm{Prob}\{\jmath = j\} = \sigma_j(u)/\|u\|_{\mathrm{nuc}}, \quad 1 \le j \le \mu,$$

and return $\xi \in \mathbb{R}^m$ defined by its components

$$\xi_i = \|u\|_{\mathrm{nuc}} \sum_{\ell=1}^{k_i} (p_{i\ell}^T e_\jmath)(q_{i\ell}^T f_\jmath), \quad 1 \le i \le m.$$

It is immediately seen that ξ is an unbiased random estimate of Au, and generating a realization of ξ given the SVD of u costs

$$\mathcal{C}_\xi = O(1)\left[\mu + \sum_{i=1}^{m} \sum_{\ell=1}^{k_i} [\#(p_{i\ell}) + \#(q_{i\ell})]\right] \text{ arithmetic operations.}$$

To gain an impression of what can be saved here, assume that $k_i = k$, $\#(p_{i\ell}) = \#(q_{i\ell}) = p$ for all i, ℓ. Then, skipping $O(1)$ factors, we obtain

$$\Theta := \frac{\mathcal{C}_\eta + \mathcal{C}_\xi}{\mathcal{C}_d} = \frac{1}{kp^2} + \frac{1}{m} + \frac{\mu}{kmp^2} + \frac{1}{p}.$$

In a meaningful range of the sizes μ, m, k and p, Θ tends to zero as the sizes increase,[24] meaning that randomization can indeed accelerate the solution process.

Finally, we remark that the noise of our random estimates of matrix–vector products can be reduced by taking the average of several realizations of the estimate. For more details on this subject, see Juditsky *et al.* (2013*a*) and Baes *et al.* (2013).

[23] Looking at the description of the MP algorithm, we see that all the u for which Au should be computed are the values (already computed when Au is to be computed) of the prox-mapping associated either with the DGF (2.9) or with the DGF (2.11). As explained in Section 2.1.1, when computing such a value u we in fact get the SVD of u. Thus, in our context, when computing Au, the SVD of u is indeed readily available.

[24] Let, for example, $\mu = \nu$, $m = O(\mu^2)$, $p = O(\mu)$ (square u, the number of observations is of order of $\dim u$, $p_{i\ell}$, $q_{i\ell}$ are dense). In this case, $\mathcal{C}_d = O(1)k\mu^4$, $\Theta \le O(1)\mu^{-1}$ tends to zero as μ grows. In addition, in the case under consideration, the cost per iteration of computing prox-mappings is $O(1)[\mu^3 + m] \le O(1)\mu^3 \le O(1)\mu^{-1}\mathcal{C}_d$, implying that the randomization reduces arithmetic complexity of an iteration by a factor of $O(\mu)$.

Appendix: Proofs

A.1. Relation (2.2)

We need the following simple lemma.

Lemma A.1. Let Z be a closed convex domain in Euclidean space E, let $\|\cdot\|_E, \omega(\cdot)$ be a proximal set-up for Z, E, let $\Psi : Z \to \mathbb{R} \cup \{+\infty\}$ be a lower-semicontinuous convex function which is finite on the relative interior of Z, and let $\phi(\cdot) : Z \to \mathbb{R}$ be a convex continuously differentiable function. Assume that z_+ is a minimizer of the function $\omega(w) + \phi(w) + \Psi(w)$ on Z. Then $z_+ \in Z^o$, and there exists $\Psi' \in \partial\Psi(z_+)$ such that $\langle \omega'(z_+) + \nabla\phi(z_+) + \Psi', w - z_+\rangle \geq 0$ for all $w \in Z$. If $\Psi(\cdot)$ is differentiable at z_*, one can take $\Psi' = \nabla\Psi(z_+)$.

Proof. Let us set $\widehat{\omega}(w) = \omega(w) + \phi(w)$. Since ϕ is continuously differentiable on Z, $\widehat{\omega}$ possesses the same smoothness properties as ω: specifically, $\widehat{\omega}$ is convex and continuous on Z,

$$\{w \in Z : \partial\widehat{\omega}(w) \neq \emptyset\} = Z^o := \{w : \partial\omega(w) \neq \emptyset\},$$

and $\widehat{\omega}'(w) := \omega'(w) + \nabla\phi(w) : Z^o \to E$ is a continuous selection of subgradients of $\widehat{\omega}$.

Without loss of generality we can assume that $\operatorname{int} Z \neq \emptyset$ and that $z_+ = 0$. Note that z_+ is a minimizer of $\widehat{\omega}(w) + \Psi(w)$ over $w \in Z$. Let $\bar{w} \in \operatorname{int} Z$, and let $w_t = t\bar{w}$, so that $w_t \in \operatorname{int} Z$ for $0 < t \leq 1$.

Let us prove first that $z_+ = 0 \in Z^o$. To this end it suffices to verify that the vectors $\widehat{\omega}'(w_t)$ remain bounded as $t \to +0$. Indeed, in this case the set of limiting points of $\widehat{\omega}'(w_t)$ as $t \to +0$ is non-empty. Since $\widehat{\omega}(\cdot)$ is continuous on Z, every one of these limiting points belongs to $\partial\widehat{\omega}(0)$, and thus $\partial\widehat{\omega}(0) \neq \emptyset$, meaning that $0 \in Z^o$. To prove that $\widehat{\omega}'(w_t)$ is bounded as $t \to +0$, let $r > 0$ be such that with $B = \{h \in E : \|h\|_E \leq r\}$ we have $B^+ = \bar{w} + B \subset \operatorname{int} Z$. All we need to prove is that the linear forms $\langle \widehat{\omega}'(w_t), h\rangle$ of $h \in E$ are upper-bounded on B uniformly in $t \in (0, 1/2)$. The convexity of $\widehat{\omega}$ says that these forms are uniformly upper-bounded on B^+, since with $\|h\|_E \leq r$ we have

$$\begin{aligned}\langle \widehat{\omega}'(w_t), \bar{w} + h\rangle &= (1-t)^{-1}\langle \widehat{\omega}'(w_t), [\bar{w} + (1-t)h] - w_t\rangle \\ &\leq (1-t)^{-1}\Big[\max_{w \in B^+} \widehat{\omega}(w) - \min_{0 \leq t \leq 1} \widehat{\omega}(w_t)\Big] \\ &< \infty.\end{aligned}$$

All we need to derive upper bounds on the linear forms $f_t(h) = \langle \widehat{\omega}'(w_t), h\rangle$ on B^+, uniform in $0 < t \leq 1/2$, is to verify lower bounds on $\langle \widehat{\omega}'(w_t), \bar{w}\rangle$ which are uniform in $0 < t \leq 1/2$. Assume, to the contrary, that for some sequence $t_i > 0$ converging to 0 as $i \to \infty$, we have $p_i := \langle \widehat{\omega}'(w_{t_i}), \bar{w}\rangle \to -\infty$, and let $q = \langle \Psi'(\bar{w}), \bar{w}\rangle$ with $\Psi'(\bar{w}) \in \partial\Psi(\bar{w})$. Then, by convexity of Ψ and

due to $w_t = t\bar{w}$, we have $\Psi(0) \geq \Psi(w_t) - qt$, and by convexity of $\widehat{\omega}(\cdot)$ we have $\widehat{\omega}(0) \geq \widehat{\omega}(w_t) - t\langle \widehat{\omega}'(w_t), \bar{w}\rangle$, so that $0 \geq \widehat{\omega}(0) + \Psi(0) - [\widehat{\omega}(w_{t_i}) + \Psi(w_{t_i})] \geq -t_i(p_i + q)$ (recall that $z_+ = 0$ is the minimizer of $\widehat{\omega} + \Psi$ on Z) for all i, which contradicts $p_i \to -\infty$, $i \to \infty$.

Now, since $0 = z_+ \in \operatorname{Dom}\Psi$ and $\operatorname{Dom}\Psi$ contains the interior of Z, from lower-semicontinuity and convexity of Ψ on the segment $[0, \bar{w}]$, it follows that Ψ is continuous on this segment. Setting $\psi(t) = \Psi(w_t)$, $\phi(t) = \widehat{\omega}(w_t)$, observe that from the smoothness properties of $\widehat{\omega}(\cdot)$ it follows that ϕ is continuously differentiable on $[0, 1]$ with the derivative $\phi'(t) = \langle \widehat{\omega}'(w_t), \bar{w}\rangle$. Since $\psi(t) + \phi(t)$ attains its minimum on $[0, 1]$ at $t = 0$, it immediately follows that the right derivative $\psi'(0)$ of ψ at $t = 0$ is $\geq -\phi'(0) = -\langle \widehat{\omega}'(0), \bar{w}\rangle$, whence $\Psi(\bar{w}) - \Psi(0) = \psi(1) - \psi(0) \geq \langle -\widehat{\omega}'(0), \bar{w}\rangle$. The relation $\Psi(\bar{w}) - \Psi(0) \geq -\langle \widehat{\omega}'(0), \bar{w}\rangle$ holds true for all $\bar{w} \in \operatorname{int} Z$ and therefore, by convexity of Ψ, for all $w \in Z$, meaning that $\Psi' := -\widehat{\omega}'(0) \in \partial\Psi(0)$, and of course $\langle \Psi' + \widehat{\omega}'(0), w\rangle \geq 0$ for all w. Finally, when $\Psi(\cdot)$ is differentiable at $z_+ = 0$, we have $\psi'(0) = \langle \nabla\Psi(0), \bar{w}\rangle$. Since we have already seen that $\psi'(0) + \phi'(0) \geq 0$, we get $\langle \nabla\Psi(0) + \widehat{\omega}'(0), \bar{w}\rangle \geq 0$. This inequality holds true for all $\bar{w} \in \operatorname{int} Z$ and thus for all $\bar{w} \in Z$. Recalling that $\widehat{\omega}'(0) = \omega'(0) + \nabla\phi(0)$, the proof is complete. □

Proof of relation (2.2). Applying Lemma A.1 with $\Psi(\cdot) \equiv 0$ and $\phi(w) = \langle \xi - \omega'(z), w\rangle$, so that $z_+ := \operatorname{Prox}_z(\xi)$ is the minimizer of $\omega(w) + \phi(w) + \Psi(w)$ over $w \in Z$, we get $z_+ \in Z^o$ and

$$\langle \omega'(z_+) + \xi - \omega'(z), w - z_+\rangle \geq 0, \quad \text{for all } w \in Z. \tag{A.1}$$

Plugging into (2.2) the definitions of Bregman distances, we immediately see that this is exactly (A.1). □

