Fast Saddle-Point Algorithm for Generalized Dantzig Selector and FDR Control with the Ordered l1-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 problem, based on a new convex-concave saddle-point (SP) reformulation. Our new formulation makes it possible to adopt 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 local acceleration to achieve the rate of $O(1/k^2)$ 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 original Dantzig selector, to use any norm $\mathcal{R}(\cdot)$ for regularization and its dual norm $\mathcal{R}^D(\cdot)$ for measuring estimation error.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

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 $c > 0$:

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

The original Dantzig selector is attained when $\mathcal{R}(\cdot) = \|\cdot\|_1$ and $\mathcal{R}^D(\cdot) = \|\cdot\|_\infty$. The GDS 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 their 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 to solve (1) based on a linearized version of 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: regarding the latter, projection was onto the dual ball defined with $\mathcal{R}^D(\cdot)$ and therefore can be easily computed via the proximal operator of $\mathcal{R}(\cdot)$ and Moreau's identity (Rockafellar, 1997). The algorithm exhibits $O(1/k)$ convergence rate when its penalty parameter is set to a value at least $\|\mathbf{X}\|_2^4$ (Chatterjee et al., 2014; Wang and Banerjee, 2014). However, its practical performance tends to be quite sensitive to the parameter, whose best value is 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 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 ADMM recently appeared achieving a better $O(1/k^2)$ rate, however, with

an assumption that the objective is strongly convex in case of ADMM (Goldstein et al., 2014; Kadkhodaie et al., 2015), or with a smoothness assumption of the part to be linearized in case of 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: we prove its false discovery rate control properties in variable selection, where the norm itself has been recently studied in other contexts (Bogdan et al., 2013, 2015; Figueiredo and Nowak, 2014).

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 $\|\cdot\|$ and inner products by $\langle\cdot,\cdot\rangle$.

2 CONVEX-CONCAVE SADDLE-POINT FORMULATION

2.1 (More) Generalized Dantzig Selector

In this paper we consider a slightly more general form of the GDS problem (1),

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

where $\mathcal{F}:\mathbb{R}^p \to (-\infty,+\infty]$ is a proper, convex, and lower-semicontinuous (l.s.c.) function, and $\mathcal{G}^D(\cdot)$ 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 $C_{\mathcal{G}^D}$ the constraint set of residuals in (2),

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

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

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

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*

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

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

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

Also, from conjugacy, $\vartheta^{\star}_{C_{\mathcal{G}^D}}(\cdot) = \sup_{\mathbf{w}'\in\mathbb{R}^p}\ \langle\mathbf{w}',\cdot\rangle - \vartheta_{C_{\mathcal{G}^D}}(\mathbf{w}') = \max_{\mathbf{w}':\mathcal{G}^D(\mathbf{w}')\le 1}\langle\mathbf{w}',\cdot\rangle$, which is by definition the dual norm of $\mathcal{G}^D(\cdot)$, 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),

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

This reformulation allows us to benefit from recent developments in saddle-point optimization, 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:

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

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 by Moreau's identity (Rockafellar, 1997), i.e., $\mathbf{z} = \text{prox}_{\mathcal{F}}(\mathbf{z}) + \text{prox}_{\mathcal{F}^\star}(\mathbf{z})$.

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 $\mathcal{V} = \{\mathbf{w}' : \mathcal{G}^D(\mathbf{w}') \le 1\}$ 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 a saddle-point reformulation for the case 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, then it shares similar issues to Nesterov's smoothing that a parameter needs to be specified depending on an 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 for structured output prediction and proposed an algorithm more memory efficient than its quadratic program alternative, based on the dual extragradient technique of Nesterov (2007). The dual extragradient method 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 an ergodic $O(1/k)$ convergence rate, or both $\partial\mathcal{F}$ and $\partial\mathcal{E}$ are bounded in which case the algorithm exhibits 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) belongs to another family of extragradient methods that can be seen as a generalization of Korpelevich's method and some extensions (Monteiro and Svaiter, 2011), and can solve CCSP problems with the same $O(1/k)$ ergodic convergence rate. 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 the aforementioned extragradient methods, although it may look similar to Nesterov's dual extragradient technique (Nesterov, 2007). In Chambolle and Pock (2011), proximal steps are taken in each of the primal and the dual spaces, then a linear gradient extrapolation is considered either in the primal or in the dual. We base our algorithm on this technique, since it has been the fastest with the smallest variations in runtime to solve the problem of our interest in its saddle-point reformulation (4). Both properties were desired in particular for studying statistical properties of the GDS based on random simulations.

