Fast Saddle-Point Algorithm for Generalized Dantzig Selector and FDR Control with the Ordered ℓ_1-Norm

Sangkyun Lee
Technische Universität Dortmund

Damian Brzyski
Jagiellonian University

Malgorzata Bogdan
University of Wroclaw

Abstract

In this paper we propose a primal-dual proximal extragradient algorithm to solve the generalized Dantzig selector (GDS) estimation, based on a new convex-concave saddle-point (SP) formulation of the GDS and a simple gradient extrapolation technique. Our reformulation makes it possible to adapt recent developments in saddle-point optimization, to achieve the optimal $O(1/k)$ rate of convergence. Compared to the optimal non-SP algorithms, ours do not require specification of sensitive parameters that affect algorithm performance or solution quality. We also provide a new analysis showing a possibility of acceleration in special cases even without strong convexity or strong smoothness. As an application, we propose a GDS equipped with the ordered ℓ_1-norm, showing its false discovery rate control properties in variable selection. Algorithm performance is compared between ours and other alternatives, including the linearized ADMM, Nesterov's smoothing, Nemirovski's mirror-prox, and the accelerated hybrid proximal extragradient techniques.

1 INTRODUCTION

The Dantzig selector (Candes and Tao, 2007) has been proposed as an alternative approach for penalized regression, mainly in the context of sparse or group sparse regression in high dimensions. A generalized Dantzig selector (GDS) (Chatterjee et al., 2014) has been recently proposed extending the Dantzig selector to use any norm $\mathcal{R}(\cdot)$ for regularization and its dual norm $\mathcal{R}_*(\cdot)$ for measuring estimation error. For linear models of the form $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}$, where $\mathbf{y} \in \mathbb{R}^n$ contains observations, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a design matrix, and $\boldsymbol{\xi}$ is an i.i.d. standard Gaussian noise vector, the GDS searches for the best parameter solving the following problem with a constant $c > 0$:

$$\min_{\mathbf{w} \in \mathbb{R}^p} \ \mathcal{R}(\mathbf{w}) \ \text{ s.t. } \ \mathcal{R}_*(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w})) \le c. \qquad (1)$$

The original Dantzig selector is attained when $\mathcal{R}(\cdot) = \|\cdot\|_1$ and $\mathcal{R}_*(\cdot) = \|\cdot\|_\infty$. In general, it requires to solve a non-separable and non-smooth convex optimization problem, which does not contain any strongly smooth part (with Lipschitz continuous gradients) required to apply (accelerated) proximal gradient methods (Nesterov, 1983; Beck and Teboulle, 2009). Subgradient methods (Shor et al., 1985) can be applied, but theirs very slow $O(1/\sqrt{k})$ convergence rate (for an iteration counter k) is not desirable for practical use.

Chatterjee et al. (2014) proposed an algorithm based on the linearized alternating direction method of multipliers (L-ADMM) (Wang and Banerjee, 2014; Wang and Yuan, 2012)), of which two subproblems are simplified to two proximal operations thanks to linearization and fast projection. Constraint sets for projection were restricted to the dual balls defined with $\mathcal{R}_*(\cdot)$ and therefore easily computed via the proximal operator of $\mathcal{R}(\cdot)$, due to Moreau's identity (Rockafellar, 1997). The algorithm exhibits $O(1/k)$ convergence rate when its penalty parameter is at least $\|\mathbf{X}\|_2^4$ (Chatterjee et al., 2014; Wang and Banerjee, 2014). However, the practical performance of this algorithm tends to be quite sensitive to the parameter, whose best values are not easy to determine a priori running the algorithm.

Recently, there have been attractive improvements in ADMM, although they are not applicable to our problem due to their extra requirements. Local linear rates of convergence has been shown for ADMM, but for the limited cases of minimizing a quadratic objective under linear constraints (Boley, 2013), or minimizing a sum of strongly convex smooth functions (Shi et al., 2014). Accelerated versions of the ADMM recently ap-

Preliminary work. Under review by AISTATS 2016. Do not distribute.

peared achieving a better $O(1/k^2)$ rate, however, with an assumption that the objective is strongly convex for ADMM (Goldstein et al., 2014; Kadkhodaie et al., 2015), or with a smoothness assumption of the part to be linearized for L-ADMM (Ouyang et al., 2015).

The GDS problem (1) can also be solved using the smoothing technique due to Nesterov (2005). It is based on creating a smooth approximation of a non-smooth function by adding a strongly convex regularizer to the conjugate of the non-smooth function, where the strong convexity is modulated by a parameter $\mu > 0$. It is shown that the smooth approximation has Lipschitz continuous gradients and therefore can be optimized via accelerated gradient methods (Nesterov, 1983). The smoothing technique achieves $O(1/k)$ rate of convergence when $\mu = O(\epsilon)$ (Nesterov, 2005; Theorem 3) for an optimality gap ϵ. However, using small values of μ to achieve a near-optimal solution tends to slow down the algorithm quite significantly in practice. Implementations of Nesterov's smoothing such as TFOCS (Becker et al., 2011) require users to specify this parameter with only little guidance.

In this paper, we propose a new convex-concave saddle point (CCSP) formulation of the GDS, in fact a slightly more generalized version of it to allow for using any convex function for regularization. Our reformulation allows us to provide a fast and simple algorithm to find solutions of GDS instances, achieving the optimal $O(1/k)$ convergence rate without relying on sensitive parameters affecting convergence or solution quality. Our algorithm is applied to a new kind of GDS defined with the ordered ℓ_1-norm, showing its false discovery rate control properties in variable selection, where the norm itself has been recently studied in other contexts (Figueiredo and Nowak, 2014; Bogdan et al., 2015).

We show that our proposed algorithm suits better than existing solvers when high-precision solutions are desired for accurate variable selection, for example in statistical simulation studies. We denote the Euclidean norm by $\|\cdot\|$.

2 CONVEX-CONCAVE SADDLE-POINT FORMULATION

2.1 (More) Generalized Dantzig Selector

In this paper we consider a slightly more general form of the generalized Dantzig selector (GDS) problem,

$$\text{(GDS)} \quad \min_{\mathbf{w}\in\mathbb{R}^p} \ \mathcal{F}(\mathbf{w}) \ \text{ s.t. } \ \mathcal{G}_*(\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})) \le 1. \quad (2)$$

where $\mathcal{F}:\mathbb{R}^p \to (-\infty,+\infty]$ is a proper, convex, and lower-semicontinuous (l.s.c.) function, and $\mathcal{G}_*(\cdot)$ is the dual norm of a norm $\mathcal{G}(\cdot)$, possibly parametrized by a vector $\boldsymbol{\lambda}$. Unlike (1), $\mathcal{G}$ is not necessarily the same as $\mathcal{F}$, and also $\mathcal{F}$ does not have to be a norm. Neither $\mathcal{F}$ nor $\mathcal{G}$ is assumed to be differentiable.

2.2 Reformulation

Denoting by $C_{\mathcal{G}_*}$ the constraint set of residuals in (2), i.e.,

$$C_{\mathcal{G}_*} := \{\mathbf{r}\in\mathbb{R}^p \ : \ \mathcal{G}_*(\mathbf{r}) \le 1\},$$

and using an indicator function $\vartheta_{C_{\mathcal{G}_*}}(\mathbf{r})$, which returns 0 if $\mathbf{r}\in C_{\mathcal{G}_*}$ or $+\infty$ otherwise, it is trivial to see the GDS problem (2) can be restated as,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \ \mathcal{F}(\mathbf{w}) + \vartheta_{C_{\mathcal{G}_*}}(\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})). \quad (3)$$

Now, we invoke a simple lemma to replace the indicator function with its adjoint form.

Lemma 1. *For any $\mathbf{w}\in\mathbb{R}^p$, we have*

$$\vartheta_{C_{\mathcal{G}_*}}(\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})) = \max_{\mathbf{v}\in\mathbb{R}^p} \left\langle \mathbf{A}\begin{bmatrix}\mathbf{y}\\ \mathbf{w}\end{bmatrix}, \mathbf{v}\right\rangle - \mathcal{G}(\mathbf{v}),$$

where $\mathbf{A} := \mathbf{X}^T\begin{bmatrix}\mathbf{I}_n & -\mathbf{X}\end{bmatrix} \in \mathbb{R}^{p\times(n+p)}$ and $\mathbf{I}_n$ is the $n\times n$ identity matrix.

Proof. Since $\vartheta_{C_{\mathcal{G}_*}}$ is an indicator function on a closed set, we have $\vartheta_{C_{\mathcal{G}_*}}(\cdot) = \vartheta^{\star\star}_{C_{\mathcal{G}_*}}(\cdot)$ with the biconjugation

$$\vartheta^{\star\star}_{C_{\mathcal{G}_*}}(\mathbf{r}) = \sup_{\mathbf{v}\in\mathbb{R}^p} \{\langle\mathbf{r},\mathbf{v}\rangle - \vartheta^{\star}_{C_{\mathcal{G}_*}}(\mathbf{v})\}.$$

Also, from conjugacy, $\vartheta^{\star}_{C_{\mathcal{G}_*}}(\cdot) = \sup_{\mathbf{w}'\in\mathbb{R}^p} \ \langle\mathbf{w}',\cdot\rangle - \vartheta_{C_{\mathcal{G}_*}}(\mathbf{w}') = \max_{\mathbf{w}':\mathcal{G}_*(\mathbf{w}')\le 1}\langle\mathbf{w}',\cdot\rangle$, which is by definition the dual norm of $\mathcal{G}_*(\cdot)$, i.e., $\mathcal{G}(\cdot)$. The result follows when we set $\mathbf{r} = \mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})$. □

The following convex-concave saddle point reformulation of the GDS (2) follows when we apply the above lemma to (3),

$$\text{(GDS-SP)} \quad \min_{\mathbf{w}\in\mathbb{R}^p}\max_{\mathbf{v}\in\mathbb{R}^p} \left\langle \mathbf{A}\begin{bmatrix}\mathbf{y}\\ \mathbf{w}\end{bmatrix}, \mathbf{v}\right\rangle + \mathcal{F}(\mathbf{w}) - \mathcal{G}(\mathbf{v}). \quad (4)$$

This reformulation allows us to benefit from recent developments in saddle-point algorithms, including our algorithm discussed later. Hereafter, we assume that both $\mathcal{F}$ and $\mathcal{G}$ are simple, so that their *proximal operator*, defined below for $\mathcal{F}$, can be computed efficiently:

$$\operatorname{prox}_{\tau\mathcal{F}}(\mathbf{z}) := \arg\min_{\mathbf{w}'} \ \left\{\frac{1}{2}\|\mathbf{w}'-\mathbf{z}\|^2 + \tau\mathcal{F}(\mathbf{w}')\right\}.$$

Note that it suffices to meet this requirement for either $\mathcal{F}$ or its conjugate $\mathcal{F}^\star$ (similarly for $\mathcal{G}$ or $\mathcal{G}^\star$), since the prox operation for one can be computed by that of the other, i.e.,

$$\mathbf{z} = \mathrm{prox}_{\mathcal{F}}(\mathbf{z}) + \mathrm{prox}_{\mathcal{F}^\star}(\mathbf{z})$$

by Moreau's identity (Rockafellar, 1997).

2.3 Related Works

It is worthwhile to note that the Tikhonov-type formulation of the GDS (3) is closely related to the popular regularized estimation problems in machine learning and statistics,

$$\min_{\mathbf{w}\in\mathcal{W}} \ \mathcal{F}(\mathbf{w}) + \mathcal{E}(\mathbf{D}\mathbf{w}),$$

where $\mathbf{D}$ is a data matrix and $\mathcal{E}$ is a proper convex l.s.c. loss function. Using biconjugation of $\mathcal{E}$ similarly to the proof of Lemma 1, this can be reformulated as the following convex-concave saddle point problem,

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} \ \phi(\mathbf{w},\mathbf{v}) := \langle \mathbf{D}\mathbf{w}, \mathbf{v}\rangle + \mathcal{F}(\mathbf{w}) - \mathcal{E}^\star(\mathbf{v}),$$

given that a maximizer in $\mathcal{V}$ can be attained (in our case it is true as $\mathcal{V} = \{\mathbf{w}' : \mathcal{G}_*(\mathbf{w}') \le 1\}$ is compact).

This type of reformulation has been studied quite recently in machine learning to design new algorithms. For example, Zhang and Xiao (2015) proposed a stochastic primal-dual coordinate descent (SPDC) algorithm based on the saddle-point reformulation for the cases when $\mathcal{F}$ is strongly convex and $\mathcal{E}$ is a sum of smooth loss functions with Lipschitz continuous gradients, in which case the conjugate $\mathcal{E}^\star$ becomes strongly convex (Rockafellar and Wets, 2004; Proposition 12.60): both do not hold in case of the GDS. Although SPDC can be extended for nonsmooth cases by augmenting strongly convex terms, but then it shares similar issues to Nesterov's smoothing that a parameter needs to be specified depending on a unknown quantity $\|\mathbf{w}^*\|$ when $(\mathbf{w}^*, \mathbf{v}^*)$ is a saddle point.

Another example is Taskar et al. (2006) who considered a saddle-point reformulation of max-margin estimation models for structured output and proposed an algorithm more memory-efficient than its QP alternative, based on the dual extragradient technique (Nesterov, 2007). The dual extragradient itself is closely related to our method, but it additionally requires that both $\mathcal{F}$ and $\mathcal{E}$ are smooth with Lipschitz continuous gradients to achieve the ergodic $O(1/k)$ convergence rate, or both $\partial\mathcal{F}$ and $\partial\mathcal{E}$ are bounded to have a slower $O(1/\sqrt{k})$ rate.

Extragradient techniques to handle the CCSP problems are of our particular interest. The mirror-prox method (Nemirovski, 2004) has extended one of the earliest extragradient algorithm of Korpelevich (1976), establishing the $O(1/k)$ ergodic (in terms of averaged iterates) rate of convergence with two proximal operations per iteration. This method however requires to choose stepsizes carefully with the knowledge of $L = \|\mathbf{A}\|$. Tseng (2008) suggested a line search procedure to find better estimates of L, which requires to compute two extra proximal operations per line search step.

The hybrid proximal extragradient (HPE) algorithm (Solodov and Svaiter, 1999a,b) is another family of extragradient methods that can solve CCSPs with the same $O(1/k)$ ergodic convergence rate, which also can be seen as a generalization of Korpelevich's method and its extensions (Monteiro and Svaiter, 2011). In each iteration of the HPE framework, an extragradient is computed by solving a subproblem with controlled inaccuracy. The subproblem itself can be solved using an accelerated method similar to Nesterov's smoothing (He and Monteiro, 2014) using three proximal operations in each inner iteration. A pitfall however is that the accuracy of solving the subproblem tends to affect the overall runtime.

Recently, Chambolle and Pock (2011) proposed a simple extragradient technique with $O(1/k)$ ergodic convergence rate, which is quite different in its nature to above extragradient methods, although it looks similar to Nesterov's dual extragradient (Nesterov, 2007). In their method, proximal steps are taken in the primal and the dual space, then a linear gradient extrapolation is considered either in the primal or in the dual space. We base our algorithm on this technique, since it has been the fastest with the smallest variations in runtime to solve the GDS problem. Both properties were desired in particular for statistical studies of the GDS using random data.