A.2. Proximal set-ups

Proof of Theorem 2.1. For $z \in Z' = \{z \in Z : z^j \neq 0\, \forall j\}$ we have

$$\gamma D\omega(z)[h] = \sum_{j=1}^{n} \|z^j\|_2^{p-2} \langle z^j, h^j\rangle$$

and

$$\begin{aligned}
\gamma D^2\omega(z)[h,h] &= -(2-p)\sum_{j=1}^{n} \|z^j\|_2^{p-4}[\langle z^j, h^j\rangle]^2 + \sum_{j=1}^{n} \|z^j\|_2^{p-2}\|h^j\|_2^2 \\
&\geq \sum_{j=1}^{n} \|z^j\|_2^{p-2}\|h^j\|_2^2 - (2-p)\sum_{j=1}^{n} \|z^j\|_2^{p-4}\|z^j\|_2^2\|h^j\|_2^2 \\
&\geq (p-1)\sum_{j=1}^{n} \|z^j\|_2^{p-2}\|h^j\|_2^2.
\end{aligned}$$

At the same time,

$$\left[\sum_j \|h^j\|_2\right]^2 = \left[\sum_{j=1}^n [\|h^j\|_2 \|z^j\|_2^{\frac{p-2}{2}}] \|z^j\|_2^{\frac{2-p}{2}}\right]^2 \le \left[\sum_{j=1}^n \|h^j\|_2^2 \|z^j\|_2^{p-2}\right] \left[\sum_{j=1}^n \|z^j\|_2^{2-p}\right],$$

and hence

$$\left[\sum_j \|h^j\|_2\right]^2 \le \left[\sum_{j=1}^n \|z^j\|_2^{2-p}\right] \frac{\gamma}{p-1} D^2\omega(z)[h,h].$$

Setting $t_j = \|z^j\|_2 \ge 0$, we have $\sum_j t_j \le 1$, whence due to $0 \le 2-p \le 1$ we have $\sum_j t_j^{2-p} \le n^{p-1}$. Thus

$$\left[\sum_j \|h^j\|_2\right]^2 \le n^{p-1} \frac{\gamma}{p-1} D^2\omega(z)[h,h], \tag{A.2}$$

while

$$\max_{z\in Z} \omega(z) - \min_{z\in Z} \omega(z) \le \frac{1}{\gamma p}. \tag{A.3}$$

With p, γ as stated in Theorem 2.1, when $n \ge 3$ we get $\frac{\gamma}{p-1} n^{p-1} = 1$, and similarly for $n = 1, 2$. Consequently,

$$\left[\sum_{j=1}^n \|h^j\|_2\right]^2 \le D^2\omega(z)[h,h], \quad \text{for all } z \in Z', h. \tag{A.4}$$

Since $\omega(\cdot)$ is continuously differentiable and the complement of Z' in Z is the union of finitely many proper linear subspaces of E, (A.4) implies that ω is strongly convex on Z, with modulus 1, with respect to the ℓ_1/ℓ_2 norm $\|\cdot\|_E$. Moreover, we have

$$\frac{1}{\gamma p} \left\{ \begin{array}{ll} = \frac{1}{2}, & n = 1 \\ = 1, & n = 2 \\ \le \mathrm{e}\ln(n), & n \ge 3 \end{array} \right\} \le O(1)\ln(n+1),$$

which combines with (A.3) to imply (2.6). □

Proof of Corollary 2.2. Let

$$w(z) = \frac{1}{\gamma p} \sum_{j=1}^n \|z^j\|_2^p : E := \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_n} \to \mathbb{R}, \tag{A.5}$$

with γ, p given by (2.5); note that $p \in (1,2]$. As we know from Theorem 2.1, the restriction of $w(\cdot)$ onto the unit ball of the ℓ_1/ℓ_2 norm $\|\cdot\|$ is continuously differentiable and strongly convex, with modulus 1, with respect to $\|\cdot\|$. Further, $w(\cdot)$ is clearly non-negative and positively homogeneous of degree p, *i.e.*, $w(\lambda z) \equiv |\lambda|^p w(z)$. Note that the outlined properties clearly imply that $w(\cdot)$ is convex and continuously differentiable on all of E. We need the following simple lemma.

Lemma A.2. Let H be a Euclidean space, let $\|\cdot\|_H$ be a norm (not necessarily the Euclidean one) on H, and let $Y_H = \{y \in H : \|y\|_H \leq 1\}$. Further, let $\zeta : H \to \mathbb{R}$ be a non-negative continuously differentiable function, absolutely homogeneous of a degree $p \in (1,2]$, and strongly convex, with modulus 1, with respect to $\|\cdot\|_H$, on Y_H. Then the function

$$\widehat{\zeta}(y) = \zeta^{2/p}(y) : H \to \mathbb{R}$$

is continuously differentiable, homogeneous of degree 2 and strongly convex, with modulus

$$\beta = \frac{2}{p}\big[\min_{y:\|y\|_H=1} \zeta^{2/p-1}(y)\big],$$

with respect to $\|\cdot\|_H$, on all of H.

Proof. It is evident that $\widehat{\zeta}(\cdot)$ is continuously differentiable, convex, and homogeneous of degree 2 on H. All we need is to verify that the function is strongly convex, with modulus β, with respect to $\|\cdot\|_H$:

$$\langle \widehat{\zeta}'(y) - \widehat{\zeta}'(y'), y - y' \rangle \geq \beta \|y - y'\|_H^2, \quad \text{for all } y, y'. \tag{A.6}$$

It is immediately seen that for a continuously differentiable function $\widehat{\zeta}$, the target property is local: to establish (A.6) it suffices to verify that every $\bar{y} \in H \backslash \{0\}$ admits a neighbourhood U such that the inequality in (A.6) holds for all $y, y' \in U$. Because of homogeneity, establishing the latter property for every $\bar{y} \in H \backslash \{0\}$ is the same as proving it when $\|\bar{y}\|_H = 1 - \epsilon$, for a fixed $\epsilon \in (0,1)$. Thus, let us fix an $\epsilon \in (0, 2/3)$ and $\bar{y} \in H$ with $\|\bar{y}\|_H = 1 - \epsilon$, and let $U = \{y' : \|y' - \bar{y}\|_H \leq \epsilon/2\}$, so that U is a compact set contained in the interior of Y_H. Now let $\phi(\cdot)$ be a non-negative C^∞ function on H which vanishes away from Y_H and has unit integral with respect to the Lebesgue measure on H. For $0 < r \leq 1$, let us set

$$\phi_r(y) = r^{-\dim H} \phi(y/r), \quad \zeta_r(\cdot) = (\zeta * \phi_r)(\cdot), \; \widehat{\zeta}_r(\cdot) = \zeta_r^{2/p}(\cdot),$$

where $*$ stands for convolution. As $r \to +0$, the non-negative C^∞ function $\zeta_r(\cdot)$ converges with the first-order derivative, uniformly on U, to $\zeta(\cdot)$, and $\widehat{\zeta}_r(\cdot)$ converges with the first-order derivative, uniformly on U, to $\widehat{\zeta}(\cdot)$. Further, when $r < \epsilon/2$, the function $\zeta_r(\cdot)$ is, along with $\zeta(\cdot)$, strongly convex, with modulus 1, with respect to $\|\cdot\|_H$, on U (since $\zeta_r'(\cdot)$ is the convolution

of $\zeta'(\cdot)$ and $\phi_r(\cdot)$). Since ζ_r is C^∞, we conclude that when $r < \epsilon/2$ we have

$$D^2\zeta_r(y)[h,h] \geq \|h\|_H^2, \quad \text{for all } y \in U,\ h \in H,$$

whence also

$$\begin{aligned} D^2\widehat{\zeta}_r(y)[h,h] &= \frac{2}{p}\left(\frac{2}{p}-1\right)\zeta_r^{2/p-2}(y)[D\zeta_r(y)[h]]^2 + \frac{2}{p}\zeta_r^{2/p-1}(y)D^2\zeta_r[h,h] \\ &\geq \frac{2}{p}\zeta_r^{2/p-1}(y)\|h\|_H^2, \quad \text{for all } h \in H,\ y \in U. \end{aligned}$$

It follows that when $r < \epsilon/2$ we have

$$\begin{gathered} \langle \widehat{\zeta}_r'(y) - \widehat{\zeta}_r'(y'), y - y' \rangle \geq \beta_r[U]\|y - y'\|_H^2, \quad \text{for all } y, y' \in U, \\ \text{where } \beta_r[U] = \frac{2}{p}\min_{u \in U}\zeta_r^{2/p-1}(u). \end{gathered}$$

As $r \to +0$, $\widehat{\zeta}_r'(\cdot)$ converges to $\widehat{\zeta}'(\cdot)$ uniformly on U, and $\beta_r[U]$ converges to $\beta[U] = \frac{2}{p}\min_{u\in U}\zeta^{2/p-1}(u) \geq (1-3\epsilon/2)^{2-p}\beta$ (recall the origin of β and U and note that ζ is absolutely homogeneous of degree p), we arrive at the relation

$$\langle \widehat{\zeta}'(y) - \widehat{\zeta}'(y'), y - y' \rangle \geq (1-3\epsilon/2)^{2-p}\beta\|y - y'\|_H^2, \quad \text{for all } y, y' \in U.$$

As we have already explained, the recently established validity of the latter relation for every $U = \{y : \|y - \bar{y}\|_H \leq \epsilon/2\}$ and every $\bar{y}$, $\|\bar{y}\|_H = 1 - \epsilon$ implies that $\widehat{\zeta}$ is strongly convex, with modulus $(1-3\epsilon/2)^{2-p}\beta$, on all of H. Since $\epsilon \in (0, 2/3)$ is arbitrary, Lemma A.2 is proved. □

Applying Lemma A.2 to E in the role of H, the ℓ_1/ℓ_2 norm $\|\cdot\|$ in the role of $\|\cdot\|_H$ and the function $w(\cdot)$ given by (A.5) in the role of ζ, we conclude that the function $\widehat{w}(\cdot) \equiv w^{2/p}(\cdot)$ is strongly convex, with modulus

$$\beta = \frac{2}{p}\Big[\min_y\{w(y) : y \in E, \|y\| = 1\}\Big]^{2/p-1},$$

with respect to $\|\cdot\|$, on the entire space E, whence the function $\beta^{-1}\widehat{w}(z)$ is continuously differentiable and strongly convex, with modulus 1, with respect to $\|\cdot\|$, on all of E. This observation, in view of the evident relation $\beta = \frac{2}{p}(n^{1-p}\gamma^{-1}p^{-1})^{2/p-1}$ and the origin of γ, p, immediately implies all the claims in Corollary 2.2. □

Proof of Theorem 2.3. Let $\mathbb{S}^n$ be the space of $n \times n$ symmetric matrices equipped with the Frobenius inner product. For $y \in \mathbb{S}^n$, let $\lambda(y)$ be the vector of eigenvalues of y (taken with their multiplicities in non-ascending order), and let $|y|_1 = \|\lambda(y)\|_1$ be the trace norm.