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 $\mathcal{G}$, respectively. 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 (5) imply that the following saddle-point inequality holds for any $(\mathbf{w}, \mathbf{v})$,

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

We present our primal-dual saddle-point (PDSP) algorithm in Algorithm 1, which solves the CCSP formulation of the GDS problem (4).

Algorithm 1: Primal-Dual Saddle-Point (PDSP)

Data : $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^n$, $L = \|\mathbf{X}^T \begin{bmatrix}\mathbf{I}_n & -\mathbf{X}\end{bmatrix}\|$;
Initialize: $(\mathbf{w}_0, \mathbf{v}_0) \in \mathbb{R}^p \times \mathbb{R}^p$, $\mathbf{w}'_0 = \mathbf{w}_0$;
Params : $\tau_0 > 0$, $\sigma_0 > 0$ satisfying $\tau_0\sigma_0 L^2 \le 1$,
$\gamma \ge 0$: strong convexity modulus of $\mathcal{G}$;
for $k = 0, 1, 2, \dots$ **do**

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

Check (both if $\gamma = 0$, only pointwise 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 define the primal-dual gap, following Chambolle and Pock (2011), restricted to the set $\mathcal{X} \times \mathcal{Y}$,

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

When $\mathcal{X}\times\mathcal{Y}$ 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

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

Then the result essentially follows from Theorem 1 of Chambolle and Pock (2011). For completeness, we provide the full proof in the supplementary material, 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 crucial for our discussion in the sequel. (We note that similar boundedness results are available for some related methods, e.g. Nemirovski (2004); Tseng (2008), but not for all). In particular, in many sparse regression scenarios in high-dimensions, we expect that $\|\mathbf{w}^*\|$ may not be very large due to the small support size (the number of nonzero components) of a true signal. As our algorithm naturally starts from the zero vector ($\mathbf{w}_0 = \mathbf{0}$), it is therefore likely from Theorem 1 (a), with some proper values of τ_0 and σ_0, that $\|\mathbf{w}_k - \mathbf{w}^*\|$ (or even $\|\mathbf{w}_k\|$) would be small as well, although we need more information about $\|\mathbf{v}_0 - \mathbf{v}^*\|$ to say it definitely.

3.1 Local 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.

Here, we claim that such an acceleration is also possible, at least locally, without strong convexity. Let us focus on $\mathcal{F}$, since our arguments here can be equally 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 modulus $\gamma > 0$ and for any $\mathbf{w}', \mathbf{w} \in \operatorname{dom}\mathcal{F}$.

Suppose that $\mathcal{F}$ is *not* strongly convex (i.e., $\gamma = 0$), as in the general GDS cases (4). Also, suppose that $\mathcal{F}$ is indeed a norm, so that it satisfies the (reverse) 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 $\mathcal{F}(\mathbf{w}^* - \mathbf{w}_k) \leq c\|\mathbf{w}^* - \mathbf{w}_k\|$ holds for some $c > 0$, where the right-hand side is bounded due to Theorem 1 (a), then we can find constants $\bar{c}, \delta > 0$ such that

$$\mathcal{F}(\mathbf{w}^*) - \mathcal{F}(\mathbf{w}_k) \geq \bar{c}\mathcal{F}(\mathbf{w}^* - \mathbf{w}_k) \geq \delta\|\mathbf{w}^* - \mathbf{w}_k\|^2, \quad (6)$$

for all $k \geq k_0$, with 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 $\mathcal{F}$, i.e., $\mathcal{F}(\mathbf{w}^*) \geq \mathcal{F}(\mathbf{w}_k) + \langle \mathbf{g}, \mathbf{w}^* - \mathbf{w}_k \rangle$ with $\mathbf{g} \in \partial\mathcal{F}(\mathbf{w}_k)$, it follows that

