Primal and Dual Predicted Decrease Approximation Methods

Amir Beck* Edouard Pauwels† Shoham Sabach*

March 22, 2017

Abstract

We introduce the notion of predicted decrease approximation (PDA) for constrained convex optimization, a flexible framework which includes as special cases known algorithms such as generalized conditional gradient, proximal gradient, greedy coordinate descent for separable constraints and working set methods for linear equality constraints with bounds. The new scheme allows the development of a unified convergence analysis for these methods. We further consider a partially strongly convex nonsmooth model and show that dual application of PDA-based methods yields new sublinear convergence rate estimates in terms of both primal and dual objectives. As an example of an application, we provide an explicit working set selection rule for SMO-type methods for training the support vector machine with an improved primal convergence analysis.

Keywords: Primal-dual methods, approximate linear oracles, conditional gradient algorithm, working set methods

1 Introduction

1.1 Context

Linear oracle based methods, such as the conditional gradient algorithm, are arguably among the simplest methods to tackle problems consisting of minimizing smooth functions over compact convex sets. Indeed, such methods amount to solving a sequence of linear programs over the constraint set [14, 11, 23, 13]. This simplicity translates into $O(1/k)$ convergence rates (k being the iteration counter) which are not improvable in general [6, 16]. Despite its apparent simplicity, it was shown in recent works that linear oracle based methods have a very elegant interpretation in the context of convex duality [3] and allow for stronger primal-dual convergence results [16].

In view of this situation, a legitimate question is whether these nice convergence properties can be generalized to more complicated models, as well as more advanced methods such as proximal splitting methods or working set-based methods for constrained optimization. Our starting point will be to take a new look at the conditional gradient algorithm and treat it as an analytical tool that will enable us to analyze convergence properties of more advanced methods.

1.2 Contributions

Our main idea is to ensure that an algorithmic step is "at least as good" as a conditional gradient step. This is the concept of "predicted decrease" which is central in this work and is very much related to the inexact oracle with multiplicative error already presented in [21, 20] in the context of

*Faculty of Industrial Engineering and Management, Technion, Haifa, Israel ({becka,ssabach}@ie.technion.ac.il).

†IRIT-IMT, Universite Paul Sabatier, Toulouse, France (epauwels@irit.fr)

conditional gradient and in [7, 15] in the context of support vector machines. As a first step, we show that this concept of predicted decrease is general enough to encompass a variety of descent methods for problems of the form

$$\min_{\mathbf{y}\in\mathbb{R}^d} \{F(\mathbf{y}) + G(\mathbf{y})\},$$

where F is smooth and convex and G is a closed, convex with compact domain. Our framework allows to unify the convergence analysis of generalized conditional gradient, proximal gradient, greedy coordinate descent for separable constraints and working set methods for linear equality constraints with bounds. For each of these methods, the analysis is based on the same concept of predicted decrease leading to explicit sublinear rates.

As a second step we focus on the partially strongly convex, nonsmooth problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{A}\mathbf{x}) + g(\mathbf{B}\mathbf{x})\},$$

where f is strongly convex, g is convex and globally Lipschitz, $\mathbf{A}$ and $\mathbf{B}$ are matrices where $\mathbf{A}$ has full row rank. We consider the application of a predicted decrease approximation method to the dual of this problem. Our analysis yields an $O(1/k)$ rate of convergence in terms of both primal and dual sequences. These results are original up to our knowledge.

As for practical applications in data analysis, we show that our results translate into improved convergence guaranties in term of primal sequence for the application of SMO-type working set methods [27, 28, 17] to the training of a Support Vector Machine (SVM) [10]. These algorithms take advantage of the structure of the SVM dual quadratic program and allow to perform iterative search with extremely sparse updates–only two coordinates are updated at each iteration. This results in extremely cheap iterations, linear of the size of the dataset, and constitutes one of the most widely used algorithm for SVM training. Convergence rates for these types of algorithms are very scarse in the litterature. We provide an $O(1/k)$ rate of convergence rate for the primal sequence which is the quantity of interest in practice. This improves uppon the $O(1/\sqrt{k})$ rate which is, up to our knowledge, the best known rate in term of primal suboptimality for these types of working set methods [15].