3 ALGORITHM

Solving the GDS-SP problem (4), we assume that there exists a saddle point $(\mathbf{w}^*, \mathbf{v}^*)$ satisfying the conditions

$$\begin{aligned} \mathbf{A}\begin{bmatrix}\mathbf{y}\\ \mathbf{w}^*\end{bmatrix} &= \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w}^* \in \partial\mathcal{G}(\mathbf{v}^*), \\ -(\mathbf{A}_{[\cdot,(n+1):(n+p)]})^T\mathbf{v}^* &= \mathbf{X}^T\mathbf{X}\mathbf{v}^* \in \partial\mathcal{F}(\mathbf{w}^*) \end{aligned} \tag{5}$$

where $\partial\mathcal{F}$ and $\partial\mathcal{G}$ are the subdifferentials of $\mathcal{F}$ and $\mathcal{G}$, resp. Denoting the objective by ϕ, i.e.,

$$\phi(\mathbf{w},\mathbf{v}) := \left\langle \mathbf{A}\begin{bmatrix}\mathbf{y}\\ \mathbf{w}\end{bmatrix}, \mathbf{v}\right\rangle + \mathcal{F}(\mathbf{w}) - \mathcal{G}(\mathbf{v}),$$

the above conditions imply that the following saddle-point inequality holds for any $(\mathbf{w}, \mathbf{v})$,

$$\phi(\mathbf{w}^*, \mathbf{v}) \le \phi(\mathbf{w}^*, \mathbf{v}^*) \le \phi(\mathbf{w}, \mathbf{v}^*).$$

Algorithm 1: SP-PLX: (Accelerated) Saddle-Point Proximal Linear eXtragradient Algorithm

Data : $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^n$
Initialize: $(\mathbf{w}_0, \mathbf{v}_0) \in \mathbb{R}^p \times \mathbb{R}^p$, $\mathbf{w}'_0 = \mathbf{w}_0$
Params : τ_0, $\sigma_0 > 0$: $\tau_0\sigma_0 L^2 \le 1$, $\gamma \ge 0$: strong convexity modulus of $\mathcal{G}$

for $k = 0, 1, 2, \dots$ **do**

$$\begin{aligned}
\mathbf{v}_{k+1} &= \mathrm{prox}_{\sigma\mathcal{G}}\left(\mathbf{v}_k + \sigma_k(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w}'_k)\right),\\
\mathbf{v}'_{k+1} &= \mathbf{v}_{k+1} \text{ (or } 2\mathbf{v}_{k+1}\text{, see Section 3.1)},\\
\mathbf{w}_{k+1} &= \mathrm{prox}_{\tau\mathcal{F}}\left(\mathbf{w}_k + \tau_k\mathbf{X}^T\mathbf{X}\mathbf{v}'_{k+1}\right),\\
\theta_k &= 1/\sqrt{1+2\gamma\tau_k},\\
\tau_{k+1} &= \theta_k\tau_k, \quad \sigma_{k+1} = \sigma_k/\theta_k,\\
\mathbf{w}'_{k+1} &= \mathbf{w}_{k+1} + \theta_k(\mathbf{w}_{k+1} - \mathbf{w}_k).
\end{aligned}$$

Check (both if $\gamma = 0$, the former if $\gamma > 0$):

- Pointwise convergence of $(\mathbf{w}_{k+1}, \mathbf{v}_{k+1})$;
- Ergodic convergence of $(\overline{\mathbf{w}}_{k+1}, \overline{\mathbf{v}}_{k+1}) = \frac{1}{k+1}\sum_{i=1}^{k+1}(\mathbf{w}_i, \mathbf{v}_i)$;

end

We present our linear extragradient algorithm in Algorithm 1, which solves the CCSP formulation of the GDS problem (4).

We define the primal-dual gap, following Chambolle and Pock (2011) restricted to the set $\mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned}
\mathcal{T}_{\mathcal{X}\times\mathcal{Y}}(\mathbf{w}, \mathbf{v}) := \max_{\mathbf{v}'\in\mathcal{Y}} &\left\{\langle \mathbf{A}\begin{bmatrix}\mathbf{y}\\ \mathbf{w}\end{bmatrix}, \mathbf{v}'\rangle + \mathcal{F}(\mathbf{w}) - \mathcal{G}(\mathbf{v}')\right\}\\
&- \min_{\mathbf{w}'\in\mathcal{X}}\left\{\langle \mathbf{A}\begin{bmatrix}\mathbf{y}\\ \mathbf{w}'\end{bmatrix}, \mathbf{v}\rangle + \mathcal{F}(\mathbf{w}') - \mathcal{G}(\mathbf{v})\right\}.
\end{aligned}$$

Note that if $\mathcal{X} \times \mathcal{Y}$ contains a saddle-point $(\mathbf{w}^*, \mathbf{v}^*)$ satisfying (5), then it is easy to check that

$$\begin{aligned}
\mathcal{T}_{\mathcal{X}\times\mathcal{Y}}(\mathbf{w}, \mathbf{v}) \ge &\left\{\langle \mathbf{A}\begin{bmatrix}\mathbf{y}\\ \mathbf{w}\end{bmatrix}, \mathbf{v}^*\rangle + \mathcal{F}(\mathbf{w}) - \mathcal{G}(\mathbf{v}^*)\right\}\\
&- \left\{\langle \mathbf{A}\begin{bmatrix}\mathbf{y}\\ \mathbf{w}^*\end{bmatrix}, \mathbf{v}\rangle + \mathcal{F}(\mathbf{w}^*) - \mathcal{G}(\mathbf{v})\right\} \ge 0.
\end{aligned}$$

Theorem 1. *Suppose that $(\mathbf{w}^*, \mathbf{v}^*)$ is a saddle-point of the GDS-SP problem (4). Then the iterates $(\mathbf{w}_k, \mathbf{v}_k)$ generated by Algorithm 1 with $\gamma = 0$ and $\theta_k = 1$ for all k (therefore $\tau_k = \tau_0$ and $\sigma_k = \sigma_0$) satisfy the following properties:*

(a) $(\mathbf{w}_k, \mathbf{v}_k)$ is bounded for any k, i.e.,

$$\begin{aligned}
&\frac{\|\mathbf{w}_k - \mathbf{w}^*\|^2}{\tau_0} + \frac{\|\mathbf{v}_k - \mathbf{v}^*\|^2}{\sigma_0}\\
&\quad\le C\left(\frac{\|\mathbf{w}_0 - \mathbf{w}^*\|^2}{\tau_0} + \frac{\|\mathbf{v}_0 - \mathbf{v}^*\|^2}{\sigma_0}\right)
\end{aligned}$$

for a constant $C \le 1/(1 - \tau_0\sigma_0 L^2)$.

(b) For averaged iterates $\overline{\mathbf{w}}_k = \frac{1}{k}\sum_{i=1}^k \mathbf{w}_i$ and $\overline{\mathbf{v}}_k = \frac{1}{k}\sum_{i=1}^k \mathbf{v}_i$, we have

$$\mathcal{T}(\overline{\mathbf{w}}_k, \overline{\mathbf{v}}_k) \le \frac{1+C}{k}\left(\frac{\|\mathbf{w}^* - \mathbf{w}_0\|^2}{2\tau_0} + \frac{\|\mathbf{v}^* - \mathbf{v}_0\|^2}{2\sigma_0}\right).$$

Moreover, limit points of $(\overline{\mathbf{w}}_k, \overline{\mathbf{v}}_k)$ are saddle-points of (4).

(c) There exists a saddle-point $(\hat{\mathbf{w}}, \hat{\mathbf{v}})$ of (4) such that $(\mathbf{w}_k, \mathbf{v}_k) \to (\hat{\mathbf{w}}, \hat{\mathbf{v}})$ as $k \to \infty$.

Proof. Define augmentations of $\mathbf{w}$'s with $\mathbf{y}$, e.g. $\mathbf{z}_k := [\mathbf{y}; \mathbf{w}_k] \in \mathbb{R}^{n+p}$, and define $\mathcal{H}(\mathbf{z}) = \mathcal{H}(\mathbf{y}, \mathbf{w}) := \mathcal{F}(\mathbf{w})$. Using these, the GDS-SP problem (4) can be written equivalently as

$$\min_{\mathbf{z}\in\mathbb{R}^{p+n}} \max_{\mathbf{w}\in\mathbb{R}^p} \langle \mathbf{A}\mathbf{z}, \mathbf{v}\rangle + \mathcal{H}(\mathbf{z}) - \mathcal{G}(\mathbf{v}).$$

Then the result essentially follows from Theorem 1 of Chambolle and Pock (2011). For completeness, we provide the full proof in Appendix, part of which will be used to show Theorem 2 as well. □

The *ergodic convergence* in Theorem 1 part (b) indicates that the primal-dual gap converges with $O(1/k)$ rate for *averaged iterates*, which is known to be the best rate in general convex-concave saddle-point solvers (Nemirovski, 2004; Tseng, 2008; Solodov and Svaiter, 1999a,b; He and Monteiro, 2014).

The part (c) states *pointwise* convergence without averaging, where its rate is unknown: one can conjecture from related extragradient methods, e.g. (He and Monteiro, 2014; Theorem 3.4), that the convergence might be at a slower rate of $O(1/\sqrt{k})$, but it is only an educated guess since the methods are not exactly the same. In fact, in our experiments the iterates tend to converge faster than averaged iterates, which we will discuss further in detail later.

The part (a) of the above theorem is indeed quite interesting. (We note that similar boundedness results are available for some methods, e.g. Nemirovski (2004); Tseng (2008), but not for all). In particular, in many sparse regression scenarios we expect that $\|\mathbf{w}^*\|$ will not be very large due to its small support size (the number of nonzero components) and weak signals (the magnitude of the components), whereas our algorithm naturally starts from the zero vector. Thus it is likely from Theorem 1 that $\|\mathbf{w}_k - \mathbf{w}^*\|$ (or even $\|\mathbf{w}_k\|$) would be very small, although we need more information about $\|\mathbf{v}_0 - \mathbf{v}^*\|$ to say it definitely.

3.1 Restricted Strong Convexity and Acceleration

When $\mathcal{F}$ or $\mathcal{G}$ is strongly convex, it can be shown that Algorithm 1 exhibits a faster $O(1/k^2)$ *pointwise* convergence rate due to Chambolle and Pock (2011), using the same trick as in the proof of Theorem 1.

Let us focus on $\mathcal{F}$, since the same argument can be applied for $\mathcal{G}$. When $\mathcal{F}$ is strongly convex, it satisfies

$$\mathcal{F}(\mathbf{w}') \geq \mathcal{F}(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle + \frac{\gamma}{2}\|\mathbf{w}' - \mathbf{w}\|^2, \;\; \mathbf{g} \in \partial\mathcal{F}(\mathbf{w}),$$

for some $\gamma > 0$ and for any $\mathbf{w}', \mathbf{w} \in \operatorname{dom} \mathcal{F}$.

Our interest here is actually the case when $\mathcal{F}$ is *not* strongly convex, the general case of the GDS problem (4). Suppose that $\mathcal{F}$ is indeed a norm, so that it satisfies the triangle inequality, $\mathcal{F}(\mathbf{w}^*) - \mathcal{F}(\mathbf{w}_k) \leq \mathcal{F}(\mathbf{w}^* - \mathbf{w}_k)$, for a solution $\mathbf{w}^*$ and an iterate $\mathbf{w}_k$ of Algorithm 1. If $\mathcal{F}(\mathbf{w}^* - \mathbf{w}_k)$ is bounded so that $\mathcal{F}(\mathbf{w}^* - \mathbf{w}_k) \leq c\|\mathbf{w}^* - \mathbf{w}_k\|$ for some $c > 0$, where the right-hand side is bounded due to Theorem 1 (a), then we can find a constant $\bar{c}, \delta > 0$ such that

$$\mathcal{F}(\mathbf{w}^*) - \mathcal{F}(\mathbf{w}_k) \geq \bar{c}\mathcal{F}(\mathbf{w}^* - \mathbf{w}_k) \geq \delta\|\mathbf{w}^* - \mathbf{w}_k\|^2, \; \forall k \geq k_0 \tag{6}$$

for some $k_0 > 0$ (note that $\mathbf{w}_k \to \mathbf{w}^*$ due to Theorem 1 (c)). Together with the inequality from the convexity of $\mathcal{F}$, $\mathcal{F}(\mathbf{w}^*) \geq \mathcal{F}(\mathbf{w}_k) + \langle \mathbf{g}, \mathbf{w}^* - \mathbf{w}_k \rangle$ with $\mathbf{g} \in \partial\mathcal{F}(\mathbf{w}_k)$, it follows that

$$\mathcal{F}(\mathbf{w}^*) \geq \mathcal{F}(\mathbf{w}_k) + \frac{1}{2}\langle \mathbf{g}, \mathbf{w}^* - \mathbf{w}_k \rangle + \frac{\delta}{2}\|\mathbf{w}^* - \mathbf{w}_k\|^2. \tag{7}$$

Comparing to the above inequality of strong convexity, this provides us a weaker condition of strong convexity restricted to the region where (6) holds. This in fact allows us to establish a pointwise convergence rate even in non-strongly convex cases.

Theorem 2. *Let the iterates* $(\mathbf{w}_k, \mathbf{v}_k)$ *be generated by Algorithm 1 with the choices of* τ_0 *and* σ_0 *such that* $2\tau_0\sigma_0 L^2 = 1$*, and* $\mathbf{v}'_{k+1} = 2\mathbf{v}_{k+1}$*. Suppose that the restricted strong convexity* (7) *holds for* $\mathcal{F}$ *with a constant* $\delta > 0$ *about* $\mathbf{w}_k$*,* $\forall k \geq k_0$ *with some* $k_0 > 0$*. Then for a saddle-point* $(\mathbf{w}^*, \mathbf{v}^*)$ *of the GDS-SP problem* (4)*, there exists* $k_1 \geq k_0$ *depending on* $\epsilon \geq 1$ *and* $\delta\tau_0$ *such that for all* $k \geq k_1$*,*

$$\|\mathbf{w}^* - \mathbf{w}_k\|^2 \leq \frac{4\epsilon}{k^2}\left(\frac{\|\mathbf{w}^* - \mathbf{w}_0\|^2}{4\delta^2\tau_0^2} + \frac{L^2}{\delta^2}\|\mathbf{v}^* - \mathbf{v}_0\|^2\right).$$

The proof is provided in Appendix due to its length. In reality, the constant $\delta > 0$ can be very small, probably enough to make the rate similar to $O(1/k)$. Also the condition (6) is not easy to check without knowing $\mathcal{F}(\mathbf{w}^*)$ a priori. Further, it implies $\mathcal{F}(\mathbf{w}^*) \geq \mathcal{F}(\mathbf{w}_k)$ for $k \geq k_0$, which is not enforced by the algorithm. Nonetheless, ours is a new result showing that local pointwise convergence with an accelerated rate $O(1/k^2)$ is possible under the restricted strong convexity. In our experience, Algorithm 1 exhibited pointwise convergence rate as fast as, or even faster than, $O(1/k)$, in surprisingly many cases, even if we chose $\mathbf{v}'_{k+1} = \mathbf{v}_{k+1}$ and $\delta = 0$: this motivated us to check both pointwise and ergodic convergence in non-strongly convex cases.