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

Comparing to the above inequality of strong convexity, this provides us a weaker notion of strong convexity in the region where (6) holds. We show that this is enough to establish a local accelerated 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 local 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 the supplementary material 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 easily verifiable without knowing $\mathcal{F}(\mathbf{w}^*)$ a priori. Further, (6) implies $\mathcal{F}(\mathbf{w}^*) \geq \mathcal{F}(\mathbf{w}_k)$ for $k \geq k_0$, which is not enforced by our algorithm. Nonetheless, our new result shows that local pointwise convergence with an accelerated rate $O(1/k^2)$ is possible without strong convexity, under some special conditions. In our experience, Algorithm 1 seemed to exhibit 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 Algorithm 1 for non-strongly convex cases.

4 DANTZIG SELECTOR WITH THE ORDERED ℓ_1-NORM

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

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

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 $J^D_{\boldsymbol{\lambda}}$ 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,

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

Although the ODS (8) can be formulated as a linear program, it requires exponentially many constraints to express the constraint set. Our algorithm can avoid handling this thanks to the fact that in our saddle-point reformulation the dual norm appears in forms of the double dual norm, i.e., $J_{\boldsymbol{\lambda}}(\cdot)$:

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

The proximal operator for $J_{\boldsymbol{\lambda}}(\cdot)$ can be computed in $O(p\log p)$ 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 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,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \ \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2+\lambda\|\mathbf{w}\|_1. \tag{9}$$

When observations follow the model $\mathbf{y}=\mathbf{X}\mathbf{w}^*+\boldsymbol{\xi}$ with orthogonal design ($\mathbf{X}^T\mathbf{X}=\mathbf{I}_p$) and noise $\boldsymbol{\xi}\sim\mathcal{N}(0,\sigma^2\mathbf{I}_n)$, one can choose $\lambda\approx\sigma\sqrt{2\log p}$ to control the *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 which replaces the ℓ_1-term in (9) with the ordered ℓ_1-norm $J_{\boldsymbol{\lambda}}(\cdot)$, it has been shown that data-adaptive *false discovery rate (FDR)* control is possible (Bogdan et al., 2015). The SLOPE follows the spirit of the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) in multiple hypothesis testing, which can adapt to 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 the Dantzig Selector. Our next theorem shows that ODS can control FDR, in orthogonal design cases.

Theorem 3. *Under the linear data model* $\mathbf{y}=\mathbf{X}\mathbf{w}^*+\boldsymbol{\xi}$ *with* $\mathbf{X}\in\mathbb{R}^{n\times p}$, $\mathbf{X}^T\mathbf{X}=\mathbf{I}_p$, *and* $\boldsymbol{\xi}\sim\mathcal{N}(0,\sigma^2\mathbf{I}_n)$, *suppose that we choose* $\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_p)$ *according to*

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

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

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

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

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

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

$$\begin{cases} \lambda_1' &= \lambda_1\\ \lambda_i' &= \lambda_i\sqrt{1+\frac{\sum_{j<i}(\lambda_j')^2}{n-i}}, \ \ i\ge 2, \end{cases}$$

and then for $t=\arg\min_i\{\lambda_i'\}$, take

$$\lambda_i^G=\begin{cases}\lambda_i', & i\le t,\\ \lambda_t & i>t.\end{cases} \tag{10}$$

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 a convex function. For details about the adjustment, we refer to (Bogdan et al., 2015; Section 3.2.2).

5 EXPERIMENTS

We demonstrate our algorithm on the ODS instances with 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.