Lemma A.3. Let $N \geq M \geq 2$, and let F be a linear subspace in $\mathbb{S}^N$ such that every matrix $y \in F$ has at most M non-zero eigenvalues. Let $q = 1/(2\ln(M))$, so that $0 < q < 1$, and let $\widehat{\omega} : \mathbb{S}^N \to \mathbb{R}$ be given by

$$\widehat{\omega}(y) = \frac{4\sqrt{\mathrm{e}}\ln(M)}{1+q} \sum_{j=1}^{N} |\lambda_j(y)|^{1+q}.$$

The function $\widehat{\omega}(\cdot)$ is continuously differentiable, convex, and its restriction on the set $Y_F = \{y \in F : |y|_1 \leq 1\}$ is strongly convex, with modulus 1, with respect to $|\cdot|_1$.

Proof. **(1)** Let $0 < q < 1$. Consider the following function of $y \in \mathbb{S}^N$:

$$\chi(y) = \frac{1}{1+q} \sum_{j=1}^{N} |\lambda_j(y)|^{1+q} = \mathrm{Tr}(f(y)), \quad f(s) = \frac{1}{1+q}|s|^{1+q}. \tag{A.7}$$

(2) The function $f(s)$ is continuously differentiable on the axis and twice continuously differentiable away from the origin. Consequently, we can find a sequence of polynomials $f_k(s)$ converging to f, as $k \to \infty$, along with their first derivatives, uniformly on every compact subset of $\mathbb{R}$ and, further, converging to f uniformly, along with the first and second derivatives, on every compact subset of $\mathbb{R}\backslash\{0\}$. Now let $y, h \in \mathbb{S}^N$, let $y = u\,\mathrm{Diag}\{\lambda\}u^T$ be the eigenvalue decomposition of y, and let $h = u\widehat{h}u^T$. For a polynomial $p(s) = \sum_{k=0}^{K} p_k s^k$, setting $P(w) = \mathrm{Tr}(\sum_{k=0}^{K} p_k w^k)$, for $\omega \in \mathbb{S}^N$, and letting γ denote a closed contour in $\mathbb{C}$ enclosing the spectrum of y, we have

$$P(y) = \mathrm{Tr}(p(y)) = \sum_{j=1}^{N} p(\lambda_j(y)), \tag{A.8a}$$

$$DP(y)[h] = \sum_{k=1}^{K} k p_k \mathrm{Tr}(y^{k-1}h) = \mathrm{Tr}(p'(y)h) = \sum_{j=1}^{N} p'(\lambda_j(y))\widehat{h}_{jj}, \tag{A.8b}$$

$$\begin{aligned} D^2P(y)[h,h] &= \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} DP(y+th)[h] = \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \mathrm{Tr}(p'(y+th)h) \qquad \text{(A.8c)} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \frac{1}{2\pi\mathrm{i}} \oint_\gamma \mathrm{Tr}(h(zI-(y+th))^{-1})p'(z)\,\mathrm{d}z \\ &= \frac{1}{2\pi\mathrm{i}} \oint_\gamma \mathrm{Tr}(h(zI-y)^{-1}h(zI-y)^{-1})p'(z)\,\mathrm{d}z \\ &= \frac{1}{2\pi\mathrm{i}} \oint_\gamma \sum_{i,j=1}^{N} \widehat{h}_{ij}^2 \frac{p'(z)}{(z-\lambda_i(y))(z-\lambda_j(y))}\,\mathrm{d}z = \sum_{i,j=1}^{n} \widehat{h}_{ij}^2 \Gamma_{ij}, \end{aligned}$$

$$\text{where} \quad \Gamma_{ij} = \begin{cases} \frac{p'(\lambda_i(y))-p'(\lambda_j(y))}{\lambda_i(y)-\lambda_j(y)}, & \lambda_i(y) \neq \lambda_j(y), \\ p''(\lambda_i(y)), & \lambda_i(y) = \lambda_j(y). \end{cases}$$

We conclude from (A.8a) and (A.8b) that as $k \to \infty$, the real-valued polynomials $F_k(\cdot) = \mathrm{Tr}(f_k(\cdot))$ on $\mathbb{S}^N$ converge, along with their first-order derivatives, uniformly on every bounded subset of $\mathbb{S}^N$, and the limit of the sequence, by (A.8a), is exactly $\chi(\cdot)$. Thus, $\chi(\cdot)$ is continuously differentiable, and (A.8b) states that

$$D\chi(y)[h] = \sum_{j=1}^{N} f'(\lambda_j(y))\widehat{h}_{jj}. \tag{A.9}$$

Further, (A.8) states that if U is a closed convex set in $\mathbb{S}^N$ which does not contain singular matrices, then $F_k(\cdot)$ converges, as $k \to \infty$, uniformly on every compact subset of U, together with its first and second derivatives. Thus $\chi(\cdot)$ is twice continuously differentiable on U, and at every point $y \in U$ we have

$$D^2\chi(y)[h,h] = \sum_{i,j=1}^{N} \widehat{h}_{ij}^2 \Gamma_{ij}, \tag{A.10}$$

$$\text{where} \quad \Gamma_{ij} = \begin{cases} \frac{f'(\lambda_i(y)) - f'(\lambda_j(y))}{\lambda_i(y) - \lambda_j(y)}, & \lambda_i(y) \neq \lambda_j(y), \\ f''(\lambda_i(y)), & \lambda_i(y) = \lambda_j(y), \end{cases}$$

and in particular $\chi(\cdot)$ is convex on U.

(3) We intend to prove that (i) $\chi(\cdot)$ is convex, and (ii) its restriction to the set Y_F is strongly convex, with a certain modulus $\alpha > 0$, with respect to the trace norm $|\cdot|_1$. Since χ is continuously differentiable, all we need to prove (i) is to verify that

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \geq 0 \tag{A.11}$$

for a dense subset (y', y'') in $\mathbb{S}^N \times \mathbb{S}^N$, *e.g.*, those with non-singular $y' - y''$. If $y' - y''$ is non-singular, then the polynomial $q(t) = \mathrm{Det}(y' + t(y'' - y'))$, $t \in \mathbb{R}$, is not identically zero, and thus has finitely many roots in $[0,1]$. In other words, we can find finitely many points $t_0 = 0 < t_1 < \cdots < t_n = 1$ such that all 'matrix intervals' $\Delta_i = (y_i, y_{i+1})$, $y_k = y' + t_k(y'' - y')$, $1 \leq i \leq n-1$, contain only non-singular matrices. Therefore χ is convex on every closed segment contained in one of the Δ_i, and since χ is continuously differentiable, (A.11) follows.

(4) Now let us prove that with properly defined $\alpha > 0$ we have

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \geq \alpha |y' - y''|_1^2, \quad \text{for all } y', y'' \in Y_F. \tag{A.12}$$

Let $\epsilon > 0$, and let Y^ϵ be a convex and open in $Y = \{y : |y|_1 \leq 1\}$ subset of Y containing Y_F and such that, for all $y \in Y^\epsilon$, at most M eigenvalues of y are of magnitude $> \epsilon$. We intend to prove that for some $\alpha_\epsilon > 0$ we have

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \geq \alpha_\epsilon |y' - y''|_1^2, \quad \text{for all } y', y'' \in Y^\epsilon. \tag{A.13}$$

As above, it suffices to verify this relation for a dense set of pairs $y', y'' \in Y^\epsilon$ in $Y^\epsilon \times Y^\epsilon$, *e.g.*, those pairs $y', y'' \in Y^\epsilon$ for which $y' - y''$ is non-singular. Defining matrix intervals Δ_i as above and taking into account continuous differentiability of χ, it suffices to verify that, if $y \in \Delta_i$ and $h = y' - y''$, then $D^2\chi(y)[h,h] \geq \alpha_\epsilon |h|_1^2$. To this end, observe that by (A.10) all we have to prove is that

$$D^2\chi(y)[h,h] = \sum_{i,j=1}^N \widehat{h}_{ij}^2 \Gamma_{ij} \geq \alpha_\epsilon |h|_1^2. \tag{A.14}$$

(5) Setting $\lambda_j = \lambda_j(y)$, observe that $\lambda_i \neq 0$ for all i due to the origin of y. We claim that if $|\lambda_i| \geq |\lambda_j|$, then $\Gamma_{ij} \geq q|\lambda_i|^{q-1}$. Indeed, the latter relation definitely holds true when $\lambda_i = \lambda_j$. Now, if λ_i and λ_j are of the same sign, then

$$\Gamma_{ij} = \frac{|\lambda_i|^q - |\lambda|_j^q}{|\lambda_i| - |\lambda_j|} \geq q|\lambda_i|^{q-1},$$

since the derivative of the concave function t^q of $t > 0$ is positive and non-increasing (recall that $0 < q \leq 1$). If λ_i and λ_j are of different signs, then

$$\Gamma_{ij} = \frac{|\lambda_i|^q + |\lambda_j|^q}{|\lambda_i| + |\lambda_j|} \geq |\lambda_i|^{q-1}$$

due to $|\lambda_j|^q \geq |\lambda_j||\lambda_i|^{q-1}$, and therefore $\Gamma_{ij} \geq q|\lambda_i|^{q-1}$. Thus, our claim is justified.