4 DANTZIG SELECTOR WITH THE ORDERED ℓ_1-NORM

Here we introduce a new kind of the GDS, defined with the ordered ℓ_1-norm: for given parameters $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, the Ordered Dantzig Selector (ODS) performs penalized estimation by solving

$$\text{(ODS)} \qquad \begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \;& J_{\boldsymbol{\lambda}}(\mathbf{w}) = \sum_{i=1}^p \lambda_i |w|_{(i)} \\ \text{s.t.} \;& J_{\boldsymbol{\lambda}*}(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w})) \leq 1 \end{aligned} \tag{8}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$, $|w|_{(i)}$ denotes the ith largest absolute value of the components of the vector $\mathbf{w} = (w_1, \ldots, w_p)$, and $J_{\boldsymbol{\lambda}*}$ is the dual norm of $J_{\boldsymbol{\lambda}}$. It has been shown that $J_{\boldsymbol{\lambda}}(\cdot)$ is indeed a norm (Bogdan et al., 2015; Proposition 1.2). Its dual norm has a rather complicated expression,

$$J_{\boldsymbol{\lambda}*}(\mathbf{w}) = \max\left\{\frac{|w|_{(1)}}{\lambda_1}, \cdots, \frac{\sum_{i=1}^p |w|_{(i)}}{\sum_{i=1}^p \lambda_i}\right\},$$

which we can ignore thanks to the fact that our algorithm only need the proximal operator involving $J_{\boldsymbol{\lambda}}(\cdot)$ to solve its convex-concave saddle-point formulation,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\mathbf{v}\in\mathbb{R}^p} \left\langle \mathbf{X}^T \begin{bmatrix}\mathbf{I} & -\mathbf{X}\end{bmatrix} \begin{bmatrix}\mathbf{y} \\ \mathbf{w}\end{bmatrix}, \mathbf{v} \right\rangle + J_{\boldsymbol{\lambda}}(\mathbf{w}) - J_{\boldsymbol{\lambda}}(\mathbf{v}).$$

The proximal operator for $J_{\boldsymbol{\lambda}}(\cdot)$ can be computed in linear $O(p)$ time using the stack-based FastProxSL1 algorithm (Bogdan et al., 2015; Algorithm 4).

4.1 False Discovery Rate Control

In high-dimensional variable selection, some types of statistical confidence information about selection is desired since otherwise the power of detection of true regressors might be very low or, on the contrary, the number false discoveries can be too large.

In the popular LASSO approach, variable selection is performed based on an ℓ_1-penalized regression,

$$\min_{\mathbf{w}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|_1.$$

When observations follow the model $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}$ with $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ and noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, it is well known that one can choose $\lambda \approx \sigma\sqrt{2\log p}$ to control *family-wise error rate* (FWER), the probability of at least one false rejection. However, this choice is non-adaptive to data as it does not depend on the sparsity and magnitude of the true signal, being likely to result in a loss of power (Bogdan et al., 2015).

In contrast, in an alternative strategy called the SLOPE, where the ℓ_1-regularizer is replaced with the ordered ℓ_1-norm, it has been shown that data-adaptive *false discovery rate (FDR)* control is possible (Bogdan et al., 2015). SLOPE follows the spirit of the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) in multiple hypothesis testing, which can adapt to the unknown signal sparsity with improved asymptotic optimality (Abramovich et al., 2006; Bogdan et al., 2011; Frommlet and Bogdan, 2013; Wu and Zhou, 2013).

Our new proposal, the ODS, shares the same motive as the SLOPE to use the ordered ℓ_1-norm, yet in a different context of Dantzig Selector. Our next theorem shows that similarly to SLOPE, ODS controls FDR for orthogonal design cases.

Theorem 3. *Suppose that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ for $\mathbf{X} \in \mathbb{R}^{n\times p}$ and that $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)$ with*

$$\lambda_i := \Phi^{-1}\left(1 - i\frac{q}{2p}\right)$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. Then the ODS problem (8) has a unique solution $\hat{\mathbf{w}}$ with the FDR controlled at the level

$$FDR = \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right] \le q \cdot \frac{p_0}{p} \le q,$$

where

$$\begin{cases} p_0 & := \left|\{i : w_i^* = 0\}\right| \ \ (\#\ \textit{true null hypotheses}) \\ V & := \left|\{i : w_i^* = 0, \hat{w}_i \neq 0\}\right| \ \ (\#\ \textit{false rejections}) \\ R & := \left|\{i : \hat{w}_i \neq 0\}\right| \ \ (\#\ \textit{all rejections}) \end{cases}$$

Proof. Our proof is based on showing the equivalence between the ODS and SLOPE estimates under the given conditions, and thereby both share the same FDR control. A full proof is provided in Appendix. □

For non-orthogonal design, we may need to use a different sequence of λ_i's. For instance, we can consider a SLOPE adjustment for Gaussian design cases,

$$\begin{cases} \lambda_1' & = \lambda_1 \\ \lambda_i' & = \lambda_i\sqrt{1 + \frac{\sum_{j<i}(\lambda_j')^2}{n-i}}, \ \ i \ge 2, \end{cases}$$
$$\text{and then for } t = \arg\min_i\{\lambda_i'\}, \text{ take} \qquad (9)$$
$$\lambda_i^G = \begin{cases} \lambda_i', & i \le t, \\ \lambda_t & i > t. \end{cases}$$

The second step is required to make the sequence $\{\lambda_i^G\}$ to be non-increasing since otherwise $J_{\boldsymbol{\lambda}^G}(\cdot)$ may not be convex. For more details, we refer to (Bogdan et al., 2015; Section 3.2.2).

5 EXPERIMENTS

We demonstrate our algorithm on the ODS instances on randomly generated data in various settings. Since the ordered ℓ_1-norm is not strongly convex, we run Algorithm 1 with $\gamma = 0$ and $\mathbf{v}_{k+1}' = \mathbf{v}_{k+1}$ unless otherwise specified in this section.

Under the data model $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi}$, we sampled each entry of the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and the noise vector $\boldsymbol{\xi}$ independently from the Gaussian distribution $\mathcal{N}(0, 1)$. The true signal $\mathbf{w} \in \mathbb{R}^p$ was generated to be an s-sparse vector, where the signal strength was set to $w_i = \sqrt{2\log p}$ for all nonzero elements i. The λ_i values were chosen according to Theorem 3 and the adjustment (9), with the target FDR level of $q = 0.1$.

The performance of our algorithm has been compared to alternatives:

Non-SP Algorithms:
− LADMM: linearized ADMM specialized for the GDS (Chatterjee et al., 2014).
− TFOCS: an implementations of Nesterov's smoothing technique (Becker et al., 2011).

SP Algorithms:
− SP-PLX: our proposal method (Algorithm 1).
− SP-HPE: accelerated hybrid proximal extragradient method (He and Monteiro, 2014).
− SP-MiP: a variant of the mirror-prox (Nemirovski, 2004) with linesearch (Tseng, 2008).

Unlike the SP algorithms, the non-SP algorithms require to specify some parameters difficult to determine: in particular, the penalty parameter $\rho \ge \|\mathbf{X}\|^4$ for the LADMM and the smoothing parameter $\mu \approx O(\epsilon)$ for the TFOCS. For LADMM, we estimated $\|\mathbf{X}\|$ using inner products $\langle\mathbf{X}, \mathbf{x}\rangle$ with a random unit vector $\mathbf{x}$. (A similar procedure was used to estimate $L := \|\mathbf{A}\| = \|\mathbf{X}^T\left[\mathbf{I} - \mathbf{X}\right]\|$ in other methods whenever required.) For TFOCS, we fixed $\mu = 10^{-2} \gg \epsilon$ where a larger value than the target optimality is usually recommended for performance.

For the comparison of variable selection quality, we chose the SLOPE approach (Bogdan et al., 2015), which minimizes least squares with the ordered ℓ_1 regularization.

All algorithms were stopped if the following condition was satisfied with the suboptimality tolerance $\epsilon = 10^{-7}$,

$$\|\mathbf{z}_k - \mathbf{z}_{k-1}\| / \max\{1, \|\mathbf{z}_k\|\} \le \epsilon,$$

for either $\mathbf{z}_k = (\mathbf{w}_k, \mathbf{v}_k)$ (pointwise) or $\mathbf{z}_k = (\overline{\mathbf{w}}_k, \overline{\mathbf{v}}_k)$ (ergodic convergence). A tight optimality threshold is typically required for accurate variable selection.

All experiments were performed on a Linux machine with a quadcore 3.20 GHz Intel Xeon CPU and 24 GB of memory, using MATLAB R2015a.

5.1 Algorithm Performance

The primary advantage of our method (SP-PLX) is its fast speed with small variation while being simple. Table 1 compares the runtime of the algorithms over 50 randomly generated ODS instances of different scenarios, i.e., the combinations of ($p < n$, $p = n$, $p > n$) and ($s = 5, 10, 15$).

Our method has been the most favorable over all cases, except for few where SP-HPE performed slightly better. However, the SP-HPE algorithm is far more complicated than ours (see Algorithm 3 and 4 in Appendix), having an iterative subproblem solver which requires to specify extra parameters.

The advantage of SP methods over non-SP counterparts looks apparent. In particular, LADMM, previously proposed for the GDS, performed well for $p < n$, but quite poorly for the other situations. Overall, TFOCS has been slower than LADMM. Note that both LADMM and TFOCS may perform better if their parameters were tuned for individual cases: which is exactly what we tried to avoid.

5.2 FDR Control

To show the FDR control property of the ODS (solved with our algorithm), and its relation to the SLOPE approach, we generated 300 random ODS instance of the dimension $n = 1000$ and $p = 1000$, and compared the power, FDR, and the number of false discoveries of the two approaches for the target FDR level $q = 0.1$.

Figure 1 shows the mean values of these quality criteria over increasing numbers of relevant features (s) in the true signal. Both approaches behaved similarly, although the ODS appeared to be slightly more conservative than the counterpart, improving FDR and the average number of false discoveries at the cost of the small decrease of power. It is important to note here that a smaller number of false discoveries at the nominal FDR level (i.e. preserving relatively high power) is attractive for some applications like finding biomarkers from high-dimensional genomic data, where false positive discoveries can be very costly for follow-up validation.

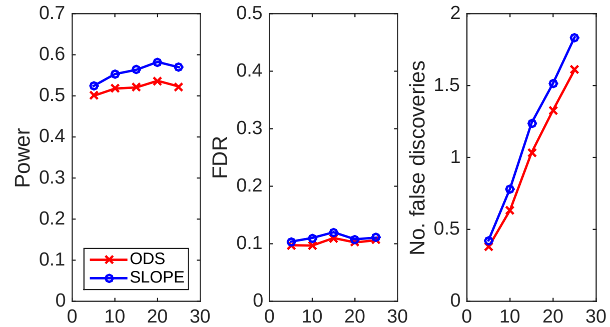

Figure 1: Variable Selection: Power, FDR, and No. of False Discoveries. Mean Values are Shown for Different No. of Relevant Features ($s = 5, 10, 15, 20, 25$).

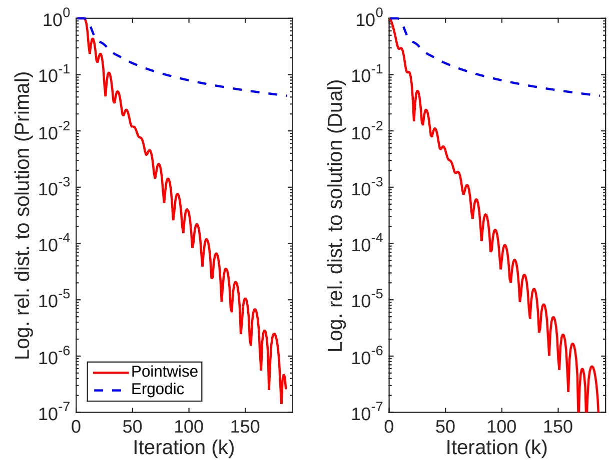

Figure 2: Different Modes in Convergence: Pointwise vs. Ergodic. (Left) Primal; (Right) Dual.

5.3 Convergence Rate

Using our algorithm SP-PLX in experiment, the pointwise convergence was almost always faster than its ergodic convergence. This was quite surprising, as the former was expected to be slower than the latter in general cases, as we discussed earlier.

Figure 2 shows one instance of the randomly generated cases with $p = n = 1000$, $s = 15$, and $q = 0.1$ (behaviors were quite similar in other settings, too). We ran our algorithm twice for the same data, 1) to obtain the primal and dual solutions, then 2) to obtain the relative distances of iterates to their corresponding

Table 1: Runtime of Algorithms Until Convergence with Optimality $\epsilon \leq 10^{-7}$: Mean (Std) Cputime in Seconds for Solving 50 Random ODS Instances.

s	p	n	SP-PLX		SP-HPE		SP-MiP		LADMM		TFOCS	
5	100	1000	0.04	(0.03)	0.05	(0.03)	0.20	(0.12)	0.10	(0.22)	2.92	(4.88)
	1000	1000	1.35	(3.38)	48.96	(339.70)	3.98	(6.56)	15.47	(29.39)	54.43	(291.03)
	1000	100	2.79	(1.63)	2.28	(1.57)	8.15	(4.05)	31.87	(19.99)	20.02	(48.73)
10	100	1000	0.19	(0.40)	54.22	(382.27)	0.74	(1.30)	0.41	(0.53)	14.33	(45.28)
	1000	1000	2.47	(6.07)	1.97	(3.97)	6.31	(11.74)	29.82	(31.77)	37.73	(85.57)
	1000	100	4.99	(5.61)	30.05	(188.19)	12.93	(11.34)	46.78	(24.39)	57.27	(101.36)
15	100	1000	0.33	(0.68)	13.95	(67.75)	1.07	(1.49)	1.32	(1.70)	27.56	(50.32)
	1000	1000	3.99	(8.35)	2.69	(5.18)	9.76	(15.43)	39.52	(32.66)	38.95	(103.08)
	1000	100	9.88	(10.70)	6.93	(8.00)	23.86	(20.82)	91.52	(33.56)	85.77	(124.23)

solution, such as $\|\mathbf{w}_k - \mathbf{w}^*\|/\|\mathbf{w}^*\|$.

As we can see, the averaged iterate (denoted by "Ergodic") showed the expected $O(1/k)$ convergence rate. In contrast, the non-averaged iterates ("pointwise") converged much faster, even exhibiting typical fluctuation patterns of Nesterov's accelerated gradient method. We believe that this behavior is caused by the restricted strong convexity and local acceleration we discussed.

In fact, in Figure 2, neither any information of restricted strong convexity nor the alternative choice of $\mathbf{v}'_{k+1} = 2\mathbf{v}_{k+1}$ was used. When these two options were used, our algorithm showed even faster pointwise convergence, but there were some cases the algorithm did not converge, since the required assumptions might have not been satisfied.

5.4 Restricted Strong Convexity

We again generated 300 random ODS instances with the same settings as in the previous experiment, to simulate how often the restricted strong convexity (7) would be fulfilled, and to what degree.

Figure 3 (top and middle) report the box-plots of the values

$$\frac{\mathcal{F}(\mathbf{w}^*) - \mathcal{F}(\mathbf{w}_k) - \frac{1}{2}\langle \mathbf{g}, \mathbf{w}^* - \mathbf{w}_k \rangle}{\|\mathbf{w}^* - \mathbf{w}_k\|^2}, \quad \mathbf{g} \in \partial \mathcal{F}(\mathbf{w}_k),$$

for the primal, and the equivalent quantify with $\mathcal{G}$ for the dual, evaluated for the last 100 iterations of each run. As we approached the last iteration ($t = 1$), these values varied more away from zero, where the chance of being positive was nearly 50% in the primal and dual, resp. In fact, for acceleration to happen, we only need the values to be positive in either of the two cases: Figure 3 (bottom) shows the chance of such events happening in each iteration during *all* of the last t iterations, which approaches the probability of one as $t \to 1$. This supports our theory, making local acceleration quite likely in many cases.