Under the data model $\mathbf{y}=\mathbf{X}\mathbf{w}+\boldsymbol{\xi}$, we sampled each entry of the Gaussian design matrix $\mathbf{X}\in\mathbb{R}^{n\times p}$ and the noise vector $\boldsymbol{\xi}$ independently from the normal 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 (10), with the target FDR level of $q=0.1$.

The performance of Algorithm 1 (PDSP) has been compared to the following alternatives:

SP Algorithms:

- HPE: accelerated hybrid proximal extragradient method (He and Monteiro, 2014).
- MPL: a variant of the mirror-prox (Nemirovski, 2004) with linesearch (Tseng, 2008).

Non-SP Algorithms:

- LADMM: linearized ADMM customized for the GDS (Chatterjee et al., 2014).
- TFOCS: an implementation of Nesterov's smoothing technique (Becker et al., 2011).

Unlike the SP algorithms, the non-SP algorithms require to specify extra parameters difficult to determine: in particular, the penalty parameter $\rho\ge\|\mathbf{X}\|^4$ for LADMM and the smoothing parameter $\mu\approx O(\epsilon)$ for TFOCS. Whenever needed, the values of $\|\mathbf{X}\|$ and $\|\mathbf{A}\|=\|\mathbf{X}^T\begin{bmatrix}\mathbf{I}_n & -\mathbf{X}\end{bmatrix}\|$ were estimated by taking inner products of the matrices with random unit vectors. For TFOCS, we fixed $\mu=10^{-2}\gg\epsilon$, since a larger value than the target optimality ϵ is usually recommended for better performance.

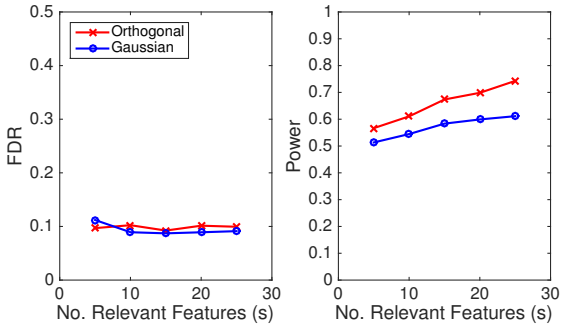

Figure 1: Mean FDR and Power Detecting Signals of Different Sparsity ($s = 5, 10, 15, 20, 25$).

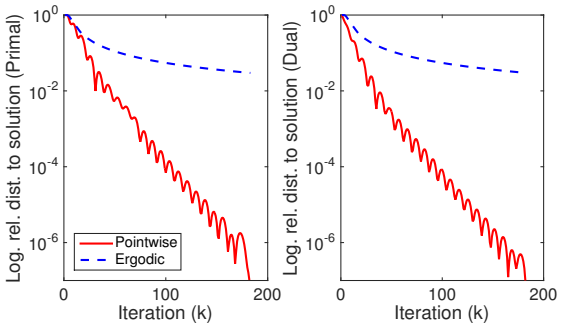

Figure 2: Pointwise and Ergodic Convergence. (Left) Primal; (Right) Dual.

All algorithms were stopped if the following condition was satisfied with an optimality threshold of $\epsilon = 10^{-7}$,

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

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 (PDSP) is its fast speed with small runtime variations while being simple to implement. Table 1 compares the runtime of the algorithms over 50 randomly generated ODS instances in different scenarios, i.e., the combinations of problem dimensions ($p < n$, $p = n$, $p > n$) and signal sparsity ($s = 5, 10, 15$).

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

The advantage of SP methods over non-SP counterparts also 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 have performed 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), we generated random ODS instances with orthogonal and Gaussian design matrices of the dimension $n = 2000$ and $p = 1000$, and compared the FDR and the power of the two cases for the target FDR level of $q = 0.1$.

Figure 1 shows the mean values of these quality criteria over 300 repetitions, for increasing numbers of relevant features (s) in the true signal (also referred to as signal sparsity). FDR was indeed controlled at the desired level of 10% in both orthogonal and Gaussian cases, as we claimed. We observed slightly improved power with orthogonal design compared to the Gaussian cases: it is natural since λ values were adjusted to control FDR resulting in larger penalty for the latter.