Without loss of generality we can assume that the positive reals $\mu_i = |\lambda_i|$, $i = 1, \ldots, N$, form a non-decreasing sequence, so that, by the above, $\Gamma_{ij} \geq q\mu_j^{q-1}$ when $i \leq j$. Further, at most M of μ_j are $\geq \epsilon$, since $y', y'' \in Y^\epsilon$ and therefore $y \in Y^\epsilon$ by convexity of Y^ϵ. Hence

$$D^2\chi(y)[h,h] \geq 2q \sum_{i<j\leq N} \widehat{h}_{ij}^2 \mu_j^{q-1} + q \sum_{j=1}^N \widehat{h}_{jj}^2 \mu_j^{q-1},$$

or equivalently, by symmetry of $\widehat{h}$, if

$$h^j = \begin{bmatrix} & & & \widehat{h}_{1j} \\ & & & \widehat{h}_{2j} \\ & & & \vdots \\ \widehat{h}_{j1} & \widehat{h}_{j2} & \cdots & \widehat{h}_{jj} \end{bmatrix}$$

and $H_j = \|h^j\|_{\mathrm{Fro}}$ is the Frobenius norm of h^j, then

$$D^2\chi(y)[h,h] \geq q \sum_{j=1}^N H_j^2 \mu_j^{q-1}. \tag{A.15}$$

Observe that the h^j are of rank ≤ 2, whence $|h^j|_1 \le \sqrt{2}\|h^j\|_{\rm Fro} = \sqrt{2}H_j$, so that

$$\begin{aligned}
|h|_1^2 = |\widehat{h}|_1^2 &\le \left(\sum_{j=1}^N |h^j|_1\right)^2 \le 2\left(\sum_{j=1}^N H_j\right)^2 = 2\left(\sum_j [H_j\mu_j^{(q-1)/2}]\mu_j^{(1-q)/2}\right)^2 \\
&\le 2\left(\sum_{j=1}^N H_j^2\mu_j^{q-1}\right)\left(\sum_{j=1}^N \mu_j^{1-q}\right) \\
&\le (2/q)D^2\chi(y)[h,h]\left(\sum_{j=1}^N \mu_j^{1-q}\right) \quad \text{(by (A.15))} \\
&\le (2/q)D^2\chi(y)[h,h]\left((N-M)\epsilon^{1-q} + [M^{-1}\sum_{j=N-M+1}^N \mu_j]^{1-q}M\right) \\
&\qquad \text{(due to } 0<q<1 \text{ and } \mu_j \le \epsilon,\ j \le N-M) \\
&\le (2/q)D^2\chi(y)[h,h]\big((N-M)\epsilon^{1-q} + M^q\big) \\
&\qquad \text{(since } \textstyle\sum_j \mu_j \le 1 \text{ due to } y \in Y^\epsilon \subset \{y : |y|_1 \le 1\}),
\end{aligned}$$

and we see that (A.14) holds with $\alpha_\epsilon = q/(2[(N-M)\epsilon + M^q])$. As a result, with the α_ϵ just defined, relation (A.13) holds true, whence (A.12) is satisfied with $\alpha = \lim_{\epsilon\to+0}\alpha_\epsilon = \frac{q}{2}M^{-q}$, that is, χ is a continuously differentiable convex function which is strongly convex, with modulus α, with respect to $|\cdot|_1$, on Y_F. Recalling the definition of q, we see that $\widehat{\omega}(\cdot) = \alpha^{-1}\chi(\cdot)$, so that $\widehat{\omega}$ satisfies the conclusion of Lemma A.3. □

We are now ready to prove Theorem 2.3. Under the premises and in the notation of the theorem, let

$$N = \mu + \nu, \quad M = 2\mu, \quad \mathcal{A}z = \frac{1}{2}\begin{bmatrix} & z \\ z^T & \end{bmatrix} \in \mathbb{S}^N, \quad \text{for } z \in \mathbb{R}^{\mu\times\nu}. \quad \text{(A.16)}$$

Observe that the image space F of $\mathcal{A}$ is a linear subspace of $\mathbb{S}^N$, and that the eigenvalues of $y = \mathcal{A}z$ are the 2μ reals $\pm\sigma_i(z)/2$, $1 \le i \le \mu$, and $N-M$ zeros, so that $\|z\|_{\rm nuc} \equiv |\mathcal{A}z|_1$ and M, F satisfy the premises of Lemma A.3. Setting

$$\omega(z) = \widehat{\omega}(\mathcal{A}z) = \frac{4\sqrt{\rm e}\ln(2\mu)}{2^q(1+q)}\sum_{i=1}^\mu \sigma_i^{1+q}(z), \quad q = \frac{1}{2\ln(2\mu)},$$

and invoking Lemma A.3, we see that ω is a convex continuously differentiable function on $\mathbb{R}^{\mu\times\nu}$ which, due to the identity $\|z\|_{\rm nuc} \equiv |\mathcal{A}z|_1$, is strongly convex, with modulus 1, with respect to $\|\cdot\|_{\rm nuc}$, on the $\|\cdot\|_{\rm nuc}$-unit ball Z. This function clearly satisfies (2.10). □

Proof of Corollary 2.4. This corollary is obtained from Theorem 2.3 in exactly the same way as Corollary 2.2 was derived from Theorem 2.1. □

A.3. Proofs for Section 2.2.2

Proof of Lemma 2.7. Let $F(w) = f(w) - L\omega(w)$. Lemma A.1 as applied to $\phi \equiv 0$, $\Psi \equiv L^{-1}F$, implies that $\bar{w} \in Z^o$ and that there exists $F' \in \partial F(\bar{w})$ such that $\langle F' + L\,\omega'(\bar{w}), w - \bar{w}\rangle \geq 0$ for all $w \in Z$. Therefore,

$$\begin{aligned} f(w) &= F(w) + L\,\omega(w) \geq F(\bar{w}) + \langle F', w - \bar{w}\rangle + L\,\omega(w) \\ &\geq F(\bar{w}) - L\,\langle \omega'(\bar{w}), w - \bar{w}\rangle + L\omega(w) = F(\bar{w}) + L\,\omega(\bar{w}) + L\,V_{\bar{w}}(w) \\ &= f(\bar{w}) + L\,V_{\bar{w}}(w). \end{aligned}$$ □

Proof of Theorem 2.9. Let

$$\ell_t(x) \stackrel{\text{def}}{=} f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + L_f V_{x_t}(x) + \Psi(x) \stackrel{(2.18)}{\geq} \phi(x).$$

Thus, $\ell_t(\cdot)$ satisfies the premises of Lemma 2.7 (note the definition of $V_{x_t}(x)$). Since x_{t+1} is the minimizer of this expression over $x \in Z$, Lemma 2.7 states that $x_{t+1} \in Z^o$ (so that the algorithm is well defined), and for every $w \in Z$ we have

$$\ell_t(x_{t+1}) + L_f V_{x_{t+1}}(w) \leq \ell_t(w).$$

Substituting $w = x_t$, we obtain

$$\phi(x_t) = \ell_t(x_t) \geq \ell_t(x_{t+1}) + L_f V_{x_{t+1}}(x_t) \geq \phi(x_{t+1}).$$

Thus, $\phi(x_{t+1}) \leq \phi(x_t)$, as claimed. Substituting $w = x^*$, we obtain

$$\ell_t(x_{t+1}) + L_f V_{x_{t+1}}(x^*) \leq \ell_t(x^*) \leq \phi(x^*) + L_f V_{x_t}(x^*),$$

whence $\phi(x_{t+1}) + L_f V_{x_{t+1}}(x^*) \leq \phi(x^*) + L_f V_{x_t}(x^*)$. Summing up these inequalities over $t = 0, \ldots, T-1$, we find

$$\sum_{t=1}^{T} (\phi(x_t) - \phi(x^*)) \leq L_f V_{x_0}(x^*) - L_f V_{x_T}(x^*) \leq L_f V_{x_0}(x^*),$$

which proves (2.20). The remaining claims of Theorem 2.9 are immediate corollaries of (2.20) due to the monotonicity of the algorithm, the inclusions $x_t \in Z$ and the convexity of ϕ. □

Proof of Theorem 2.11. Define the functions

$$\psi_t(x) = \sum_{i=0}^{t} [f(x_i) + \langle \nabla f(x_i), x - x_i\rangle] + (t+1)\Psi(x) + L_f V_{z_\omega}(x).$$

Let $\psi_t^* = \min_{w \in Z} \psi_t(w)$. Note that $\psi_t^* = \psi_t(x_{t+1})$ by construction of the algorithm, and that $x_t, y_t \in Z^o$ for all t by Lemma A.1.

Taking into account that $\psi_t(\cdot) - L_f\omega(\cdot)$ is convex, and that $x_{t+1} \in \arg\min_Z \psi_t$, by Lemma 2.7 we have

$$\psi_t^* + L_f V_{x_{t+1}}(x) \le \psi_t(x) \le (t+1)\phi(x) + L_f V_{z_\omega}(x), \quad \text{for } x \in Z. \quad \text{(A.17)}$$

Let us now prove that

$$\psi_t^* \ge \sum_{i=0}^{t} \phi(y_i). \quad \text{(A.18)}$$

For $t = 0$, we have $x_0 = z_\omega$. Therefore,

$$\begin{aligned}
\psi_0^* &= \min_{w\in Z}[f(x_0) + \langle \nabla f(x_0), w - x_0\rangle + \Psi(x) + L_f V_{x_0}(w)] \\
&= f(x_0) + \langle \nabla f(x_0), y_0 - x_0\rangle + \Psi(y_0) + L_f V_{x_0}(y_0) \\
&\overset{(2.18)}{\ge} \phi(y_0).
\end{aligned}$$

Assume now that (A.18) is true for some $t \ge 0$. Then

$$\begin{aligned}
\psi_{t+1}^* &= \min_{w\in Z}\{\psi_t(w) + f(x_{t+1}) + \langle \nabla f(x_{t+1}), w - x_{t+1}\rangle + \Psi(w)\} \\
&\overset{(A.17)}{\ge} \min_{w\in Z}\{\psi_t^* + L_f V_{x_{t+1}}(w) + f(x_{t+1}) + \langle \nabla f(x_{t+1}), w - x_{t+1}\rangle + \Psi(w)\} \\
&\overset{(A.18)}{\ge} \sum_{i=0}^{t} \phi(y_i) + \min_{w\in Z}\{L_f V_{x_{t+1}}(w) + f(x_{t+1}) \\
&\qquad\qquad + \langle \nabla f(x_{t+1}), w - x_{t+1}\rangle + \Psi(w)\} \\
&= \sum_{i=0}^{t} \phi(y_i) + f(x_{t+1}) + \langle \nabla f(x_{t+1}), y_{t+1} - x_{t+1}\rangle \\
&\qquad\qquad + L_f V_{x_{t+1}}(y_{t+1}) + \Psi(y_{t+1}) \\
&\overset{(2.18)}{\ge} \sum_{i=0}^{t+1} \phi(y_i).
\end{aligned}$$

Thus, we prove (A.18) for all $t \ge 0$.

Combining (A.18) and (A.17) and setting $x = x^*$, we arrive at (2.22). The remaining claims of Theorem 2.11 are immediate corollaries of (2.22) due to the inclusions $y_t \in Z$ and the convexity of ϕ. □

Proof of Theorem 2.14. Observe that by construction

$$\psi_t(\cdot) = L_f\omega(\cdot) + \ell_t(\cdot) + \alpha_t\Psi(\cdot),$$

with non-negative α_t and affine $\ell_t(\cdot)$, whence the algorithm is well defined and maintains the inclusions $z_t \in Z^o$, $x_t, y_t \in Z$. Further, let us introduce

a convenient notation for the coefficients of Algorithm 2.13. Namely, let

$$a_{t+1} = \frac{t+2}{2}, \quad A_t = \sum_{i=1}^{t} a_i = \frac{t(t+3)}{4}, \quad t \geq 0.$$

Then $A_0 = 0$ and $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. It is important that

$$\tau_t^2 A_{t+1} \leq 1, \quad t \geq 0. \tag{A.19}$$

Let $\psi_t^* = \psi_t(z_t)$. Let us prove by induction that

$$\psi_t^* \geq A_t \phi(y_t), \quad t \geq 0. \tag{A.20}$$

For $t = 0$ this inequality is valid. Suppose that it is true for some $t \geq 0$. Let us prove that it is valid for index $t+1$.