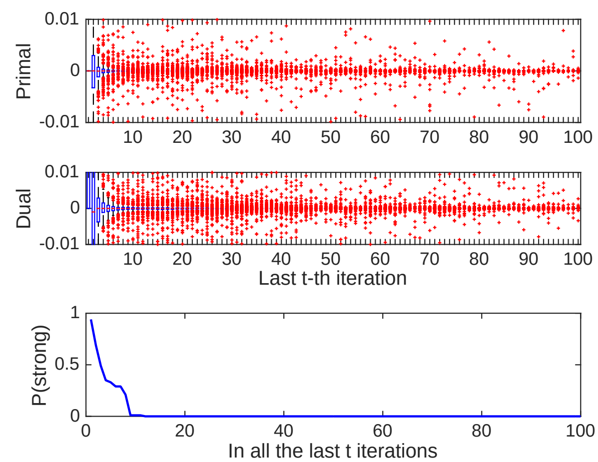

Figure 3: (Top; Middle) Estimation of the Restricted Strong Convexity Constants for $\mathcal{F}$ (Primal) and $\mathcal{G}$ (Dual). (Bottom) Probability of Restricted Strong Convexity.

6 CONCLUSION

We proposed a fast and simple proximal linear extragradient algorithm to solve the generalized Dantzig selector estimation, based on a new convex-concave saddle-point formulation. Our algorithm achieved the optimal convergence rate, while taking the advantage of restricted strong convexity for local acceleration, a new notion that would be useful to find new perspectives of other algorithms. We also introduced the ordered Dantzig selector with FDR control, a new instance of the GDS, which we hope will foster further research in variable selection and signal recovery.

References

F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics*, 34(2):584–653, 2006.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

S. R. Becker, E. J. Candés, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3:165–218, 2011.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1):289–300, 1995.

M. Bogdan, A. Chakrabarti, F. Frommlet, and J. K. Ghosh. Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *Annals of Statistics*, 39(3):1551–1579, 2011.

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candes. SLOPE – adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 2015.

D. Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *SIAM Journal on Optimization*, 23 (4):2183–2207, 2013.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

S. Chatterjee, S. Chen, and A. Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems 27*, pages 1934–1942. 2014.

M. Figueiredo and R. Nowak. Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. *arXiv:1409.4005*, 2014.

F. Frommlet and M. Bogdan. Some optimality properties of fdr controlling rules under sparsity. *Electronic Journal of Statistics*, 7:1328–1368, 2013.

T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge mathematical library. Cambridge University Press, 1994.

Y. He and R. D. C. Monteiro. An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *Optimization Online*, April 2014.

M. Kadkhodaie, K. Christakopoulou, M. Sanjabi, and A. Banerjee. Accelerated alternating direction method of multipliers. In *Proceedings of the 21st ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD, 2015.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.

R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of tseng's modified f-b splitting and korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

A. Nemirovski. 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 Journal on Optimization*, 15(1):229–251, 2004.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Math. Dokl.*, 27(2), 1983.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103:127–152, 2005.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2–3):319–344, 2007.

Y. Ouyang, Y. Chen, G. Lan, and E. P. Jr. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 2015.

R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1997.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *A Series of Comprehensive Studies in Mathematics*. Springer, 2nd edition, 2004.

W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7), 2014.

N. Z. Shor, K. C. Kiwiel, and A. Ruszcayǹski. *Minimization methods for non-differentiable functions*. Springer-Verlag New York, Inc., NY, USA, 1985.

M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4):323–345, 1999a.

M. V. Solodov and B. F. Svaiter. A hybrid projection-proximal point algorithm. *Journal of Convex Analysis*, 6(1), 1999b.

B. Taskar, S. Lacoste-Julien, and M. I. Jordan. Structured prediction, dual extragradient and bregman projections. *Journal of Machine Learning Research*, 7:1627–1653, 2006.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

H. Wang and A. Banerjee. Bregman alternating direction method of multipliers. In *Advances in Neural Information Processing Systems 27*, pages 2816–2824. 2014.

X. Wang and X. Yuan. The linearized alternating direction method of multipliers for dantzig selector. *SIAM Journal on Scientific Computing*, 34(5): A2792–A2811, 2012.

Z. Wu and H. Zhou. Model selection and sharp asymptotic minimaxity. *Probability Theory and Related Fields*, 156(1–2):165–191, 2013.

Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Appendix

Proof of Theorem 1

This proof essentially follows that of Chambolle and Pock (2011; Theorem 1), with a small trick to reformulate the GDS-SP to the formulation discussed in the original proof.

From the definition of the proximal steps in Algorithm 1,

$$\begin{aligned}\mathbf{v}_{k+1} &= (I+\sigma\partial\mathcal{G})^{-1}(\mathbf{v}_k + \sigma(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w}'_k))\\ \mathbf{w}_{k+1} &= (I+\tau\partial\mathcal{F})^{-1}(\mathbf{w}_k + \tau\mathbf{X}^T\mathbf{X}\mathbf{v}'_{k+1}),\end{aligned}$$

it is implied that

$$\begin{aligned}\partial\mathcal{G}(\mathbf{v}_{k+1}) &\ni \frac{\mathbf{v}_k - \mathbf{v}_{k+1}}{\sigma} + \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w}'_k\\ \partial\mathcal{F}(\mathbf{w}_{k+1}) &\ni \frac{\mathbf{w}_k - \mathbf{w}_{k+1}}{\tau} + \mathbf{X}^T\mathbf{X}\mathbf{v}'_{k+1}.\end{aligned} \tag{10}$$

Since $\mathcal{G}$ and $\mathcal{H}$ are convex functions, it follows for any $(\mathbf{w},\mathbf{v})$,

$$\begin{aligned}\mathcal{G}(\mathbf{v}) &\geq \mathcal{G}(\mathbf{v}_{k+1}) + \sigma^{-1}\langle \mathbf{v}_k - \mathbf{v}_{k+1}, \mathbf{v} - \mathbf{v}_{k+1}\rangle\\ &\quad + \langle \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w}'_k, \mathbf{v} - \mathbf{v}_{k+1}\rangle\\ \mathcal{F}(\mathbf{w}) &\geq \mathcal{F}(\mathbf{w}_{k+1}) + \tau^{-1}\langle \mathbf{w}_k - \mathbf{w}_{k+1}, \mathbf{w} - \mathbf{w}_{k+1}\rangle\\ &\quad + \langle \mathbf{X}^T\mathbf{X}(\mathbf{w} - \mathbf{w}_{k+1}), \mathbf{v}'_{k+1}\rangle.\end{aligned}$$

We define augmentations of $\mathbf{w}$'s with $\mathbf{y}$, e.g. $\mathbf{z}_k := [\mathbf{y};\mathbf{w}_k] \in \mathbb{R}^{n+p}$, and $\mathcal{H}(\mathbf{z}) := \mathcal{H}(\mathbf{y},\mathbf{w}) = \mathcal{F}(\mathbf{w})$. Then the preceding inequalities can be rewritten as follows,

$$\begin{aligned}\mathcal{G}(\mathbf{v}) &\geq \mathcal{G}(\mathbf{v}_{k+1}) + \sigma^{-1}\langle \mathbf{v}_k - \mathbf{v}_{k+1}, \mathbf{v} - \mathbf{v}_{k+1}\rangle\\ &\quad + \langle \mathbf{A}\mathbf{z}'_k, \mathbf{v} - \mathbf{v}_{k+1}\rangle\\ \mathcal{H}(\mathbf{z}) &\geq \mathcal{H}(\mathbf{z}_{k+1}) + \tau^{-1}\langle \mathbf{z}_k - \mathbf{z}_{k+1}, \mathbf{z} - \mathbf{z}_{k+1}\rangle\\ &\quad - \langle \mathbf{A}(\mathbf{z} - \mathbf{z}_{k+1}), \mathbf{v}'_{k+1}\rangle.\end{aligned} \tag{11}$$

Summing both inequalities and using an elementary result $\langle a-c, b-c\rangle = \|a-c\|^2/2 + \|b-c\|^2/2 - \|a-b\|^2/2$, we have

$$\begin{aligned}&\frac{\|\mathbf{v}-\mathbf{v}_k\|^2}{2\sigma} + \frac{\|\mathbf{z}-\mathbf{z}_k\|^2}{2\tau}\\ &\geq [\langle \mathbf{A}\mathbf{z}_{k+1}, \mathbf{v}\rangle - \mathcal{G}(\mathbf{v}) + \mathcal{H}(\mathbf{z}_{k+1})]\\ &\quad - [\langle \mathbf{A}\mathbf{z}, \mathbf{v}_{k+1}\rangle - \mathcal{G}(\mathbf{v}_{k+1}) + \mathcal{H}(\mathbf{z})]\\ &\quad + \frac{\|\mathbf{v}-\mathbf{v}_{k+1}\|^2}{2\sigma} + \frac{\|\mathbf{z}-\mathbf{z}_{k+1}\|^2}{2\tau}\\ &\quad + \frac{\|\mathbf{v}_k-\mathbf{v}_{k+1}\|^2}{2\sigma} + \frac{\|\mathbf{z}_k-\mathbf{z}_{k+1}\|^2}{2\tau}\\ &\quad + \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}'_k), \mathbf{v}_{k+1} - \mathbf{v}\rangle\\ &\quad - \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}), \mathbf{v}_{k+1} - \mathbf{v}'_{k+1}\rangle.\end{aligned} \tag{12}$$

Replacing the extragradient steps,

$$\begin{cases}\mathbf{z}'_k &= \mathbf{z}_k + \theta(\mathbf{z}_k - \mathbf{z}_{k-1})\\ \mathbf{v}'_{k+1} &= \mathbf{v}_{k+1},\end{cases}$$

the last two lines of (12) can be bounded as follows,

$$\begin{aligned}&\langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}'_{k+1}, \mathbf{v}_{k+1} - \mathbf{v}\rangle\\ &\geq \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{v}_{k+1} - \mathbf{v}\rangle\\ &\quad - \theta\langle \mathbf{A}(\mathbf{z}_k - \mathbf{z}_{k-1})), \mathbf{v}_k - \mathbf{v}\rangle\\ &\quad - \theta L\|\mathbf{z}_k - \mathbf{z}_{k-1}\|\|\mathbf{v}_{k+1} - \mathbf{v}_k\|\\ &\geq \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{v}_{k+1} - \mathbf{v}\rangle\\ &\quad - \theta\langle \mathbf{A}(\mathbf{z}_k - \mathbf{z}_{k-1})), \mathbf{v}_k - \mathbf{v}\rangle\\ &\quad - \theta^2\sqrt{\sigma\tau}L\frac{\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2}{2\tau} - \sqrt{\sigma\tau}L\frac{\|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2}{2\sigma}\end{aligned}$$

where the first inequality is due to Cauchy-Schwarz and $L := \|\mathbf{A}\|$, and the second is using $|ab| \leq (\alpha/2)a^2 + 1/(2\alpha)b^2$ for any $\alpha > 0$ (in this case $\alpha = \sqrt{\sigma/\tau}$).

Combining this with the full inequality (12), we have

$$\begin{aligned}&\frac{\|\mathbf{v}-\mathbf{v}_k\|^2}{2\sigma} + \frac{\|\mathbf{z}-\mathbf{z}_k\|^2}{2\tau}\\ &\geq [\langle \mathbf{A}\mathbf{z}_{k+1}, \mathbf{v}\rangle - \mathcal{G}(\mathbf{v}) + \mathcal{H}(\mathbf{z}_{k+1})]\\ &\quad - [\langle \mathbf{A}\mathbf{z}, \mathbf{v}_{k+1}\rangle - \mathcal{G}(\mathbf{v}_{k+1}) + \mathcal{H}(\mathbf{z})]\\ &\quad + \frac{\|\mathbf{v}-\mathbf{v}_{k+1}\|^2}{2\sigma} + \frac{\|\mathbf{z}-\mathbf{z}_{k+1}\|^2}{2\tau}\\ &\quad \left(1-\sqrt{\sigma\tau}L\right)\frac{\|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2}{2\sigma}\\ &\quad + \frac{\|\mathbf{z}_k-\mathbf{z}_{k+1}\|^2}{2\tau} - \theta^2\sqrt{\sigma\tau}L\frac{\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2}{2\tau}\\ &\quad + \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{v}_{k+1} - \mathbf{v}\rangle\\ &\quad - \theta\langle \mathbf{A}(\mathbf{z}_k - \mathbf{z}_{k-1})), \mathbf{v}_k - \mathbf{v}\rangle\end{aligned}$$

If we choose parameters so that $\theta = 1$ and $\sigma\tau L^2 \leq 1$, and define

$$\Delta_k := \frac{\|\mathbf{v}-\mathbf{v}_k\|^2}{2\sigma} + \frac{\|\mathbf{z}-\mathbf{z}_k\|^2}{2\tau},$$

it follows that

$$\begin{aligned}\Delta_k &\geq [\langle \mathbf{A}\mathbf{z}_{k+1}, \mathbf{v}\rangle - \mathcal{G}(\mathbf{v}) + \mathcal{H}(\mathbf{z}_{k+1})]\\ &\quad - [\langle \mathbf{A}\mathbf{z}, \mathbf{v}_{k+1}\rangle - \mathcal{G}(\mathbf{v}_{k+1}) + \mathcal{H}(\mathbf{z})]\\ &\quad + \Delta_{k+1}\\ &\quad + \left(1-\sqrt{\sigma\tau}L\right)\frac{\|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2}{2\sigma}\\ &\quad + \frac{\|\mathbf{z}_k-\mathbf{z}_{k+1}\|^2}{2\tau} - \sqrt{\sigma\tau}L\frac{\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2}{2\tau}\\ &\quad + \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{v}_{k+1} - \mathbf{v}\rangle\\ &\quad - \langle \mathbf{A}(\mathbf{z}_k - \mathbf{z}_{k-1})), \mathbf{v}_k - \mathbf{v}\rangle.\end{aligned}$$

Summing up the above inequality for $k = 0$ to $t-1$, with $\mathbf{z}_{-1} = \mathbf{z}_0$ and $\mathbf{v}_{-1} = \mathbf{v}_0$, we get

$$\begin{aligned}\Delta_0 \geq& \sum_{k=1}^{t} \Big\{ [\langle \mathbf{A}\mathbf{z}_k, \mathbf{v}\rangle - \mathcal{G}(\mathbf{v}) + \mathcal{H}(\mathbf{z}_k)] \\ &\quad - [\langle \mathbf{A}\mathbf{z}, \mathbf{v}_k\rangle - \mathcal{G}(\mathbf{v}_k) + \mathcal{H}(\mathbf{z})] \Big\} \\ &+ \Delta_t \\ &+ \big(1 - \sqrt{\sigma\tau}L\big) \sum_{k=1}^{t} \frac{\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2}{2\sigma} \\ &+ \big(1 - \sqrt{\sigma\tau}L\big) \sum_{k=1}^{t-1} \frac{\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2}{2\tau} + \frac{\|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2}{2\tau} \\ &+ \langle \mathbf{A}(\mathbf{z}_t - \mathbf{z}_{t-1}), \mathbf{v}_t - \mathbf{v}\rangle \\ \geq& \sum_{k=1}^{t} \Big\{ [\langle \mathbf{A}\mathbf{z}_k, \mathbf{v}\rangle - \mathcal{G}(\mathbf{v}) + \mathcal{H}(\mathbf{z}_k)] \\ &\quad - [\langle \mathbf{A}\mathbf{z}, \mathbf{v}_k\rangle - \mathcal{G}(\mathbf{v}_k) + \mathcal{H}(\mathbf{z})] \Big\} \\ &+ \Delta_t \\ &+ \big(1 - \sqrt{\sigma\tau}L\big) \sum_{k=1}^{t} \frac{\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2}{2\sigma} \\ &+ \big(1 - \sqrt{\sigma\tau}L\big) \sum_{k=1}^{t-1} \frac{\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2}{2\tau} + \frac{\|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2}{2\tau} \\ &- \frac{\|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2}{2\tau} - \tau L^2 \frac{\|\mathbf{v}_t - \mathbf{v}\|^2}{2}.\end{aligned}$$