Comparing to SLOPE using the same λ values under Gaussian design, ODS appeared to be slightly more conservative, improving FDR and the average number of false discoveries at the cost of a small decrease in power (data not shown). So ODS would be appealing for applications like finding biomarkers from high-dimensional genomic data where false positive discoveries can cost much for follow-up validation. We leave more precise comparison to SLOPE as future work.

5.3 Convergence Rate

Using our algorithm in experiments, we observed fast pointwise convergence in almost every case. This was quite surprising, since pointwise convergence rate is not explained by the existing analysis in Theorem 1, and also expected to be slow, as we discussed earlier.

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

Table 1: Algorithm Runtimes (Suboptimality $\epsilon \leq 10^{-7}$). Mean (Std) in Seconds over 50 Random ODS Instances.

s	p	n	Saddle-Point Algorithm						Non Saddle-Point			
			PDSP		HPE		MPL		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)

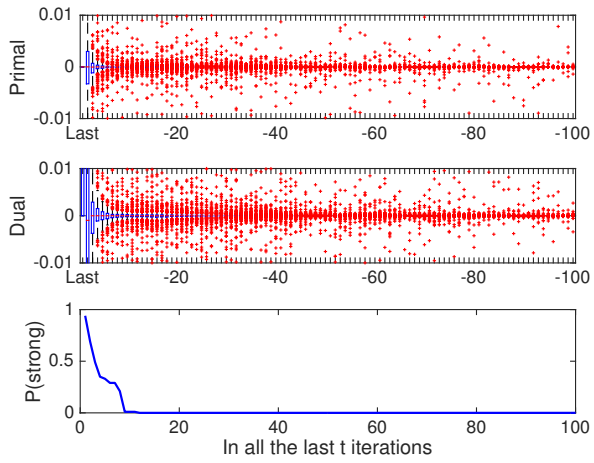

Figure 3: (Top; Middle) Local Strong Convexity Estimates for Primal: $\mathcal{F}$ and Dual: $\mathcal{G}$ in the last 100 iterations. (Bottom) Probability of Local Strong Convexity in Primal or in Dual, for all of the last t iterations.

In contrast, the non-averaged iterates ("pointwise") converged much faster, even exhibiting typical fluctuation patterns of accelerated gradient method. We believe that this behavior is closely related to the local strong convexity and acceleration we discussed.

In fact, in Figure 2, neither any information of local strong convexity nor the alternative choice of $\mathbf{v}'_{k+1} = 2\mathbf{v}_{k+1}$ was used. When the latter option was used, our algorithm showed even faster pointwise convergence, but there were some cases the algorithm did not converge, which would be when the required local strong convexity assumption was not satisfied.

5.4 Local Strong Convexity

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

Figure 3 (top and middle) reports the box-plots of the local strong convexity estimates:

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

in the primal, and equivalent quantities regarding $\mathcal{G}$ in the dual, evaluated for the last 100 iterations of each run. As we approached the last iteration, 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, it is very likely from Theorem 2 that the values need to be positive in either primal or dual: Figure 3 (bottom) shows the chance of such events to happen, in *all* of the last t iterations: the probability seemed to approach one as $t \to 1$. This indicates that local acceleration near an optimal solution would be highly likely.

6 CONCLUSION

We proposed PDSP, a fast and simple primal-dual algorithm to solve the saddle-point formulation of the generalized Dantzig selector. While achieving the known optimal convergence rate, we showed that our algorithm can exhibit a faster rate, taking the advantage of local acceleration. 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.

Acknowledgements

SL was supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876, project C1. D.B. and M.B. were supported by European Union's 7th Framework Programme under Grant Agreement no 602552 and by the Polish Ministry of Science and Higher Education under grant agreement 2932/7.PR/2013/2.

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, W. Su, and E. J. Candes. Statistical estimation and testing via the sorted l1 norm. *arXiv:1310.1969*, 2013.

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*, 9(3):1103–1140, 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.

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.