Taking into account the structure of $\psi_t(\cdot)$ just outlined, and applying Lemma A.1, for all $w \in Z$ we have

$$\begin{aligned}\psi_t(w) &\overset{(2.19)}{\geq} \psi_t^* + L_f V_{z_t}(w) \overset{(A.20)}{\geq} A_t\phi(y_t) + L_f V_{z_t}(w) \\ &\geq A_t\big[f(x_{t+1}) + \langle \nabla f(x_{t+1}), y_t - x_{t+1}\rangle + \Psi(y_t)\big] + L_f V_{z_t}(w).\end{aligned}$$

Therefore, for all $w \in Z$,

$$\begin{aligned}\psi_{t+1}(w) \geq\ & L_f V_{z_t}(w) + A_t\big[f(x_{t+1}) + \langle \nabla f(x_{t+1}), y_t - x_{t+1}\rangle + \Psi(y_t)\big] \\ &+ a_{t+1}\big[f(x_{t+1}) + \langle \nabla f(x_{t+1}), w - x_{t+1}\rangle + \Psi(w)\big].\end{aligned}$$

Since $A_t(y_t - x_{t+1}) - a_{t+1}x_{t+1} \overset{\text{Alg. 2.13(b)}}{=} -a_{t+1}z_t$, we conclude that

$$\begin{aligned}\psi_{t+1}(w) \geq\ & L_f V_{z_t}(w) + A_{t+1}f(x_{t+1}) \\ &+ A_t\Psi(y_t) + a_{t+1}[\langle \nabla f(x_{t+1}), w - z_t\rangle + \Psi(w)]\end{aligned}$$

for all $w \in Z$. Thus,

$$\begin{aligned}\psi_{t+1}^* \quad \geq\quad & \min_{w\in Z}\big\{L_f V_{z_t}(w) + A_{t+1}f(x_{t+1}) + A_t\Psi(y_t) \\ &\qquad + a_{t+1}[\langle \nabla f(x_{t+1}), w - z_t\rangle + \Psi(w)]\big\} \\ \overset{\text{Alg. 2.13(c)}}{=}\ & L_f V_{z_t}(\hat{x}_{t+1}) + A_{t+1}f(x_{t+1}) + A_t\Psi(y_t) \\ &\qquad + a_{t+1}[\langle \nabla f(x_{t+1}), \hat{x}_{t+1} - z_t\rangle + \Psi(\hat{x}_{t+1})] \\ \overset{\text{Alg. 2.13(d)}}{\geq}\ & \frac{L_f}{2}\|\hat{x}_{t+1} - z_t\|_E^2 + A_{t+1}f(x_{t+1}) \\ &\qquad + a_{t+1}\langle \nabla f(x_{t+1}), \hat{x}_{t+1} - z_t\rangle + A_{t+1}\Psi(y_{t+1}).\end{aligned}$$

It remains to note that $\hat{x}_{t+1} - z_t \overset{\text{Alg. 2.13(b,d)}}{=} (y_{t+1} - x_{t+1})/\tau_t$. Therefore,

$$\begin{aligned}
\psi^*_{t+1} &\geq \frac{L_f}{2\tau_t^2}\|y_{t+1} - x_{t+1}\|_E^2 + A_{t+1}f(x_{t+1}) \\
&\qquad + A_{t+1}\langle \nabla f(x_{t+1}), y_{t+1} - x_{t+1}\rangle + A_{t+1}\Psi(y_{t+1}) \\
&\overset{(A.19)}{\geq} A_{t+1}\left[\frac{L_f}{2}\|y_{t+1} - x_{t+1}\|_E^2 + f(x_{t+1})\right. \\
&\qquad \left. + \langle \nabla f(x_{t+1}), y_{t+1} - x_{t+1}\rangle + \Psi(y_{t+1})\right] \\
&\overset{(2.17)}{\geq} A_{t+1}\phi(y_{t+1}).
\end{aligned}$$

Thus, we have proved inequality (A.20) for all $t \geq 0$. Since $\psi^*_t \leq A_t\phi(x^*) + L_f V_{z_\omega}(x^*)$, we obtain the estimate (2.25). □

A.4. Proof of Lemma 3.1

Verifying (3.9). Let $\bar{s} \geq 1$ be such that termination does not occur at stage $\bar{s}$. For $1 \leq s \leq \bar{s}$, let t_s be the last iteration of stage s, and let $\Phi_s = \Phi^{t_s,+}$, $\phi_s = \Phi^{t_s,-}$, so that $\phi_s \geq \Phi_s/\theta > 0$. By construction of $\Phi^{t,-}$, we have $\phi_s = \Phi^-_\tau(\rho_s)$ for some $\tau = \tau(s) \leq t_s$. Taking into account that $\Phi^-_\tau(\cdot)$ underestimates $\Phi(\cdot)$, $\Phi(\rho_*) = 0$, $\rho_s > \rho_*$ and $\Phi^-_\tau(\rho_s) > 0$ and setting $p_s = \frac{\mathrm{d}}{\mathrm{d}\rho}\Phi^-_\tau(\cdot)$, we get $p_s > 0$, and the root $r_\tau = \rho_s - \Phi^-_\tau(\rho_s)/p_s = \rho_s - \phi_s/p_s$ of $\Phi^-_\tau(\cdot)$ is a point of R_{t_s}, whence $\rho_{s+1} \leq r_\tau$. Further, assuming $s > 1$, we have $\phi_s + p_s(\rho_{s-1} - \rho_s) = \Phi^-_\tau(\rho_{s-1}) \leq \Phi(\rho_{s-1}) \leq \Phi_{s-1}$. The bottom line is that, for $1 \leq s \leq \bar{s}$, we have

$$\begin{aligned}
&0 < \phi_s \leq \Phi(\rho_s) \leq \Phi_s, && (a_s)\\
&\Phi_s/\phi_s \leq \theta, && (b_s)\\
&\rho_s - \rho_{s+1} \geq \phi_s/p_s, && (c_s)\\
&\Phi_{s-1} \geq \phi_s + p_s(\rho_{s-1} - \rho_s) \quad \text{when } s > 1. && (d_s)
\end{aligned} \tag{A.21}$$

As a result, for $1 \leq s < \bar{s}$ and using the notation of (A.21) henceforth,

$$\Phi_s - \phi_{s+1} \underset{(d_{s+1})}{\geq} p_{s+1}(\rho_s - \rho_{s+1}) \underset{(c_s)}{\geq} p_{s+1}\phi_s/p_s \underset{(b_s)}{\geq} \theta^{-1}p_{s+1}\Phi_s/p_s,$$

which taken together with (a_{s+1}) implies

$$\Phi_s - \theta^{-1}\Phi_{s+1} \geq \theta^{-1}p_{s+1}\Phi_s/p_s,$$

and hence

$$\Phi_s \geq \theta^{-1}\frac{p_{s+1}\Phi_s + p_s\Phi_{s+1}}{p_s} \geq 2\theta^{-1}\sqrt{p_{s+1}\Phi_s p_s\Phi_{s+1}}/p_s,$$

and finally

$$\Phi_{s+1}p_{s+1} \leq [\theta/2]^2\Phi_s p_s.$$

Thus,

$$\Phi_{\bar{s}} p_{\bar{s}} \le [\theta/2]^{2(\bar{s}-1)} \Phi_1 p_1. \tag{A.22}$$

Now $\Phi(\cdot)$ is Lipschitz-continuous with constant $2\pi(b)$, $\Phi_1 \le \theta\phi_1 \le \theta\Phi(\rho_1) \le 2\theta\pi(b)(\rho_1 - \rho_*)$ (since $\Phi(\rho_*) = 0$), whence $\Phi_1 \le 2\theta\pi(b)\rho_1$. Further, by construction $p_1 > 0$ is the slope of an affine function which underestimates $\Phi(\cdot)$ on the ray $\rho \ge 0$, whence $p_1 \le \lim_{\rho\to\infty} \Phi'(\rho) \le \pi(b) - \delta \le \pi(b)$, and thus $\Phi_1 p_1 \le 2\theta\pi^2(b)\rho_1$. Further, invoking $(c_{\bar{s}})$ and $(b_{\bar{s}})$ we have $p_{\bar{s}}(\rho_{\bar{s}} - \rho_{\bar{s}+1}) \ge \phi_{\bar{s}} \ge \theta^{-1}\Phi_{\bar{s}}$, so that $p_{\bar{s}} \ge \theta^{-1}\Phi_{\bar{s}}/(\rho_{\bar{s}} - \rho_{\bar{s}+1}) \ge \theta^{-1}\Phi_{\bar{s}}/\rho_1$, and thus $\Phi_{\bar{s}} p_{\bar{s}} \ge \theta^{-1}\Phi_{\bar{s}}^2/\rho_1$. We see that (A.22) implies the bound

$$\Phi_{\bar{s}}^2 \le 2\theta^2 [\theta/2]^{2(\bar{s}-1)} \pi^2(b) \rho_1^2.$$

On the other hand, our root finding does not terminate at stage $\bar{s}$, implying, due to the termination rule, that $\Phi_{\bar{s}} > \epsilon\rho_{\bar{s}} \ge \epsilon\rho_*$. We arrive at the bound

$$[2/\theta]^{\bar{s}} \le 2\sqrt{2}\pi(b)\rho_1/(\epsilon\rho_*).$$

Taking into account that $2/\theta > 1$ and $\bar{s}$ is the number of an arbitrary non-terminal stage, the total number of stages admits the bound (3.9).