That is, for any $(\mathbf{z}, \mathbf{v})$,

$$\begin{aligned}&(1 - \tau\sigma L^2)\frac{\|\mathbf{v} - \mathbf{v}_t\|^2}{\sigma} + \frac{\|\mathbf{z} - \mathbf{z}_t\|^2}{\tau} \\ &\quad + 2\sum_{k=1}^{t} \Big\{ [\langle \mathbf{A}\mathbf{z}_k, \mathbf{v}\rangle - \mathcal{G}(\mathbf{v}) + \mathcal{H}(\mathbf{z}_k)] \\ &\quad - [\langle \mathbf{A}\mathbf{z}, \mathbf{v}_k\rangle - \mathcal{G}(\mathbf{v}_k) + \mathcal{H}(\mathbf{z})] \Big\} \\ &\quad + \big(1 - \sqrt{\sigma\tau}L\big) \sum_{k=1}^{t} \frac{\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2}{\sigma} \\ &\quad + \big(1 - \sqrt{\sigma\tau}L\big) \sum_{k=1}^{t-1} \frac{\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2}{\tau} \\ &\leq \frac{\|\mathbf{v} - \mathbf{v}_0\|^2}{\sigma} + \frac{\|\mathbf{z} - \mathbf{z}_0\|^2}{\tau}.\end{aligned} \tag{13}$$

For any saddle point $(\mathbf{z}^*, \mathbf{v}^*)$ satisfying the conditions in (5), we observe that

$$\begin{aligned}&[\langle \mathbf{A}\mathbf{z}_k, \mathbf{v}^*\rangle - \mathcal{G}(\mathbf{v}^*) + \mathcal{H}(\mathbf{z}_k)] \\ &\geq [\langle \mathbf{A}\mathbf{z}^*, \mathbf{v}^*\rangle - \mathcal{G}(\mathbf{v}^*) + \mathcal{H}(\mathbf{z}^*)] \\ &\geq [\langle \mathbf{A}\mathbf{z}^*, \mathbf{v}_k\rangle - \mathcal{G}(\mathbf{v}_k) + \mathcal{H}(\mathbf{z}^*)].\end{aligned} \tag{14}$$

That is, for $(\mathbf{z}, \mathbf{v}) = (\mathbf{z}^*, \mathbf{v}^*)$ the first summation in (13) is non-negative, and therefore $(\mathbf{z}_k, \mathbf{v}_k)$ is bounded, showing the property (a).

Also, from (13) due the convexity of $\mathcal{G}$ and $\mathcal{H}$, we have the following result for $\overline{\mathbf{z}}_t = \frac{1}{t}\sum_{k=1}^{t} \mathbf{z}_k$ and $\overline{\mathbf{v}}_t = \frac{1}{t}\sum_{k=1}^{t} \mathbf{v}_k$,

$$\begin{aligned}&[\langle \mathbf{A}\overline{\mathbf{z}}_t, \mathbf{v}\rangle - \mathcal{G}(\mathbf{v}) + \mathcal{H}(\overline{\mathbf{z}}_t)] - [\langle \mathbf{A}\mathbf{z}, \overline{\mathbf{v}}_t\rangle - \mathcal{G}(\overline{\mathbf{v}}_t) + \mathcal{H}(\mathbf{z})] \\ &\leq \frac{1}{t}\left(\frac{\|\mathbf{v} - \mathbf{v}_0\|^2}{2\sigma} + \frac{\|\mathbf{z} - \mathbf{z}_0\|^2}{2\tau}\right) \\ &\leq \frac{1}{t}\left(\frac{\|\mathbf{v} - \mathbf{v}^*\|^2}{2\sigma} + \frac{\|\mathbf{v}^* - \mathbf{v}_0\|^2}{2\sigma} \right. \\ &\quad \left. + \frac{\|\mathbf{z} - \mathbf{z}^*\|^2}{2\tau} + \frac{\|\mathbf{z}^* - \mathbf{z}_0\|^2}{2\tau}\right) \\ &\leq \frac{1 + C}{t}\left(\frac{\|\mathbf{v}^* - \mathbf{v}_0\|^2}{2\sigma} + \frac{\|\mathbf{z}^* - \mathbf{z}_0\|^2}{2\tau}\right),\end{aligned} \tag{15}$$

where $C = 1/(1 - \tau\sigma L^2)$ and the last inequality is due to the previous result (a). This shows the first part of (b).

Furthermore, with $t \to \infty$, (15) implies for any limit point $(\hat{\mathbf{z}}, \hat{\mathbf{v}})$ of $(\overline{\mathbf{z}}_t, \overline{\mathbf{v}}_t)$ that

$$[\langle \mathbf{A}\hat{\mathbf{z}}, \mathbf{v}\rangle - \mathcal{G}(\mathbf{v}) + \mathcal{H}(\hat{\mathbf{z}})] - [\langle \mathbf{A}\mathbf{z}, \hat{\mathbf{v}}\rangle - \mathcal{G}(\hat{\mathbf{v}}) + \mathcal{H}(\mathbf{z})] \leq 0$$

since $\mathcal{H}$ and $\mathcal{G}$ are l.s.c., i.e.

$$\liminf_{t\to\infty} \mathcal{H}(\overline{\mathbf{z}}_t) \geq \mathcal{H}(\hat{\mathbf{z}}), \quad \liminf_{t\to\infty} \mathcal{G}(\overline{\mathbf{v}}_t) \geq \mathcal{G}(\hat{\mathbf{v}}).$$

This implies that $(\hat{\mathbf{z}}, \hat{\mathbf{v}})$ is a saddle-point of (4), the second claim in (b).

Finally, since $(1 - \sqrt{\sigma\tau}L) \geq 0$, (13) also implies that $\|\mathbf{v}_k - \mathbf{v}_{k-1}\| \to 0$ and $\|\mathbf{z}_k - \mathbf{z}_{k-1}\| \to 0$ as $k \to \infty$, and therefore $(\mathbf{z}_k, \mathbf{v}_k)$ should have a limit, say $(\tilde{\mathbf{z}}, \tilde{\mathbf{v}})$. Then again from (13), we have at the limit

$$[\langle \mathbf{A}\tilde{z}, \mathbf{v}^*\rangle - \mathcal{G}(\mathbf{v}^*) + \mathcal{H}(\tilde{z})] - [\langle \mathbf{A}\mathbf{z}^*, \tilde{\mathbf{v}}\rangle - \mathcal{G}(\tilde{\mathbf{v}}) + \mathcal{H}(\mathbf{z}^*)] = 0,$$

which tells that $(\tilde{\mathbf{z}}, \tilde{\mathbf{v}})$ is also a saddle-point, showing the last claim (c). □

Proof of Theorem 2

Hence $\mathcal{F}$ satisfies the restricted strong convexity (7), so does the function $\mathcal{H}(\mathbf{z}) = \mathcal{H}(\mathbf{y}, \mathbf{w}) = \mathcal{F}(\mathbf{w})$ on augmented vectors $\mathbf{z} = (\mathbf{y}, \mathbf{w})$, and from the expression of subgradients (10) we have (with $\tau_k' = 2\tau_k$)

$$\begin{aligned}\mathcal{H}(\mathbf{z}^*) \geq& \mathcal{H}(\mathbf{z}_{k+1}) + \left\langle \frac{\mathbf{z}_k - \mathbf{z}_{k+1}}{\tau_k'}, \mathbf{z}^* - \mathbf{z}_{k+1} \right\rangle \\ &- \left\langle \mathbf{A}(\mathbf{z}^* - \mathbf{z}_{k+1}), \frac{\mathbf{v}_{k+1}'}{2} \right\rangle + \frac{\delta}{2}\|\mathbf{z}^* - \mathbf{z}_{k+1}\|^2.\end{aligned}$$

Also, from the convexity of $\mathcal{G}$ (11),

$$\begin{aligned}\mathcal{G}(\mathbf{v}^*) \geq& \mathcal{G}(\mathbf{v}_{k+1}) + \left\langle \frac{\mathbf{v}_k - \mathbf{v}_{k+1}}{\sigma_k}, \mathbf{v}^* - \mathbf{v}_{k+1} \right\rangle \\ &+ \langle \mathbf{A}\mathbf{z}_k', \mathbf{v}^* - \mathbf{v}_{k+1}\rangle.\end{aligned}$$

With these, the previous inequality (12) modifies as follows,

$$\begin{aligned}
&\frac{\|\mathbf{v}^* - \mathbf{v}_k\|^2}{2\sigma_k} + \frac{\|\mathbf{z}^* - \mathbf{z}_k\|^2}{2\tau'_k} \\
&\geq [\langle \mathbf{A}\mathbf{z}_{k+1}, \mathbf{v}^*\rangle - \mathcal{G}(\mathbf{v}^*) + \mathcal{H}(\mathbf{z}_{k+1})] \\
&\quad - [\langle \mathbf{A}\mathbf{z}^*, \mathbf{v}_{k+1}\rangle - \mathcal{G}(\mathbf{v}_{k+1}) + \mathcal{H}(\mathbf{z}^*)] \\
&\quad + \frac{\delta}{2}\|\mathbf{z}^* - \mathbf{z}_{k+1}\|^2 + \frac{\|\mathbf{v}^* - \mathbf{v}_{k+1}\|^2}{2\sigma_k} + \frac{\|\mathbf{z}^* - \mathbf{z}_{k+1}\|^2}{2\tau'_k} \\
&\quad + \frac{\|\mathbf{v}_k - \mathbf{v}_{k+1}\|^2}{2\sigma_k} + \frac{\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2}{2\tau'_k} \\
&\quad + \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}'_k), \mathbf{v}_{k+1} - \mathbf{v}^*\rangle \\
&\quad - \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}^*), \mathbf{v}_{k+1} - \frac{\mathbf{v}'_{k+1}}{2}\rangle.
\end{aligned}$$

From (14), the expression in the first two terms of the right-hand side is bounded below by zero.

Choosing the extragradient steps as follows,

$$\begin{cases} \mathbf{z}'_k &= \mathbf{z}_k + \theta_{k-1}(\mathbf{z}_k - \mathbf{z}_{k-1}) \\ \mathbf{v}'_{k+1} &= 2\mathbf{v}_{k+1}, \end{cases}$$

it follows that

$$\begin{aligned}
&\frac{\|\mathbf{v}^* - \mathbf{v}_k\|^2}{2\sigma_k} + \frac{\|\mathbf{z}^* - \mathbf{z}_k\|^2}{2\tau'_k} \\
&\geq \frac{\delta}{2}\|\mathbf{z}^* - \mathbf{z}_{k+1}\|^2 + \frac{\|\mathbf{v}^* - \mathbf{v}_{k+1}\|^2}{2\sigma_k} + \frac{\|\mathbf{z}^* - \mathbf{z}_{k+1}\|^2}{2\tau'_k} \\
&\quad + \frac{\|\mathbf{v}_k - \mathbf{v}_{k+1}\|^2}{2\sigma_k} + \frac{\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2}{2\tau'_k} \\
&\quad + \langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{v}_{k+1} - \mathbf{v}^*\rangle \\
&\quad - \theta_{k-1}\langle \mathbf{A}(\mathbf{z}_k - \mathbf{z}_{k-1}), \mathbf{v}_k - \mathbf{v}^*\rangle \\
&\quad - \theta_{k-1}L\|\mathbf{z}_k - \mathbf{z}_{k-1}\|\|\mathbf{v}_{k+1} - \mathbf{v}_k\|.
\end{aligned}$$

The rest of the proof is very similar to that of Section 5 of Chambolle and Pock (2011), from eq. (39) to (42) therein. From the above, we have

$$\begin{aligned}
&\frac{\|\mathbf{v}^* - \mathbf{v}_k\|^2}{\sigma_k} + \frac{\|\mathbf{z}^* - \mathbf{z}_k\|^2}{\tau'_k} \\
&\geq (1 + \delta\tau'_k)\frac{\tau'_{k+1}}{\tau'_k}\frac{\|\mathbf{z}^* - \mathbf{z}_{k+1}\|^2}{\tau'_{k+1}} + \frac{\sigma_{k+1}}{\sigma_k}\frac{\|\mathbf{v}^* - \mathbf{v}_{k+1}\|^2}{\sigma_{k+1}} \\
&\quad + \frac{\|\mathbf{v}_k - \mathbf{v}_{k+1}\|^2}{\sigma_k} + \frac{\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2}{\tau'_k} - \frac{\|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2}{\sigma_k} \\
&\quad + 2\langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{v}_{k+1} - \mathbf{v}^*\rangle \\
&\quad - 2\theta_{k-1}\langle \mathbf{A}(\mathbf{z}_k - \mathbf{z}_{k-1}), \mathbf{v}_k - \mathbf{v}^*\rangle \\
&\quad - \theta_{k-1}^2 L^2\sigma_k\tau'_{k-1}\frac{\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2}{\tau'_{k-1}}.
\end{aligned}$$

We choose sequences τ'_k and σ_k such that

$$(1 + \delta\tau'_k)\frac{\tau'_{k+1}}{\tau'_k} = \frac{\sigma_{k+1}}{\sigma_k} = \frac{1}{\theta_k} = \frac{\tau'_k}{\tau'_{k+1}} > 1.$$

Denoting

$$\Delta_k := \frac{\|\mathbf{v}^* - \mathbf{v}_k\|^2}{\sigma_k} + \frac{\|\mathbf{z}^* - \mathbf{z}_k\|^2}{\tau'_k},$$

dividing the both sides of the above inequality by τ'_k, and using $L^2\sigma_k\tau'_k = L^2\sigma_0\tau'_0 \leq 1$, we get

$$\begin{aligned}
\frac{\Delta_k}{\tau'_k} &\geq \frac{\Delta_{k+1}}{\tau'_{k+1}} + \frac{\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2}{(\tau'_k)^2} - \frac{\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2}{(\tau'_{k-1})^2} \\
&\quad + \frac{2}{\tau'_k}\langle \mathbf{A}(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{v}_{k+1} - \mathbf{v}^*\rangle \\
&\quad - \frac{2}{\tau'_{k-1}}\langle \mathbf{A}(\mathbf{z}_k - \mathbf{z}_{k-1}), \mathbf{v}_k - \mathbf{v}^*\rangle.
\end{aligned}$$

Applying this inequality for $k = 0, \cdots, (t-1)$, and using $\mathbf{z}_{-1} = \mathbf{z}_0$, it leads to

$$\begin{aligned}
\frac{\Delta_0}{\tau'_0} &\geq \frac{\Delta_t}{\tau'_t} + \frac{\|\mathbf{z}_{t-1} - \mathbf{z}_t\|^2}{(\tau'_{t-1})^2} \\
&\quad + \frac{2}{\tau'_{t-1}}\langle \mathbf{A}(\mathbf{z}_t - \mathbf{z}_{t-1}), \mathbf{v}_t - \mathbf{v}^*\rangle \\
&\geq \frac{\Delta_t}{\tau'_t} + \frac{\|\mathbf{z}_{t-1} - \mathbf{z}_t\|^2}{(\tau'_{t-1})^2} \\
&\quad - \frac{\|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2}{(\tau'_{t-1})^2} - L^2\|\mathbf{v}_t - \mathbf{v}^*\|^2.
\end{aligned}$$

Rearranging terms, using $L^2\sigma_k\tau'_k = L^2\sigma_0\tau'_0$, and replacing $\tau'_k = 2\tau_k$, we finally obtain

$$\begin{aligned}
&4\tau_t^2\frac{1 - 2L^2\sigma_0\tau_0}{2\sigma_0\tau_0}\|\mathbf{v}^* - \mathbf{v}_t\|^2 + \|\mathbf{z}^* - \mathbf{z}_t\|^2 \\
&\leq 4\tau_t^2\left(\frac{\|\mathbf{z}^* - \mathbf{z}_0\|^2}{4\tau_0^2} + \frac{\|\mathbf{v}^* - \mathbf{v}_0\|^2}{2\sigma_0\tau_0}\right).
\end{aligned}$$

If we choose σ_0, τ_0 so that $2\sigma_0\tau_0L^2 = 1$, then

$$\|\mathbf{z}^* - \mathbf{z}_t\|^2 \leq 4\tau_t^2\left(\frac{\|\mathbf{z}^* - \mathbf{z}_0\|^2}{4\tau_0^2} + L^2\|\mathbf{v}^* - \mathbf{v}_0\|^2\right).$$

The proof completes if we show $\tau_t \sim t^{-1}$ for all $t \geq t_0$. Note that our choice of τ_k satisfies (with $\tau'_k = 2\tau_k$),

$$(1 + 2\delta\tau_k)\frac{\tau_{k+1}}{\tau_k} = \frac{\tau_k}{\tau_{k+1}},$$

which is an identical choice to that of Lemma 1 of Chambolle and Pock (2011), replacing $\gamma = \delta$ therein. Therefore Corollary 1 of Chambolle and Pock (2011) applies as it is, which shows that $\lim_{t\to\infty} t\delta\tau_t = 1$. □

FDR Control of the ODS

Properties of $J_{\boldsymbol{\lambda}}$ and $J_{\boldsymbol{\lambda}^*}$

Proposition 1. *If* $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^p$ *are such that* $|\mathbf{a}| \preceq |\mathbf{b}|$, *then the vectors sorted in decreasing component magnitudes satisfy* $|\mathbf{a}|_{(\cdot)} \preceq |\mathbf{b}|_{(\cdot)}$.