Verifying (3.11). If t is the number of a non-terminal iteration of stage s, then by construction $\Phi^{t,+} > \epsilon\rho_s$ and $\Phi^{t,-} < \theta^{-1}\Phi^{t,+}$, implying that

$$\Phi^{t,+} - \Phi^{t,-} > \frac{\theta - 1}{\theta} \epsilon\rho_s \ge \frac{\theta - 1}{\theta} \epsilon\rho_*.$$

As we have seen when describing our root-finding routine, the left-hand side of this inequality does not exceed $\mathrm{DualityGap}^s(z^t)$, and we see that for our t we have

$$[\varepsilon_t \ge]\mathrm{DualityGap}^s(z^t) > \frac{\theta - 1}{\theta} \epsilon\rho_*,$$

which implies the bound (3.11). □

A.5. Proof of Theorem 3.3

We shall apply (2.2), first setting (a) $z = z_\tau$, $\xi = \gamma_\tau F(w_\tau)$ (so that $z_+ = z_{\tau+1}$), and then (b) $z = z_\tau, \xi = \gamma_\tau F(z_\tau)$ (so that $z_+ = w_\tau$). For (a), we obtain, for all $w \in Z$,

$$\gamma_\tau \langle F(w_\tau), z_{\tau+1} - w \rangle \le V_{z_\tau}(w) - V_{z_{\tau+1}}(w) - V_{z_\tau}(z_{\tau+1}),$$

and hence

$$\begin{aligned} &\gamma_\tau \langle F(w_\tau), w_\tau - w \rangle \\ &\quad \le V_{z_\tau}(w) - V_{z_{\tau+1}}(w) + \underbrace{[\gamma_\tau \langle F(w_\tau), w_\tau - z_{\tau+1} \rangle - V_{z_\tau}(z_{\tau+1})]}_{=\delta_\tau}, \end{aligned}$$

For (b), we find

$$\gamma_\tau\langle F(z_\tau), w_\tau - z_{\tau+1}\rangle \le V_{z_\tau}(z_{\tau+1}) - V_{w_\tau}(z_{\tau+1}) - V_{z_\tau}(w_\tau) \tag{A.23}$$

and so

$$\begin{aligned}
\delta_\tau &= \gamma_\tau\langle F(w_\tau) - F(z_\tau), w_\tau - z_{\tau+1}\rangle - V_{z_\tau}(z_{\tau+1}) + \gamma_\tau\langle F(z_\tau), w_\tau - z_{\tau+1}\rangle \\
&\le \gamma_\tau\langle F(w_\tau) - F(z_\tau), w_\tau - z_{\tau+1}\rangle - V_{w_\tau}(z_{\tau+1}) - V_{z_\tau}(w_\tau) \quad \text{(by (A.23))} \\
&\le \gamma_\tau\|F(w_\tau) - F(z_\tau)\|_{E,*}\|w_\tau - z_{\tau+1}\|_E - \frac{1}{2}\|w_\tau - z_{\tau+1}\|_E^2 - \frac{1}{2}\|w_\tau - z_\tau\|_E^2 \\
&\le \frac{\gamma_\tau^2}{2}\|F(w_\tau) - F(z_\tau)\|_{E,*}^2 - \frac{1}{2}\|w_\tau - z_\tau\|_E^2 \\
&\le \frac{1}{2}[\gamma_\tau^2 L^2 - 1]\|w_\tau - z_\tau\|_E^2 \quad \text{(by (3.16))}.
\end{aligned}$$

We see that, first, $\gamma_\tau \le 1/L$ indeed ensures that $\delta_\tau \le 0$, and, second, whenever the latter relation holds for all τ we have

$$\gamma_\tau\langle F(w_\tau), w_\tau - w\rangle \le V_{z_\tau}(w) - V_{z_{\tau+1}}(w), \quad \text{for all } \tau, w \in Z.$$

Summing up the latter inequalities over $\tau = 1, \dots, t$ and taking into account that $V_{z_1}(w) = V_{z_\omega}(w) \le \frac{1}{2}\Omega^2[Z']$ for $w \in Z'$, while $V_{z_{t+1}}(w) \ge 0$, we obtain

$$\max_{w\in Z'} \sum_{\tau=1}^t \lambda_\tau^t \langle F(w_\tau), w_\tau - w\rangle \le \frac{\frac{1}{2}\Omega^2[Z']}{\sum_{\tau=1}^t \gamma_\tau}, \quad \lambda_\tau^t = \frac{\gamma_\tau}{\sum_{s=1}^t \gamma_s}, \tag{A.24}$$

for all t. Denoting $w_\tau = [u_\tau; v_\tau]$, $z^t = [u^t; v^t]$, we have

$$\begin{aligned}
&\max_{w=[u;v]\in Z'} \sum_{\tau=1}^t \lambda_\tau^t \langle F(w_\tau), w_\tau - w\rangle \\
&\quad = \max_{[u;v]\in Z'} \sum_{\tau=1}^t \lambda_\tau^t \big[\langle \nabla_u \phi(u_\tau, v_\tau), u_\tau - u\rangle + \langle \nabla_v \phi(u_\tau, v_\tau), v - v_\tau\rangle\big] \\
&\quad \ge \max_{[u;v]\in Z'} \sum_{\tau=1}^t \lambda_\tau^t \big[[\phi(u_\tau, v_\tau) - \phi(u, v_\tau)] + [\phi(u_\tau, v) - \phi(u_\tau, v_\tau)]\big] \\
&\qquad\qquad \text{(since } \phi \text{ is convex–concave)} \\
&\quad = \max_{[u;v]\in Z'} \sum_{\tau=1}^t \lambda_\tau^t \big[\phi(u_\tau, v) - \phi(u, v_\tau)\big] \\
&\quad \ge \max_{[u;v]\in Z'} \big[\phi(u^t, v) - \phi(u, v^t)\big] \\
&\qquad\qquad \text{(since } \phi \text{ is convex–concave and by definition of } z^t).
\end{aligned}$$

The resulting inequality combines with (A.24) to imply (3.18). □

A.6. Good proximal set-ups

Our goal here is to understand to what extent the set-ups introduced in Section 2.1.1 are the 'best' ones for the applications considered in this paper. Let Z be a closed and bounded convex subset of a Euclidean space E which affinely spans E, so that the set $Z^+ = \frac{1}{2}(Z - Z)$ is the unit ball of some norm $\|\cdot\|_Z$ on E uniquely defined by Z. Also, let $\|\cdot\|_{Z,*}$ be the norm conjugate to $\|\cdot\|_Z$. We need the following simple observation.

Lemma A.4. With Z and E as above, let $\|\cdot\|_E, \omega(\cdot)$ be a proximal set-up for Z, E, let $F(x) = Ax+b : E \to E$ be an affine vector field, and let $L_E(F)$, $L_Z(F)$ be the Lipschitz constants of $F\big|_Z$ with respect to $(\|\cdot\|_E, \|\cdot\|_{E,*})$ and with respect to $(\|\cdot\|_Z, \|\cdot\|_{Z,*})$, respectively:

$$L_E(F) = \sup_{x,y\in Z, x\neq y} \frac{\|F(x) - F(y)\|_{E,*}}{\|x - y\|_E},$$
$$L_Z(F) = \sup_{x,y\in Z, x\neq y} \frac{\|F(x) - F(y)\|_{Z,*}}{\|x - y\|_Z}.$$

Then

$$L_E(F)\Omega^2[Z, \omega(\cdot)] \geq L_Z(F),$$

where $\Omega[Z, \omega(\cdot)]$ is the $\omega(\cdot)$-radius of Z.

Proof. Since Z affinely spans E and $F(x) = Ax + b$ is affine, we clearly have for all $x, y, u, v \in Z$:

$$\langle A(x - y), u - v\rangle \leq L_E(F)\|x - y\|_E\|u - v\|_E \leq 4L_E(F)\Omega^2[Z, \omega(\cdot)], \quad \text{(A.25)}$$

where the concluding inequality is given by (2.4) with $Z' = Z$. On the other hand, as x, y, u, v run through Z, $(x - y, u - v)$ can be made equal to any desired pair (g, h) with $g, h \in 2Z^+$. Consequently, the maximum of the left-hand side of (A.25) over $x, y, u, v \in Z$ is $4\max_{g\in E:\|g\|_Z\leq 1} \|Ag\|_{Z,*} \geq 4L_Z(F)$, and we obtain $4L_Z(F) \leq 4L_E(F)\Omega^2[Z, \omega(\cdot)]$, as claimed. □

Now assume that we are given a proximal set-up $\|\cdot\|_E, \omega(\cdot)$ for Z, E such that $\omega'(z_w) = 0$,[25] and

- either we are using the set-up in question to solve a problem (2.16) with the domain Z and a convex quadratic $f(x)$ by one of the methods from Section 2.2.2 (with the primal gradient method initialized by $z_0 = z_\omega$)
- or we are using the set-up to solve a *bilinear* saddle-point problem (3.13) with the domain $Z = U \times V$ by the MP algorithm from Section 3.2.

[25] This is without loss of generality, since we lose nothing (in fact we even gain by reducing the ω-radius of Z) when passing from the original DGF $\omega(\cdot)$ to the DGF $\omega(z) - \langle\omega'(z_\omega), z - z_\omega\rangle$.

In the first case, the guaranteed efficiency estimate of the algorithm is

$$\epsilon_t \le O(1)\Omega^2[Z,\omega(\cdot)]L_E(F)t^{-\kappa}, \tag{A.26}$$

where ϵ_t is the error after t steps, in terms of the objective of (2.16). Here $F(z) \equiv \nabla f(z)$, $\kappa = 1$ for the primal and dual gradient methods, and $\kappa = 2$ for the fast gradient method.[26] In the second case, invoking (3.19), the efficiency estimate is again (A.26), now with $\epsilon_t = \mathrm{DualityGap}(z^t)$, $\kappa = 1$ and the vector field (3.14) in the role of F. Note that this field is affine, due to the assumed bilinearity of ϕ.

Now assume that we have at our disposal an alternative proximal set-up $\|\cdot\|, \omega_Z(\cdot)$ for Z, E, with $\|\cdot\|$ θ-compatible with $\|\cdot\|_Z$:

$$\theta^{-1}\|z\| \le \|z\|_Z \le \theta\|z\|, \quad \text{for all } z.$$

Let us say that this alternative set-up is *good* if both θ and the quantity $\Omega_Z = \Omega[Z,\omega_Z(\cdot)]$ are 'moderate', *i.e.*, absolute constants, as is the case for the Euclidean set-up and the unit Euclidean ball Z ($\theta = \Omega_Z = 1$), or are universal functions of $\dim Z$ which grow with $\dim Z$ *logarithmically*, as is the case for the set-ups from Section 2.1.1, where $\theta = 1$ and $\Omega_Z \le O(1)\sqrt{\ln(\dim Z)}$. What happens when we pass from the proximal set-up $\|\cdot\|_E$, $\omega(\cdot)$ to this alternative set-up? The answer is clear: since $\|\cdot\|$ is within a factor θ of $\|\cdot\|_Z$, the Lipschitz constant L of F with respect to $\|\cdot\|, \|\cdot\|_*$ does not exceed $\theta^2 L_Z(F)$, and the right-hand side $O(1)\Omega_Z L t^{-\kappa}$ in the efficiency estimate associated with our alternative set-up will be $\le O(1)(\theta\Omega_Z)^2 L_Z(F)t^{-\kappa}$. That is, it will be at worst a factor $O(1)\theta^2\Omega_Z^2$ larger than the right-hand side of (A.26): see Lemma A.4. Since $\theta\Omega_Z$ is moderate, passing from the original set-up to the alternative one cannot 'spoil' the efficiency estimate by more than a factor of 'nearly $O(1)$'. In other words, as far as our algorithms are concerned, a good proximal set-up for the domain of interest is nearly optimal in terms of the efficiency estimate, with the given formal interpretation of the word 'nearly'.