Proof. Without loss of generality assume that $\mathbf{a}$ and $\mathbf{b}$ are nonnegative and that $a_1 \geq \ldots \geq a_p$. We will show that $a_k \leq b_{(k)}$ for $k \in \{1, \ldots, p\}$. Fix such k and consider the set $S_k := \{b_i : \ b_i \geq a_k\}$. It is enough to show that $|S_k| \geq k$, where $|S_k|$ is the number of elements in S_k. For each $j \in \{1, \ldots, k\}$ we have

$$b_j \geq a_j \geq a_k \implies b_j \in S_k,$$

what proves the last statement. □

Corollary 1. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ and $|\mathbf{a}| \preceq |\mathbf{b}|$. Then $J_{\boldsymbol{\lambda}}(\mathbf{a}) \leq J_{\boldsymbol{\lambda}}(\mathbf{b})$ since $J_{\boldsymbol{\lambda}}(\mathbf{a}) = \boldsymbol{\lambda}^T|\mathbf{a}|_{(\cdot)} \leq \boldsymbol{\lambda}^T|\mathbf{b}|_{(\cdot)} = J_{\boldsymbol{\lambda}}(\mathbf{b})$.*

Proposition 2 (Pulling to zero)**.** *For fixed sequence $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$, let $\mathbf{w} \in \mathbb{R}^p$ be such that $w_j > 0$ for some $j \in \{1, \ldots, p\}$. For $\varepsilon \in (0, w_j]$ define:*

$$(w_\varepsilon)_i = \begin{cases} w_j - \varepsilon, & i = j \\ w_i, & \textit{otherwise} \end{cases}.$$

Then:
(i) $J_\lambda(\mathbf{w}_\varepsilon) \leq J_\lambda(\mathbf{w})$,
(ii) if $\boldsymbol{\lambda} \succ 0$, then $J_\lambda(\mathbf{w}_\varepsilon) < J_\lambda(\mathbf{w})$.

Proof. Let $\pi : \{1, \ldots, p\} \longrightarrow \{1, \ldots, p\}$ be permutation such as $\sum_{i=1}^p \lambda_i (w_\varepsilon)_{(i)} = \sum_{i=1}^p \lambda_{\pi(i)} (w_\varepsilon)_i$ for each i in $\{1, \ldots, p\}$. Using the rearrangement inequality (Hardy et al., 1994)

$$J_{\boldsymbol{\lambda}}(\mathbf{w}) - J_{\boldsymbol{\lambda}}(\mathbf{w}_\varepsilon) = \sum_{i=1}^p \lambda_i w_{(i)} - \sum_{i=1}^p \lambda_{\pi(i)} (w_\varepsilon)_i$$
$$\geq \sum_{i=1}^p \lambda_{\pi(i)} w_i - \sum_{i=1}^p \lambda_{\pi(i)} (w_\varepsilon)_i = \varepsilon \lambda_{\pi(j)} \geq 0.$$

If $\boldsymbol{\lambda} \succ 0$, then the last inequality is strict. □

Proposition 3 (Pulling to the mean)**.** *For fixed sequence $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$, let $\mathbf{w} \in \mathbb{R}^p$ be such that $\mathbf{w} \succeq 0$ and $w_j > w_l$ for some $j, l \in \{1, \ldots, p\}$. For $0 < \varepsilon \leq \frac{w_j - w_l}{2}$ define:*

$$(w_\varepsilon)_i = \begin{cases} w_l + \varepsilon, & i = l \\ w_j - \varepsilon, & i = j \\ w_i, & \textit{otherwise} \end{cases}.$$

Then:
(i) $J_{\boldsymbol{\lambda}}(\mathbf{w}_\varepsilon) \leq J_{\boldsymbol{\lambda}}(\mathbf{w})$,
(ii) if $\lambda_1 > \ldots > \lambda_p$, then $J_{\boldsymbol{\lambda}}(\mathbf{w}_\varepsilon) < J_{\boldsymbol{\lambda}}(\mathbf{w})$.

Proof. Let $\pi : \{1, \ldots, p\} \longrightarrow \{1, \ldots, p\}$ be permutation such as $\sum_{i=1}^p \lambda_i (w_\varepsilon)_{(i)} = \sum_{i=1}^p \lambda_{\pi(i)} (w_\varepsilon)_i$ for each i in $\{1, \ldots, p\}$ and $\lambda_{\pi(j)} \geq \lambda_{\pi(l)}$. From the rearrangement inequality,

$$\begin{aligned} &J_{\boldsymbol{\lambda}}(\mathbf{w}) - J_{\boldsymbol{\lambda}}(\mathbf{w}_\varepsilon) \\ &= \sum_{i=1}^p \lambda_i w_{(i)} - \sum_{i=1}^p \lambda_i (w_\varepsilon)_{(i)} \\ &= \sum_{i=1}^p \lambda_i w_{(i)} - \sum_{i=1}^p \lambda_{\pi(i)} (w_\varepsilon)_i \\ &\geq \sum_{i=1}^p \lambda_{\pi(i)} w_i - \sum_{i=1}^p \lambda_{\pi(i)} (w_\varepsilon)_i \\ &= \varepsilon\big(\lambda_{\pi(j)} - \lambda_{\pi(l)}\big) \geq 0. \end{aligned}$$

If $\lambda_1 > \ldots > \lambda_p$, then the last inequality is strict. □

Proposition 4. *Suppose that $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. For arbitrary $\mathbf{x} \in \mathbb{R}^p$, $\varepsilon > 0$, and $l, j \in \{1, \ldots, p\}$ define:*

$$(x_\varepsilon)_i = \begin{cases} x_l + \varepsilon, & i = l \\ x_j - \varepsilon, & i = j \\ x_i, & \textit{otherwise} \end{cases},$$

$$(\widetilde{x}_\varepsilon)_i = \begin{cases} x_j - \varepsilon, & i = j \\ x_i, & \textit{otherwise} \end{cases}.$$

Then:
i) if $x_j > x_l \geq 0$, then for $\varepsilon \in \big(0, (x_j - x_l)/2\big]$ it holds that $J_{\boldsymbol{\lambda}*}(\mathbf{x}_\varepsilon) \leq J_{\boldsymbol{\lambda}*}(\mathbf{x})$,
ii) if $x_j > 0$, then for $\varepsilon \in \big(0, x_j\big]$ it holds $J_{\boldsymbol{\lambda}*}(\widetilde{\mathbf{x}}_\varepsilon) \leq J_{\boldsymbol{\lambda}*}(\mathbf{x})$.

Proof. It is easy to observe that for any $\mathbf{x} \in \mathbb{R}^p$, the dual to the sorted ℓ_1-norm could be represented as $J_{\boldsymbol{\lambda}*}(\mathbf{x}) = \max\Big\{J_{\lambda^k}(x),\ k \leq p\Big\}$ for

$$\lambda_i^k := \begin{cases} \big(\sum_{j=1}^k \lambda_j\big)^{-1}, & i \leq k \\ 0, & \text{otherwise} \end{cases}.$$

The claim is therefore the straightforward consequence of Propositions 2 and 3. □

Proposition 5. *If $|\mathbf{w}| \preceq |\widetilde{\mathbf{w}}|$, then $J_{\boldsymbol{\lambda}*}(\mathbf{w}) \leq J_{\boldsymbol{\lambda}*}(\widetilde{\mathbf{w}})$.*

Proof. Claim follows simply from Corollary 1, analogously as in proof of previous proposition. □

Properties of the ODS with an Orthogonal Design

Before proving Theorem 3, we will recall results concerning the SLOPE problem,

$$\min_{\mathbf{w}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + J_{\boldsymbol{\lambda}}(\mathbf{w}). \tag{16}$$

and derive some properties useful in analysis of the ODS problem (8). Since $\frac{1}{2}\big\|\mathbf{y} - \mathbf{X}\mathbf{w}\big\|_2^2 = \frac{1}{2}\big\|\widetilde{\mathbf{y}} - \mathbf{w}\big\|_2^2 +$

const and $\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w}) = \widetilde{\mathbf{y}} - \mathbf{w}$ under orthogonal design, for $\widetilde{\mathbf{y}} := \mathbf{X}^T\mathbf{y}$, without loss of generality we will consider the case $\mathbf{X} = \mathbf{I}_p$ (hereafter we also denote $\tilde{\mathbf{y}}$ by $\mathbf{y}$ to simplify notations). Let I_1 be some subset of $\{1,\ldots,p\}$ and $\overline{\mathbf{w}_{I_1}}$ denotes the arithmetic mean of subvector $\mathbf{w}_{I_1}$, for any $\mathbf{w} \in \mathbb{R}^p$. Bogdan et al. (2015) showed that the unique solution to (16) (uniqueness follows from strict convexity of objective function) is output of the following Algorithm 2, which terminates in at most n steps.

The same authors showed that (16) can be casted as a quadratic program,

$$\begin{array}{ll} \min_{\mathbf{w}\in\mathbb{R}^p} & \frac{1}{2}\|\mathbf{y}-\mathbf{w}\|_2^2 + \boldsymbol{\lambda}^T\mathbf{w} \\ \text{subject to} & w_1 \geq \ldots \geq w_p \geq 0 \ . \end{array}$$

We aim to show that the ODS (8) can be reformulated as a linear program, and then to show that its solution is unique and can be given by the same Algorithm 2 for SLOPE.

Taking into account that $\mathbf{X} = \mathbf{I}_p$, the ODS for the orthogonal design can be rewritten as

$$\min_{\mathbf{w}} \ J_{\boldsymbol{\lambda}}(\mathbf{w}) \quad \text{subject to} \quad \mathbf{y}-\mathbf{w} \in C_{\boldsymbol{\lambda}}, \tag{17}$$

where $C_{\boldsymbol{\lambda}}$ is the unit ball in terms of the dual norm, $J_{\boldsymbol{\lambda}*}$. It is easy to see that if $\mathbf{y}-\mathbf{w} \in C_{\boldsymbol{\lambda}}$, then also $\mathbf{P}_\pi(\mathbf{y}-\mathbf{w}) \in C_{\boldsymbol{\lambda}}$ and $\mathbf{S}(\mathbf{y}-\mathbf{w}) \in C_{\boldsymbol{\lambda}}$ for any permutation π and diagonal matrix $\mathbf{S}$ with $|S_{ii}| = 1$, $i = 1,\ldots,p$.

Proposition 6. *Suppose that $\mathbf{w}^*$ is solution to (17), π is arbitrary permutation of $\{1,\ldots,p\}$ and $\mathbf{S}$ is diagonal matrix such as $|S_{ii}| = 1$ for all i. Then*
i) $\mathbf{w}^*_\pi := \mathbf{P}_\pi\mathbf{w}^*$ is a solution to

$$\min_{\mathbf{w}} J_{\boldsymbol{\lambda}}(\mathbf{w}) \ s.t. \ \mathbf{P}_\pi\mathbf{y}-\mathbf{w} \in C_{\boldsymbol{\lambda}},$$

ii) $\mathbf{w}^*_{\mathbf{S}} := \mathbf{S}\mathbf{w}^*$ is a solution to

$$\min_{\mathbf{w}} J_{\boldsymbol{\lambda}}(\mathbf{w}) \quad s.t. \ \mathbf{S}\mathbf{y}-\mathbf{w} \in C_{\boldsymbol{\lambda}}.$$

Proof. Suppose that there exists $\mathbf{w}_0 \in \mathbb{R}^p$ such as $J_{\boldsymbol{\lambda}}(\mathbf{w}_0) < J_{\boldsymbol{\lambda}}(\mathbf{w}^*_\pi)$ and $\mathbf{P}_\pi\mathbf{y}-\mathbf{w}_0 \in C_{\boldsymbol{\lambda}}$. Then we have $\mathbf{y}-\mathbf{P}_\pi^{-1}\mathbf{w}_0 \in C_{\boldsymbol{\lambda}}$ and $J_{\boldsymbol{\lambda}}(\mathbf{P}_\pi^{-1}\mathbf{w}_0) < J_{\boldsymbol{\lambda}}(\mathbf{w}^*)$, which contradicts the optimality of $\mathbf{w}^*$. The second part can be shown similarly. □

Thanks to Proposition 6, without loss of generality we assume that $y_1 \geq \ldots \geq y_p \geq 0$. Indeed, if for an arbitrary $\mathbf{y} \in \mathbb{R}^p$, the solution $\mathbf{w}^{|\mathbf{y}|_{(\cdot)}}$ to (17) for ordered magnitudes of observations is known, the original solution could be immediately recovered by $\mathbf{w}^* = \mathbf{S}\mathbf{P}_\pi\mathbf{w}^{|\mathbf{y}|_{(\cdot)}}$ with $\mathbf{P}_\pi$ and $\mathbf{S}$ satisfying $\mathbf{y} = \mathbf{S}\mathbf{P}_\pi|\mathbf{y}|_{(\cdot)}$. This coincides with the analogous property of the SLOPE:

Proposition 7. *Assume that $\lambda_1 > \ldots > \lambda_p > 0$, $y_1 \geq \ldots \geq y_p \geq 0$ and let $\mathbf{w}^*$ be solution to (17). Then, for any $j \in \{1,\ldots,p\}$ we have $0 \leq w^*_j \leq y_j$.*

Proof. Suppose first that for some j it occurs $w^*_j < 0$ and define

$$(w_\varepsilon)_i := \begin{cases} |w^*_j| - \varepsilon, & i = j \\ |w^*_i|, & \text{otherwise} \end{cases} .$$

Fix $\varepsilon = |w^*_j|$. Then $|\mathbf{y}-\mathbf{w}_\varepsilon| \preceq |\mathbf{y}-\mathbf{w}^*|$, hence Proposition 5 yields $J_{\boldsymbol{\lambda}*}(\mathbf{y}-\mathbf{w}_\varepsilon) \leq J_{\boldsymbol{\lambda}*}(\mathbf{y}-\mathbf{w}^*) \leq 1$ and $\mathbf{w}_\varepsilon$ is feasible. Moreover, Proposition 2 gives $J_{\boldsymbol{\lambda}}(\mathbf{w}_\varepsilon) < J_{\boldsymbol{\lambda}}(|\mathbf{w}^*|) = J_{\boldsymbol{\lambda}}(\mathbf{w}^*)$, which leads to contradiction.