The analysis above justifies the proximal set-ups used to handle the ℓ_1/ℓ_2 and nuclear norm Dantzig selector problems (1.1b). Indeed, using the approach of Section 3 we obtain bilinear saddle-point problems on bounded domains Z; the set-ups we use are good for these domains (and in fact result in the smallest values of $\theta\Omega_Z$ known to us, up to a factor of $O(1)$), and are thus nearly optimal within the framework of our algorithmic scheme.

The situation with the ℓ_1/ℓ_2 and nuclear norm lasso problems (1.1a), solved via composite minimization, is less clear. When posed in the form (2.16), these problems 'have trivial domain' $Z = E$, which makes the above

[26] To see that (A.26) is indeed the case, look at (2.21), (2.23) and (2.25), and note that x^* can be an arbitrary point of Z, while $\max_{x^*\in Z} V_{z_\omega}(x^*) = \frac{1}{2}\Omega^2[Z,\omega(\cdot)]$ due to $\omega'(z_w) = 0$.

analysis inapplicable and indeed makes the choice of proximal set-up a difficult task. Basically all gradient methods for lasso-type problems proposed in the literature, *e.g.*, the very popular algorithm FISTA of Beck and Teboulle (2009*a*), operate with the standard Euclidean set-up $\|z\| = \|z\|_2 := \sqrt{\langle z, z\rangle}$, $\omega(z) = \frac{1}{2}\langle z, z\rangle$. We intend to show, however, that the 'non-Euclidean' set-ups used in Section 2, while not being 'provably nearly optimal', in some cases have significant potential advantages compared to the Euclidean set-up. For the sake of definiteness, let us focus on the ℓ_1 lasso problem (1.1a) with $m \times n$ matrix A (the situations with the ℓ_1/ℓ_2 and nuclear norms are similar). Assume that the true signal x_* underlying observations b is sparse, *i.e.*, it has at most $s \ll n = \dim x_*$ non-zero entries, which is the standard motivation when using the ℓ_1 lasso.[27] For the s-sparse signal x_* we have $\|x_*\|_1 \le \sqrt{s}\|x_*\|_2$, and we can expect something similar from the optimal solution x^* to the lasso problem – at least in the low-noise regime, where x^* should be close to x_*, for otherwise why bother using lasso in the first place? Thus, it is natural to assume that $\|x^*\|_1 \le O(1)\sqrt{s}\|x^*\|_2$ with some $s \ll n$. This assumption allows us to compare the efficiency estimates

$$\epsilon_t \le O(1)\big(\sqrt{\ln(n)}\|x^*\|_1\|A\|_{\|\cdot\|_1,\|\cdot\|_2}\big)^2 t^{-\kappa}$$

(see (2.27)) of the gradient methods associated with the 'ℓ_1 proximal set-up' (the one given by Corollary 2.2 in the case of one-dimensional blocks z^j) with the estimates

$$\epsilon_t \le O(1)\big(\|x^*\|_2\|A\|_{\|\cdot\|_2,\|\cdot\|_2}\big)^2 t^{-\kappa}$$

(see (2.28)) associated with the Euclidean set-up. The ratio of the right-hand sides of these efficiency estimates is

$$\mathcal{R} = O(1)\Bigg(\underbrace{\frac{\sqrt{\ln(n)}\|x^*\|_1}{\|x^*\|_2}}_{P}\,\underbrace{\frac{\max_j \|\mathrm{Col}_j[A]\|_2}{\sigma_1(A)}}_{Q}\Bigg)^2.$$

The factor P is always ≥ 1 and thus is 'against' the ℓ_1 set-up. However, we are in the situation when P is not too large: $P \le O(1)\sqrt{\ln(n)s}$. In contrast, the factor Q is 'in favour of' the ℓ_1 set-up: it is always ≤ 1 and can be as small as $1/\sqrt{n}$ (look what happens when A is the $m \times n$ all-ones matrix). Thus, from the worst-case perspective, in the range $\sqrt{\ln(n)s} \ll \sqrt{n}$ the ℓ_1 set-up is significantly preferable to the Euclidean one. This being said, note that in some applications, *e.g.*, in compressive sensing, Q is typically significantly larger than its worst-case sharp lower bound $1/\sqrt{n}$. For example, matrices popular in compressive sensing are filled by independent

[27] An alternative motivation is to get a 'nearly sparse' approximation to x_* by penalizing the approximant with a 'sparsity-promoting' ℓ_1 norm; this motivation leads to the same conclusions as the one where x_* itself is sparse.

realizations of $\mathcal{N}(0, 1/m)$ random variables, or random variables taking values $\pm 1/\sqrt{m}$ with probability 1/2. For an $m \times n$ matrix A of this type, a typical value of $\max_j \|\mathrm{Col}_j[A]\|_2$ is $O(1)$, while a typical value of $\sigma_1(A)$ is $O(1)\sqrt{n/m}$, which yields $Q = O(1)\sqrt{m/n}$. When $m/n = O(1)$, which is often the case, $Q = O(1)\sqrt{m/n}$ certainly resolves the 'competition' between the ℓ_1 and the Euclidean set-ups in favour of the latter.

REFERENCES[28]

A. Argyriou, T. Evgeniou and M. Pontil (2008), 'Convex multi-task machine learning', *Mach. Learning* **73**, 243–272.

S. Arora and S. Kale (2007), A combinatorial, primal–dual approach to semidefinite programs. In *Proc. 39th Annual ACM Symposium on Theory of Computing*, ACM, pp. 227–236.

S. Arora, E. Hazan and S. Kale (2005), Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *Proc. 46th Annual IEEE Symposium on Foundations of Computer Science*, IEEE Computer Society, pp. 339–348.

A. W. d'Aspremont (2011), 'Subsampling algorithms for semidefinite programming', *Stochastic Systems* **1**, 274–305.

F. R. Bach, J. Mairal and J. Ponce (2008), Convex sparse matrix factorizations. Preprint, Laboratoire d'Informatique de l'Ecole Normale Supérieure. arXiv:0812.1869v1

M. Baes, M. Bürgisser and A. Nemirovski (2013), Randomized Mirror-Prox method for solving structured large-scale matrix saddle-point problems. *SIAM J. Optim.*, to appear.

R. Baraniuk, V. Cevher, M. F. Duarte and C. Hegde (2010), 'Model-based compressive sensing', *IEEE Trans. Inform. Theory* **56**, 1982–2001.

A. Beck and M. Teboulle (2003), 'Mirror descent and nonlinear projected subgradient methods for convex optimization', *Oper. Res. Lett.* **31**, 167–175.

A. Beck and M. Teboulle (2009*a*), 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems', *SIAM J. Imaging Sci.* **2**, 183–202.

A. Beck and M. Teboulle (2009*b*), 'Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems', *IEEE Trans. Image Processing* **18**, 2419–2434.

S. Becker, E. Candès and M. Grant (2010), 'Templates for convex cone problems with applications to sparse signal recovery', *Math. Prog. Comput.* **3**, 165–218.

P. Bickel, Y. Ritov and A. Tsybakov (2008), 'Simultaneous analysis of Lasso and Dantzig selector', *Ann. Statist.* **37**, 1705–1732.

J.-F. Cai, E. Candès and Z. Shen (2008), 'A singular value thresholding algorithm for matrix completion', *SIAM J. Optim.* **20**, 1956–1982.

[28] The URLs cited in this work were correct at the time of going to press, but the publisher and the authors make no undertaking that the citations remain live or are accurate or appropriate.

E. Candès (2006), Compressive sampling. In *Proc. International Congress of Mathematicians 2006*, Vol. 3, *Invited Lectures* (M. Sanz-Solé, J. Soria, J. L. Varona and J. Verdera, eds), European Mathematical Society, pp. 1433–1452.

E. Candès and B. Recht (2008), 'Exact matrix completion via convex optimization', *Found. Comput. Math.* **9**, 717–772.

E. Candès and T. Tao (2006), 'Decoding by linear programming', *IEEE Trans. Inform. Theory* **51**, 4203–4215.

E. Candès and T. Tao (2007), 'The Dantzig selector: Statistical estimation when p is much larger than n', *Ann. Statist.* **35**, 2313–23516.

E. Candès and T. Tao (2009), 'The power of convex relaxation: Near-optimal matrix completion', *IEEE Trans. Inform. Theory* **56**, 2053–2080.

E. Candès, J. Romberg and T. Tao (2006), 'Stable signal recovery from incomplete and inaccurate measurements', *Comm. Pure Appl. Math.* **59**, 1207–1223.

A. Chambolle (2004), 'An algorithm for total variation minimization and applications', *J. Math. Imaging Vision* **20**, 89–107.

V. Chandrasekaran, B. Recht, P. Parrilo and A. Willsky (2012), 'The convex geometry of linear inverse problems', *Found. Comput. Math.* **12**, 805–849.

V. Chandrasekaran, D. Sanghavi, P. Parrilo and A. Willsky (2011), 'Rank-sparsity incoherence for matrix decomposition', *SIAM J. Optim.* **21**, 572–596.

C. Chesneau and M. Hebiri (2008), 'Some theoretical results on the grouped variables lasso', *Math. Methods Statist.* **27**, 317–326.

A. Cohen, W. Dahmen and R. DeVore (2009), 'Compressed sensing and best k-term approximation', *J. Amer. Math. Soc.* **22**, 211–231.

W. Dai, M. Sheikh, O. Milenkovic, and R. Baraniuk (2009), 'Compressive sensing DNA microarrays', *EURASIP J. Bioinformatics and Systems Biology* **2009**.

O. Devolder (2011), Stochastic first order methods in smooth convex optimization. CORE Discussion paper 2011/70.

D. Donoho and J. Tanner (2005), 'Sparse nonnegative solutions of underdetermined linear equations by linear programming', *Proc. Natl Acad. Sci. USA* **102**, 9446–9451.

D. Donoho, M. Elad and V. Temlyakov (2006), 'Stable recovery of sparse overcomplete representations in the presence of noise', *IEEE Trans. Inform. Theory* **53**, 6–18.

Y. Eldar, P. Kuppinger and H. Bölcskei (2010), 'Block-sparse signals: uncertainty relations and efficient recovery', *IEEE Trans. Signal Processing* **58**, 3042–3054.

E. Elhamifar and R. Vidal (2012), 'Block-sparse recovery via convex optimization', *IEEE Trans. Signal Processing* **60** 4094–4107.

D. Goldfarb and S. Ma (2011), 'Convergence of fixed-point continuation algorithms for matrix rank minimization', *Found. Comput. Math.* **11**, 183–210.

D. Goldfarb and W. Yin (2009), 'Parametric maximum flow algorithms for fast total variation minimization', *SIAM J. Sci. Comput.* **31**, 3712–3743.