Suppose now that $w^*_j > y_j$. This gives that $w^*_j > 0$. Define

$$(w_\varepsilon)_i := \begin{cases} w^*_j - \varepsilon, & i = j \\ w^*_i, & \text{otherwise} \end{cases} ,$$

and fix $\varepsilon := w^*_j - y_j$. As before $\mathbf{w}_\varepsilon$ is feasible. Using again Proposition 2, we get $J_{\boldsymbol{\lambda}}(\mathbf{w}_\varepsilon) < J_{\boldsymbol{\lambda}}(\mathbf{w}^*)$ which contradicts the optimality of $\mathbf{w}^*$. □

Proposition 8. *Assume that $\lambda_1 > \ldots > \lambda_p > 0$, $y_1 \geq \ldots \geq y_p \geq 0$ and let $\mathbf{w}^*$ be solution to (17). Then it occurs that $w^*_1 \geq \ldots \geq w^*_p \geq 0$ and $y_1 - w^*_1 \geq \ldots \geq y_p - w^*_p \geq 0$.*

Proof. From the previous propositions we have that $\mathbf{w}^* \succeq \mathbf{0}$ and $\mathbf{y}-\mathbf{w}^* \succeq \mathbf{0}$. We will show that for $1 \leq j < l \leq p$ it holds $w^*_j \geq w^*_l$ and $y^*_j - w^*_j \geq y_l - w^*_l$. Suppose first that $w^*_j < w^*_l$ for some $j < l$ and denote $\mathbf{x} := \mathbf{y}-\mathbf{w}^*$. Since $y_j \geq y_l$, we have $x_j > x_l$. Define

$$(w_\varepsilon)_i = \begin{cases} w^*_j + \varepsilon, & i = j \\ w^*_l - \varepsilon, & i = l \\ w^*_i, & \text{otherwise} \end{cases} ,$$

$$(x_\varepsilon)_i = \begin{cases} x_j - \varepsilon, & i = j \\ x_l + \varepsilon, & i = l \\ x_i, & \text{otherwise} \end{cases} .$$

Then $\mathbf{x}_\varepsilon = \mathbf{y}-\mathbf{w}_\varepsilon$. From Propositions 3 and 4, there exist $t_1, t_2 > 0$ such that $J_{\boldsymbol{\lambda}}(\mathbf{w}_\varepsilon) < J_{\boldsymbol{\lambda}}(\mathbf{w}^*)$ for $\varepsilon \in (0, t_1)$ and $J_{\boldsymbol{\lambda}*}(\mathbf{x}_\varepsilon) \leq J_{\boldsymbol{\lambda}*}(\mathbf{x})$ for $\varepsilon \in (0, t_2)$. Define $t := \min\{t_1, t_2\}$ and fix $\varepsilon \in (0, t)$. Then, we have $J_{\boldsymbol{\lambda}*}(\mathbf{y}-\mathbf{w}_\varepsilon) = J_{\boldsymbol{\lambda}*}(\mathbf{x}_\varepsilon) \leq J_{\boldsymbol{\lambda}*}(\mathbf{x}) \leq 1$, hence $\mathbf{w}_\varepsilon$ is feasible with smaller value of objective.

Suppose that $y_j - w^*_j < y_l - w^*_l$ for $j < l$. This gives $x_j < x_l$ and $w^*_j > w^*_l$. The feasible vector, $\mathbf{w}_\varepsilon$, with smaller value of objective, could be now constructed in an analogous manner, again yielding the contradiction with optimality of $\mathbf{w}^*$. □

Proposition 8 states that including inequality constraints $w_1 \geq \ldots \geq w_p \geq 0$ and $y_1 - w_1 \geq \ldots \geq$

Algorithm 2: BogB15

]Solution to SLOPE in the Orthogonal Case (Bogdan et al., 2015; Algorithm 3) **Input**: Nonnegative and nonincreasing sequences y and λ;

while $\mathbf{y} - \boldsymbol{\lambda}$ *is not nonincreasing* **do**

Identify strictly increasing subsequences, i.e. segments $I_i := \{j, \dots, l\}$ such that

$$y_j - \lambda_j < \ldots < y_l - \lambda_l.$$

For each $k \in I_i$ replace the values of y and λ by their average value over such segments

$$y_k \leftarrow \overline{y_{I_i}}, \qquad \lambda_k \leftarrow \overline{\lambda_{I_i}}$$

end

output: $\mathbf{w}_S = (\mathbf{y} - \boldsymbol{\lambda})_+$;

$y_p - w_p \geq 0$ to the problem (17) does not change the set of solutions. These additional restraints simplify objective function and as a result the task takes form of minimizing linear function $\boldsymbol{\lambda}^T\mathbf{w}$. Moreover the condition $\mathbf{y} - \mathbf{w} \in C_{\boldsymbol{\lambda}}$ can now be represented by p affine constraints of the form $\sum_{i=1}^k (y_i - \lambda_i) \leq \sum_{i=1}^k w_i$. That is, the ODS can be casted as a linear program. We will now show that after the transformation, one can omit the conditions $y_1 - w_1 \geq \ldots \geq y_p - w_p \geq 0$, yielding an equivalent formulation

$$\min_{\mathbf{w}} \ \boldsymbol{\lambda}^T\mathbf{w} \qquad \text{s.t.} \begin{cases} \sum_{i=1}^k (y_i - \lambda_i) \leq \sum_{i=1}^k w_i, & k = 1, \dots, p \\ w_1 \geq \ldots \geq w_p \geq 0 \end{cases}. \tag{18}$$

Proposition 9. *Let $\mathbf{w}^*$ be a solution to (18), for $\lambda_1 > \ldots > \lambda_p > 0$ and $y_1 \geq \ldots \geq y_p \geq 0$. Then*

i) $\sum_{i=1}^j (y_i - \lambda_i) = \sum_{i=1}^j w_i^*$ *or* $w_j^* = w_{j+1}^*$,
ii) $y_j - w_j^* \geq y_{j+1} - w_{j+1}^*$,

for all $j \in \{1, \dots, p\}$, with the convention that $w_{p+1}^* := 0$ and $y_{p+1} := 0$.

Proof. Fix $j \in \{1, \dots, p\}$ and suppose that $\mathbf{w} \in \mathbb{R}^p$ is feasible vector of problem (18) such that $\sum_{i=1}^j (y_i - \lambda_i) < \sum_{i=1}^j w_i$ and $w_j > w_{j+1}$, with the convention that $w_{p+1} = 0$. There exists $\varepsilon > 0$, such as

$$\sum_{i=1}^j (y_i - \lambda_i) < \left(\sum_{i=1}^j w_i\right) - \varepsilon \ \text{ and } \ w_j - \varepsilon > w_{j+1} + \varepsilon. \tag{19}$$

Define $\mathbf{w}_\varepsilon \in \mathbb{R}^p$ by putting $(w_\varepsilon)_j := w_j - \varepsilon$, $(w_\varepsilon)_{j+1} := w_{j+1} + \varepsilon$ and $(w_\varepsilon)_i := w_i$ for $i \notin \{j, j+1\}$. Thanks to (19), $\mathbf{w}_\varepsilon$ is feasible (note that $\sum_{i=1}^k w_i = \sum_{i=1}^k (w_\varepsilon)_i$ for $k \neq j$). Now, with convention $\lambda_{p+1} := 0$, it holds $\boldsymbol{\lambda}^T w - \boldsymbol{\lambda}^T w_\varepsilon = \varepsilon(\lambda_j - \lambda_{j+1}) > 0$, which shows that $\mathbf{w}$ is not optimal.

To prove ii), let $\mathbf{w}$ be a feasible point such that $y_j - w_j < y_{j+1} - w_{j+1}$ for some $j \in \{1, \dots, p\}$. Considering the case $j = 1$, from the feasibility of $\mathbf{w}$ we get $y_1 - w_1 < \frac{(y_1 - w_1) + (y_2 - w_2)}{2} \leq \frac{\lambda_1 + \lambda_2}{2} < \lambda_1$. For $j \in \{2, \dots, p\}$ we have $\sum_{i=1}^{j-1} (y_i - w_i) \leq \sum_{i=1}^{j-1} \lambda_i$, $\sum_{i=1}^{j+1} (y_i - w_i) \leq \sum_{i=1}^{j+1} \lambda_i$ (with $\lambda_{p+1} := 0$). Adding both sides of these inequalities and dividing by 2 yields

$$\sum_{i=1}^{j-1} (y_i - w_i) + \frac{(y_j - w_j) + (y_{j+1} - w_{j+1})}{2} \leq \sum_{i=1}^{j-1} \lambda_i + \frac{\lambda_j + \lambda_{j+1}}{2} < \sum_{i=1}^{j} \lambda_i.$$

Due to the assumption $y_j - w_j < y_{j+1} - w_{j+1}$, it follows that

$$\sum_{i=1}^{j} (y_i - w_i) < \sum_{i=1}^{j-1} (y_i - w_i) + \frac{(y_j - w_j) + (y_{j+1} - w_{j+1})}{2} < \sum_{i=1}^{j} \lambda_i.$$

To sum up, we always have $\sum_{i=1}^j (y_i - \lambda_i) < \sum_{i=1}^j w_i$. Moreover, $y_j \geq y_{j+1}$ and $y_j - w_j < y_{j+1} - w_{j+1}$ give that $w_j > w_{j+1}$. Therefore, from (i), the vector $\mathbf{w}$ can not be optimal. □

We now show that the LP (18) has a unique solution. For $k \in \mathbb{N}$, define $k \times k$ upper and lower triangular matrices $\mathbf{S}_k$ and $\mathbf{V}_k$ as

$$\left(S_k\right)_{j,l} = \begin{cases} 1, & j \leq l \\ 0, & \textit{otherwise} \end{cases}, \qquad \left(V_k\right)_{j,l} = \begin{cases} 1, & l = j \\ -1, & j = l+1 \\ 0, & \textit{otherwise} \end{cases}. \tag{20}$$

It could be easily verified, that $\mathbf{S}_k^{-1} = \mathbf{V}_k^T$. We are now ready to prove the following lemma.

Lemma 2. *The LP (18) has a unique solution for $\lambda_1 > \cdots > \lambda_p > 0$.*

Proof. Denote the columns of matrices $\mathbf{S}_p$ and $\mathbf{V}_p$ by $\mathbf{s}_1, \ldots, \mathbf{s}_p$ and $\mathbf{v}_1, \ldots, \mathbf{v}_p$, respectively. From the condition $\mathbf{S}_p \mathbf{V}_p^T = \mathbf{I}_p$, we get that $\mathbf{s}_i$ is orthogonal to $\mathbf{v}_j$ whenever $i \neq j$. Hence, since $\mathbf{S}_p$ and $\mathbf{V}_p$ are nonsingular, the matrix $\left[(\mathbf{S}_p)_{I_1} \middle| (\mathbf{V}_p)_{I_2}\right]$ is nonsingular as well, for any partition $\{I_1, I_2\}$ of the set $\{1, \ldots, p\}$. This means that the set

$$\mathrm{SOL} := \Big\{ \mathbf{w} \in \mathbb{R}^p : \ \left[(\mathbf{S}_p)_{I_1} \middle| (\mathbf{V}_p)_{I_2}\right]^T \mathbf{w} = c^{I_1, I_2}, \\ I_1 \cup I_2 = \{1, \ldots, p\},\ I_1 \cap I_2 = \emptyset \Big\}$$

is finite, where $c_j^{I_1,I_2} := \sum_{i=1}^{j} (y_i - \lambda_i)$, for $j \in I_1$, and $c_j^{I_1,I_2} := 0$, for $j \in I_2$. Let $\mathbf{w}^*$ be any solution to the ODS (18). From Proposition 9 i), for all $j \in \{1, \ldots, p\}$ we have $\mathbf{s}_j^T \mathbf{w}^* = 0$ or $\mathbf{v}_j^T \mathbf{w}^* = 0$, which gives that $\mathbf{w}^* \in \mathrm{SOL}$. Since a feasible LP can have either one of infinitely many solutions, this immediately gives the claim. ☐

Lemma 3. *Consider perturbed version of the LP (18) with the same feasible set but with a new objective function $f_\mu(\mathbf{w}) := \boldsymbol{\lambda}^T \mathbf{w} + \frac{1}{2}\mu \|\mathbf{w}\|_2^2$ with $\mu > 0$. Let $\mathbf{w}^*$ be solution to perturbed problem (which is unique thanks to strong convexity). Then for any $y_1 \geq \ldots \geq y_p \geq 0$ and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ (i.e. coefficients of $\boldsymbol{\lambda}$ do not have to be strictly decreasing and positive), it occurs $w_j^* = w_{j+1}^*$ or $\sum_{i=1}^{j} (y_i - \lambda_i) = \sum_{i=1}^{j} w_i^*$ for all $j \in \{1, \ldots, p\}$ (with $w_{p+1}^* := 0$ and $y_{p+1} := 0$).*

Proof. Take any feasible $\mathbf{w}$ and assume that for some $j \in \{1, \ldots, p\}$ we have $w_j > w_{j+1}$ and $\sum_{i=1}^{j} (y_i - \lambda_i) < \sum_{i=1}^{j} w_i$. Let $\mathbf{w}_\varepsilon$ be feasible vector constructed as in proof of Proposition 9 i). Then

$$\begin{aligned} & f_\mu(\mathbf{w}) - f_\mu(\mathbf{w}_\varepsilon) \\ &= \varepsilon(\lambda_j - \lambda_{j+1}) + \frac{1}{2}\mu\big(w_j^2 + w_{j+1}^2\big) \\ &\quad - \frac{1}{2}\mu\big((w_j - \varepsilon)^2 + (w_{j+1} + \varepsilon)^2\big) \\ &= \varepsilon(\lambda_j - \lambda_{j+1}) + \mu\varepsilon\big((w_j - w_{j+1}) - \varepsilon\big) > 0, \end{aligned}$$

for sufficiently small $\varepsilon > 0$. Hence, $\mathbf{w}$ can not be optimal. ☐

Lemma 4. *Consider perturbed version of the LP (18) as in the previous lemma, with objective function $f_\mu(\mathbf{w})$ and a solution $\mathbf{w}^*$. Moreover, assume that for some $j, l \in \{1, \ldots, p\}$, $j < l$ we have $y_j - \lambda_j \leq y_{j+1} - \lambda_{j+1} \leq \ldots \leq y_l - \lambda_l$. Let I_1 denote the set $\{j, \ldots, l\}$ and $\overline{\mathbf{w}_{I_1}}$ denote the arithmetic mean of subvector $\mathbf{w}_{I_1}$, for any $\mathbf{w} \in \mathbb{R}^p$. Then for any $y_1 \geq \ldots \geq y_p \geq 0$ and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$:*

i) solution is constant on the segment I_1, i.e. $w_j^* = w_{j+1}^* = \ldots = w_l^*$,
ii) $\mathbf{w}^*$ is solution to perturbed problem with $\mathbf{y}$ and $\boldsymbol{\lambda}$ replaced respectively by $\widetilde{\mathbf{y}}$ and $\widetilde{\boldsymbol{\lambda}}$,

where

$$\widetilde{\lambda}_i := \begin{cases} \overline{\lambda_{I_1}}, & i \in I_1 \\ \lambda_i, & \text{otherwise} \end{cases}, \quad \widetilde{y}_i := \begin{cases} \overline{y_{I_1}}, & i \in I_1 \\ y_i, & \text{otherwise} \end{cases}. \tag{21}$$

Proof. To prove i), suppose that $w_k^* > w_{k+1}^*$ for $k \in \{j, \ldots, l-1\}$. Using the convention that $y_0 := w_0^* := \lambda_0 := 0$, from feasibility of $\mathbf{w}^*$ we have

$$\sum_{i=0}^{k-1} (y_i - \lambda_i) \leq \sum_{i=0}^{k-1} w_i^* \quad \text{and} \quad \sum_{i=0}^{k+1} (y_i - \lambda_i) \leq \sum_{i=0}^{k+1} w_i^*.$$

Adding both sides of these inequalities and dividing by 2 yields

$$\begin{aligned} & \sum_{i=0}^{k-1} (y_i - \lambda_i) + \frac{(y_k - \lambda_k) + (y_{k+1} - \lambda_{k+1})}{2} \\ & \leq \sum_{i=0}^{k-1} w_i^* + \frac{w_k^* + w_{k+1}^*}{2} < \sum_{i=1}^{k} w_i^*. \end{aligned}$$

Therefore

$$\begin{aligned} \sum_{i=1}^{k} (y_i - \lambda_i) &\leq \sum_{i=0}^{k-1} (y_i - \lambda_i) + \frac{(y_k - \lambda_k) + (y_{k+1} - \lambda_{k+1})}{2} \\ &< \sum_{i=1}^{k} w_i^*, \end{aligned}$$

which yields contradiction with Lemma 3.

To show that the aforementioned modification of $\boldsymbol{\lambda}$ and $\mathbf{y}$ does not affect the solution, we will first show that feasible sets of both problems are identical. Let D and $\widetilde{D}$ denote the feasible sets for, respectively, initial parameters $(\mathbf{y}, \boldsymbol{\lambda})$ and $(\widetilde{\mathbf{y}}, \widetilde{\boldsymbol{\lambda}})$ given by (21). We start with proving that $\widetilde{D} \subset D$. Let $\mathbf{w}$ be any vector from $\widetilde{D}$. Since $\sum_{i=1}^{k} (y_i - \lambda_i) = \sum_{i=1}^{k} (\widetilde{y}_i - \widetilde{\lambda}_i)$, for $k < j$ and $k \geq l$, the task reduces to showing that $\sum_{i=1}^{k} w_i \geq \sum_{i=1}^{k} (y_i - \lambda_i)$ for any $k \in \{j, \ldots, l-1\}$. Since $\{y_i - \lambda_i\}_{i=j}^{l}$ increases, we simply have $\overline{y_{I_1}} - \overline{\lambda_{I_1}} \geq \frac{1}{k-j+1} \sum_{i=j}^{k} (y_i - \lambda_i)$. Using the convention $y_0 := \lambda_0 := 0$, we get

$$\begin{aligned} & \sum_{i=1}^{k} w_i \geq \sum_{i=1}^{k} (\widetilde{y}_i - \widetilde{\lambda}_i) \\ & = \sum_{i=0}^{j-1} (y_i - \lambda_i) + (k - j + 1) \cdot \left(\overline{y_{I_1}} - \overline{\lambda_{I_1}}\right) \geq \sum_{i=1}^{k} (y_i - \lambda_i). \end{aligned}$$

Now, take any $\mathbf{w} \in D$. After defining $w_0 := 0$ we have $(l - k) \cdot \sum_{i=0}^{j-1} w_i \geq (l - k) \cdot \sum_{i=0}^{j-1} (y_i - \lambda_i)$ and

$(k-j+1)\cdot\sum_{i=0}^{l} w_i \geq (k-j+1)\cdot\sum_{i=0}^{l}(y_i-\lambda_i)$. Adding both sides of these inequalities and dividing by $(l-j+1)$ yields

$$\sum_{i=0}^{j-1} w_i+(k-j+1)\cdot\overline{\mathbf{w}_{I_1}} \geq \sum_{i=0}^{j-1}(y_i-\lambda_i)+(k-j+1)\cdot\big(\overline{\mathbf{y}_{I_1}}-\overline{\boldsymbol{\lambda}_{I_1}}\big).$$

From the monotonicity of $\mathbf{w}$, it occurs $\sum_{i=j}^{k} w_i \geq (k-j+1)\cdot\overline{\mathbf{w}_{I_1}}$. Consequently,

$$\begin{aligned}\sum_{i=1}^{k} w_i &\geq \sum_{i=0}^{j-1} w_i + (k-j+1)\cdot\overline{\mathbf{w}_{I_1}}\\ &\geq \sum_{i=0}^{j-1}(y_i-\lambda_i)+(k-j+1)\cdot\big(\overline{\mathbf{y}_{I_1}}-\overline{\boldsymbol{\lambda}}_{I_1}\big)\\ &= \sum_{i=1}^{k}(\widetilde{y}_i-\widetilde{\lambda}_i)\end{aligned}$$

and $D\subset\widetilde{D}$ as a result.

Suppose now that $\mathbf{w}^*$ is solution for initial parameters $(\mathbf{y},\boldsymbol{\lambda})$, $\widetilde{\mathbf{b}}^*$ is solution for $(\widetilde{\mathbf{y}},\widetilde{\boldsymbol{\lambda}})$ and that

$$\frac{1}{2}\mu\|\widetilde{\mathbf{w}}^*\|_2^2+\widetilde{\boldsymbol{\lambda}}^T\widetilde{\mathbf{w}}^* < \frac{1}{2}\mu\|\mathbf{w}^*\|_2^2+\widetilde{\boldsymbol{\lambda}}^T\mathbf{w}^* \tag{22}$$

From i) we have $w_j^*=\ldots=w_l^*$ and $\widetilde{w}_j^*=\ldots=\widetilde{w}_l^*$, which yields $\widetilde{\boldsymbol{\lambda}}^T\widetilde{\mathbf{w}}^*=\boldsymbol{\lambda}^T\widetilde{\mathbf{w}}^*$ and $\widetilde{\boldsymbol{\lambda}}^T b^*=\boldsymbol{\lambda}^T\mathbf{w}^*$. Therefore from (22) we have $f_\mu(\widetilde{\mathbf{w}}^*)<f_\mu(\mathbf{w}^*)$, which contradicts the optimality of $\mathbf{w}^*$. □

Proof of Theorem 3

Without loss of generality we can assume that we are starting with ordered and nonzero observations. Basing on Propositions 8 and 9, each solution to (8) is also a solution to (18). Since such solution is unique, this immediately gives the uniqueness of (8). Consider perturbed version of (18), with objective f_μ for sufficiently small μ, such as solutions to (18) and its perturbed version coincide (the existence of such μ is guaranteed by (Becker et al., 2011; Theorem 1). Modifying $\mathbf{y}$ and $\boldsymbol{\lambda}$ as in Algorithm 2, after finite number of iterations we finish with converted $\mathbf{y}$ and $\boldsymbol{\lambda}$ such as

$$y_1-\lambda_1\geq\ldots\geq y_p-\lambda_p. \tag{23}$$

From Lemma 4, we know that such modifications do not have an impact on the solution. Therefore, it is enough to show that, when assumption (23) is in use, the solution to SLOPE, i.e. $\mathbf{w}_S=(\mathbf{y}-\boldsymbol{\lambda})_+$, is also the unique solution, to (18).

With $\mathbf{S}_p$ and $\mathbf{V}_p$ defined in (20), the perturbed problem has following convex optimization form with affine inequality constraints

$$\min_{\mathbf{w}}\ \frac{1}{2}\mu\|\mathbf{w}\|_2^2+\boldsymbol{\lambda}^T\mathbf{w} \quad \text{s.t.}\ \begin{cases}\mathbf{S}_p^T(\mathbf{y}-\boldsymbol{\lambda}-\mathbf{w})\preceq 0,\\ -\mathbf{V}_p^T\mathbf{w}\preceq 0\end{cases}. \tag{24}$$

If $\mathbf{y}-\boldsymbol{\lambda}\prec 0$, put $I_1:=\emptyset$, $I_2:=\{1,\ldots,p\}$. Otherwise, let s be the maximal index such that $y_s-\lambda_s\geq 0$ and define $I_1:=\{1,\ldots,s\}$, $I_2:=\{1,\ldots,p\}\backslash I_1$. The KKT conditions for (24) are given by

$$\begin{cases}\mu\mathbf{w}+\boldsymbol{\lambda}=\mathbf{S}_p\boldsymbol{\nu}+\mathbf{V}_p\boldsymbol{\tau}, \quad \text{(Stationary)}\\ \nu_i\big(\mathbf{S}_p^T(\mathbf{y}-\boldsymbol{\lambda}-\mathbf{w})\big)_i=0,\ \tau_i\big(\mathbf{V}_p^T\mathbf{w}\big)_i=0\\ \qquad \text{for each } i\in\{1,\ldots,p\}, \quad \text{(Complementary slackness)}\\ \mathbf{S}_p^T(\mathbf{y}-\boldsymbol{\lambda}-\mathbf{w})\preceq 0,\ -\mathbf{V}_p^T\mathbf{w}\preceq 0, \quad \text{(Primal feasibility)}\\ \boldsymbol{\nu}\succeq 0,\ \boldsymbol{\tau}\succeq 0. \quad \text{(Dual feasibility)}\end{cases}$$

We now show that $(\mathbf{w}^*,\boldsymbol{\nu}^*,\boldsymbol{\tau}^*)$ satisfy the KKT conditions, where $\mathbf{w}^*=(\mathbf{y}-\boldsymbol{\lambda})_+$ and $\boldsymbol{\nu}^*$, $\boldsymbol{\tau}^*$ are given by

$$\boldsymbol{\nu}_{I_1}^*:=\mathbf{V}_s^T(\mu\mathbf{w}^*+\boldsymbol{\lambda})_{I_1},\ \boldsymbol{\nu}_{I_2}^*:=\mathbf{0},\ \boldsymbol{\tau}_{I_1}^*:=\mathbf{0},\\ \boldsymbol{\tau}_{I_2}^*:=\mathbf{S}_{p-s}^T(\mu\mathbf{w}^*+\boldsymbol{\lambda})_{I_2}.$$

It is easy to see that $\mathbf{w}^*$ is primal feasible. Since coefficients of $\mu\mathbf{w}^*+\boldsymbol{\lambda}$ create a nonnegative and nonincreasing sequence, we have $\boldsymbol{\nu}^*\succeq 0$, $\boldsymbol{\tau}^*\succeq 0$. Moreover, $(\mathbf{S}_p)_{I_1}^T(\mathbf{y}-\boldsymbol{\lambda}-\mathbf{w}^*)=0$ and $(\mathbf{V}_p)_{I_2}^T b^*=0$, which shows that complementary slackness conditions are satisfied. Furthermore, we have

$$\begin{aligned}\mathbf{S}_p\boldsymbol{\nu}^*+\mathbf{V}_p\boldsymbol{\tau}^* &= \mathbf{S}_{I_1}\boldsymbol{\nu}_{I_1}^*+\mathbf{V}_{I_2}\boldsymbol{\tau}_{I_2}^*\\ &= \left[\begin{array}{c|c}\mathbf{S}_s & 0\\ \hline 0 & \mathbf{V}_{p-s}\end{array}\right]\begin{bmatrix}\boldsymbol{\nu}_{I_1}^*\\ \boldsymbol{\tau}_{I_2}^*\end{bmatrix}=\mu\mathbf{w}^*+\boldsymbol{\lambda},\end{aligned}$$

which shows that stationary condition is met and finishes the proof. □

SP-HPE Algorithm

For a reference, we provide the SP-HPE algorithm. For the GDS problem, we use $f=0$, $L_f=0$, $g_1=\mathcal{F}$, and $g_2=\mathcal{G}$.

Algorithm 3: SP-HPE Algorithm

input : $(x_0, y_0) \in \mathcal{X}, \mathcal{Y}$, $\eta > 0$, and $\sigma \in (0, 1)$.
for $k = 1, 2, \ldots$ **do**

/* Solve the HPE error condition */

$$(\tilde{u}, \tilde{v}, \tilde{r}^u, \tilde{r}^v, \tilde{\epsilon}) = \text{HPE-Error-Cond}(f = \eta f, A = \eta A, g_1 = \eta g_1, g_2 = \eta g_2, (u_0, v_0) = (x_{k-1}, y_{k-1}), L_f = \eta L_f)$$

/* Update */

$$(\tilde{x}_k, \tilde{y}_k) := (\tilde{u}, \tilde{v}), \quad \tilde{r}_k = (\tilde{r}_k^x, \tilde{r}_k^y) := \frac{1}{\eta}(\tilde{r}^u, \tilde{r}^v), \quad \epsilon_k = \frac{1}{\eta}\tilde{\epsilon}$$
$$x_k = x_{k-1} - \eta \tilde{r}_k^x, \quad y_k = y_{k-1} - \eta \tilde{r}_k^y$$

/* Check Convergence */

end

Algorithm 4: HPE-Error-Cond$(f, A, g_1, g_2, (u_0, v_0), L_f)$ Subroutine

input : f, A, g_1, g_2, (u_0, v_0), and L_f.
Set $L = L_f + \|A\|^2$, $t_0 = 0$, $\tilde{u}_0 = w_0 = P_\Omega(u_0)$, and $\tilde{v}_0 = 0$.
for $k = 1, 2, \ldots$ **do**

/* Update */

$$t_k = t_{k-1} + \frac{1 + \sqrt{1 + 4Lt_{k-1}}}{2L}$$
$$u_k = \frac{t_{k-1}}{t_k}\tilde{u}_{k-1} + \frac{t_k - t_{k-1}}{t_k} w_{k-1}$$
$$v(u_k) = \arg\max_v \; \langle Au, v\rangle - g_2(v) - \frac{1}{2}\|v - v_0\|^2 = \text{prox}_{g_2}(v_0 + Au)$$
$$\tilde{v}_k = \frac{t_{k-1}}{t_k}\tilde{v}_{k-1} + \frac{t_k - t_{k-1}}{t_k} v(u_k)$$
$$w_k = \arg\min_u \; \langle A^* \tilde{v}_k, u\rangle + g_1(u) + \frac{c_k}{2}\|u - u_0\|^2 = \text{prox}_{c_k^{-1} g_1}(u_0 - c_k^{-1} A^* \tilde{v}_k), \quad c_k = 1 + \frac{1}{t_k}$$
$$\tilde{u}_k = \frac{t_{k-1}}{t_k}\tilde{u}_{k-1} + \frac{t_k - t_{k-1}}{t_k} w_k$$

/* Set */

$$\tilde{\epsilon}_k = \frac{1}{2t_k}\|\tilde{u}_k - u_0\|^2, \quad \tilde{r}_k^u = c_k(u_0 - w_k), \quad \tilde{r}_k^v = v_0 - v(\tilde{u}_k) = v_0 - \text{prox}_{g_2}(v_0 + A\tilde{u}_k)$$

/* Check Convergence */ **if** $\|\tilde{r}_k^u + \tilde{u}_k - u_0\|^2 + \|\tilde{r}_k^v + \tilde{v}_k - v_0\|^2 + 2\tilde{\epsilon}_k \le \sigma^2(\|\tilde{u}_k - u_0\|^2 + \|\tilde{v}_k - v_0\|^2)$ **then**
| stop.
end

end