D. Goldfarb, K. Scheinberg and B. Xi (2011), Fast first order methods for composite convex optimization with backtracking.
http://www.optimization-online.org/DB_HTML/2011/04/3004.html

M. D. Grigoriadis and L. G. Khachiyan (1995), 'A sublinear-time randomized approximation algorithm for matrix games', *Oper. Res. Lett.* **18**, 53–58.

M. Herman and T. Strohmer (2007), 'High-resolution radar via compressed sensing', *IEEE Trans. Signal Process.* **57**, 2275–2284.

B. Huang, S. Ma and D. Goldfarb (2013), 'Accelerated linearized Bregman method', *J. Sci. Comput.* **54**, 428–453.

J. Huang and T. Zhang (2010), 'The benefit of group sparsity', *Ann. Statist.* **38**, 1978–2004.

M. Jaggi and M. Sulovský (2010), A simple algorithm for nuclear norm regularized problems. In *ICML 2010: Proc. 27th International Conference on Machine Learning, June 2010*, Omnipress, pp. 471–478.
www.icml2010.org/papers/196.pdf

A. Juditsky and A. Nemirovski (2011*a*), 'On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization', *Math. Program.* B **127**, 57–88.

A. Juditsky and A. Nemirovski (2011*b*), First-order methods for nonsmooth large-scale convex minimization: I General purpose methods; II Utilizing problem's structure. In *Optimization for Machine Learning* (S. Sra, S. Nowozin and S. Wright, eds), The MIT Press, pp. 121–183.

A. Juditsky and A. Nemirovski (2011*c*), 'Accuracy guarantees for ℓ_1 recovery', *IEEE Trans. Inform. Theory* **57**, 7818–7839.

A. Juditsky, F. Kılınç Karzan and A. Nemirovski (2011*a*), 'Verifiable conditions of ℓ_1-recovery of sparse signals with sign restrictions', *Math. Program.* B **127**, 89–122.

A. Juditsky, F. Kılınç Karzan and A. Nemirovski (2013*a*), Randomized first order algorithms with applications to ℓ_1-minimization. *Math. Program.*, to appear.

A. Juditsky, F. Kılınç Karzan, A. S. Nemirovski and B. T. Polyak (2011*b*), On the accuracy of ℓ_1-filtering of signals with block-sparse structure. In *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor *et al.*, eds), Neural Information Processing Systems Foundation (NIPS), pp. 1260–1268.

A. Juditsky, F. Kılınç Karzan, A. Nemirovski and B. Polyak (2013*b*), Accuracy guarantees for ℓ_1 recovery of block-sparse signals. *Ann. Statist.*, to appear.

A. Juditsky, A. Nemirovski and C. Tauvel (2011*c*), 'Solving variational inequalities with Stochastic Mirror Prox algorithm', *Stochastic Systems* **1**, 17–58.

J. Lee, B. Recht, R. Salakhutdinov, N. Srebro and J. Tropp (2010), Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems 23* (J. Lafferty *et al.*, eds), NIPS, pp. 1297–1305.

C. Lemarechal, A. Nemirovski and Y. Nesterov (1995), 'New variants of bundle methods', *Math. Program.* **69**, 111–148.

H. Liu and J. Zhang (2009), 'Estimation consistency of the group lasso and its applications', *J. Mach. Learning Res. Proc. Track* **5**, 376–383.

H. Liu, J. Shang, X. Jiang and J. Liu (2010), 'The group Dantzig selector', *J. Mach. Learning Res. Proc. Track* **9**, 461–468.

Y.-J. Liu, D. Sun and K.-C. Toh (2012), 'An implementable proximal point algorithmic framework for nuclear norm minimization', *Math Program.* **133**, 399–436.

K. Lounici, M. Pontil, S. van de Geer and A. Tsybakov (2011), 'Oracle inequalities and optimal inference under group sparsity', *Ann. Statist.* **39**, 2164–2204.

M. Lustig, D. Donoho and J. M. Pauly (2007), 'Sparse MRI: The application of compressed sensing for rapid MR imaging', *Magnetic Resonance in Medicine* **56**, 1182–1195.
S. Ma, D. Goldfarb and L. Chen (2011), 'Fixed point and Bregman iterative methods for matrix rank minimization', *Math. Program.* **128**, 321–353.
L. Meier, S. van de Geer and P. Bühlmann (2008), 'The group lasso for logistic regression', *J. Roy. Statist. Soc.* B **70**, 53–71.
N. Meinshausen and B. Yu (2009), 'Lasso-type recovery of sparse representations for high-dimensional data', *Ann. Statist.* **37**, 246–270.
H. Men, N. Nguyen, R. Freund, K. Lim, P. Parrilo and J. Peraire (2011), 'Design of photonic crystals with multiple and combined band gaps', *Phys. Rev.* E **83**, 046703.
Y. Nardi and A. Rinaldo (2008), 'On the asymptotic properties of the group lasso estimator for linear models', *Electron. J. Statist.* **2**, 605–633.
A. Nemirovski (1992), 'Information-based complexity of linear operator equations', *J. Complexity* **8**, 153–175.
A. Nemirovski (2004), 'Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex–concave saddle point problems', *SIAM J. Optim.* **15**, 229–251.
A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro (2009), 'Stochastic approximation approach to stochastic programming', *SIAM J. Optim.* **19**, 1574–1609.
Y. Nesterov (1983), 'A method for solving a convex programming problem with rate of convergence $O(1/k^2)$', *Soviet Math. Doklady* **27**, 372–376.
Y. Nesterov (2005*a*), 'Smooth minimization of non-smooth functions', CORE Discussion Paper 2003/12 and *Math. Program.* **103**, 127–152.
Y. Nesterov (2005*b*), 'Excessive gap technique in nonsmooth convex minimization', *SIAM J. Optim.* **16**, 235–239.
Y. Nesterov (2007*a*), 'Dual extrapolation and its application for solving variational inequalities and related problems', *Math. Program.* **109**, 319–344.
Y. Nesterov (2007*b*), Gradient methods for minimizing composite objective functions. CORE Discussion Paper 2007/76.
Y. Nesterov (2009), 'Primal–dual subgradient methods for convex problems', *Math. Program.* B **120**, 221–259.
Y. Nesterov and A. Nemirovski (1994), *Interior Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics.
Y. Nesterov and B. Polyak (2006), 'Cubic regularization of Newton's method and its global performance', *Math. Program.* **108**, 177–205.
S. Osher, Y. Mao, B. Dong and W. Yin (2010), 'Fast linearized Bregman iteration for compressive sensing and sparse denoising', *Commun. Math. Sci.* **8**, 93–111.
F. Parvaresh, H. Vikalo, S. Misra and B. Hassibi (2008), 'Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays', *IEEE J. Selected Topics Signal Process.* **2**, 275–285.
Z. Qin and D. Goldfarb (2012), Structured sparsity via alternating direction methods. *J. Mach. Learning Res.* **13**, 1435–1468.
B. Recht and C. Ré (2011), Parallel stochastic gradient algorithms for large-scale matrix completion.
http://www.optimization-online.org/DB_HTML/2011/04/3012.html

B. Recht, M. Fazel and P. Parrilo (2010), 'Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization', *SIAM Rev.* **52**, 471–501.

B. Recht, C. Ré, S. Wright and F. Niu (2011*a*), Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor *et al.*, eds), NIPS, pp. 693–701.

B. Recht, W. Xuy and B. Hassibi (2011*b*), 'Null space conditions and thresholds for rank minimization', *Math. Program.* B **127**, 175–211.

L. Rudin, S. Osher and E. Fatemi (1992), 'Nonlinear total variation based noise removal algorithms', *Physica* D **60**, 259–268.

S. Santosa and W. Symes (1986), 'Linear inversion of band-limited reflection seismograms', *SIAM J. Sci. Comput.* **7**, 1307–1330.

K. Scheinberg, S. Ma and D. Goldfarb (2010), Sparse inverse covariance selection via alternating linearization methods. In *Advances in Neural Information Processing Systems 23* (J. Lafferty *et al.*, eds), NIPS, pp. 2101–2109.

J. Shi, W. Yin, S. Osher and P. Sajda (2010), 'A fast hybrid algorithm for large-scale ℓ_1-regularized logistic regression', *J. Mach. Learning Res.* **11**, 713–741.

N. Srebro and A. Shraibman (2005), Rank, trace-norm and max-norm. In *Proc. 18th Annual Conference on Learning Theory* (COLT), Vol. 3559 of *Lecture Notes in Computer Science*, Springer, pp. 545–560.

V. Studer, J. Bobin, M. Chahid, H. Mousavi, E. Candès and M. Dahan (2012), 'Compressive fluorescence microscopy for biological and hyperspectral imaging', *Proc. Natl Acad. Sci. USA* **109**, E1679–E1687.

H. Taylor, S. Banks and J. McCoy (1979), 'Deconvolution with the ℓ_1-norm', *Geophys.* **44**, 39–52.

R. Tibshirani (1996), 'Regression shrinkage and selection via the lasso', *J. Roy. Statist. Soc.* B **58**, 267–288.

J. Tropp (2006), 'Just relax: Convex programming methods for identifying sparse signals', *IEEE Trans. Inform. Theory* **51**, 1030–1051.

P. Tseng (2000), 'A modified forward–backward splitting method for maximal monotone mappings', *SIAM J. Control Optim.* **38**, 431–446.

P. Tseng (2008), On accelerated proximal gradient methods for convex–concave optimization. Technical report. Submitted to *SIAM J. Optim.* www.math.washington.edu/~tseng/papers/apgm.pdf

S. Vasanawala, M. T. Alley, B. A. Hargreaves, R. A. Barth, J. M. Pauly and M. Lustig (2010), 'Improved pediatric MR imaging with compressed sensing', *Radiology* **256**, 607–616.

G. Wagner, P. Schmieder, A. Stern and J. Hoch (1993), 'Application of nonlinear sampling schemes to cosy-type spectra', *J. Biomolecular NMR* **3**, 569–576.

Z. Wen, W. Yin, D. Goldfarb and Y. Zhang (2010), 'A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation', *SIAM J. Sci. Comput.* **32**, 1832–1857.

Z. Wen, W. Yin, H. Zhang and D. Goldfarb (2012), 'On the convergence of an active-set method for ℓ_1 minimization', *Optim. Software* **27**, 1127–1146.

J. Yang and X. Yuan (2013), 'Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization', *Math. Comp.* **82**, 301–329.

W. Yin, S. Osher, D. Goldfarb and J. Darbon (2008), 'Bregman iterative algorithms for L_1-minimization with applications to compressed sensing', *SIAM J. Imaging Sci.* **1**, 143–168.

M. Yuan and Y. Lin (2006), 'Model selection and estimation in regression with grouped variables', *J. Roy. Statist. Soc.* B **68**, 49–67.