1.3 Relations with previous works

SMO-type working set methods for SVM training. See [33, Section 6.2] for an overview and [8] for implementation details and extensions to broader machine learning settings [12, 35, 30]. Typical convergence results for the dual SVM problem rely on a concept that is directly related to our predicted decrease [15, 31] and yields an $O(1/k)$ convergence rate estimate in terms of the dual SVM objective. Interestingly, primal convergence guaranties for these approaches are very scarce in the literature. In this respect, our result improves upon available results given in [15, 25, 24]. On the practical side, we specify a new working set selection rule for the dual SVM problem which is completely explicit and whose complexity is linear in the number of training examples.

Box plus linear equality constraints. This model generalizes that of the dual SVM to larger number of linear equality constraints [25, 33, 2]. A byproduct of our analysis provides a working set selection based on the fundamental theorem of linear programming. The additional computational cost compared to a single call to the linear oracle of the conditional gradient method is proportional to the dimension. This leads to a search direction which has the same sparsity level as the number of linear equalities while retaining a multiplicative error inversely proportional to the dimension. A similar but less explicit construction was proposed in [25] for the same model.

Approximate linear oracle methods. As we consider linear oracle based methods as our basic analytical tool, some parts of the technical machinery involved in the convergence analysis is inspired by known results for such methods. In particular, our convergence analysis utilizes the artificial introduction of the step size proposed in [21], and non-uniform averaging schemes for primal sequence computation [22, 1, 21]. The analysis that we propose extends to more general models and allows to treat partially strongly convex primal problems.

1.4 Organization of the paper

We introduce the "Predicted Decrease Approximation" (PDA) framework in Section 2 where we show that this scheme encompasses many algorithms as special cases including the proximal gradient method and working set methods for linear equality constrained models. We also give a first sublinear convergence rate estimate. In Section 3, we introduce our partially strongly convex primal model and investigate the application of our PDA framework to the Lagrangian dual. This yields sublinear convergence rates in terms of primal and dual objective sequences. We demonstrate numerically in Section 4 the efficiency of the proposed approach to a synthetic 1D inpainting problem and SVM training.

1.5 Notation

For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $[\mathbf{x}, \mathbf{y}]$ denotes the line segment between them, which is defined by $[\mathbf{x}, \mathbf{y}] = \{\lambda\mathbf{x} + (1-\lambda)\mathbf{y} : 0 \leq \lambda \leq 1\}$. For a given vector $\mathbf{x}$, x_i denotes its ith entry. For a vector $\mathbf{v} \in \mathbb{R}^K$, the norm $\|\mathbf{v}\|$ is the l_2 norm, while for a given matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\|\mathbf{A}\|$ stands for the spectral norm of $\mathbf{A}$. The vector $\mathbf{e}_i$ is ith vector of the canonical basis, meaning that its ith component is one, while all other components are zeros. The $n \times n$ identity matrix is denoted by $\mathbf{I}_n$, where the subscript will be omitted whenever the dimension is clear from the context. For any two vectors $\mathbf{x}, \mathbf{y}$ of the same dimension, $\mathbf{x} \circ \mathbf{y}$ denotes their componentwise product (or Hadmard product), which can also be expressed as $\mathrm{diag}(\mathbf{y})\mathbf{x}$ where for a vector $\mathbf{y}$, $\mathrm{diag}(\mathbf{y})$ denotes the square diagonal matrix whose diagonal elements are the entries of $\mathbf{y}$. We denote by $\mathbf{y}^\dagger$ as the vector for which $y_i^\dagger = 1/y_i$ whenever $y_i \neq 0$ and $y_i^\dagger = 0$ otherwise. For a matrix $\mathbf{A}$, $\mathrm{im}(\mathbf{A})$ denotes its image space (the subspace spanned by its columns). The l_0 norm of a vector (which is actually not a norm) is the number of nonzero elements in $\mathbf{x}$, that is, $\|\mathbf{x}\|_0 = \#\{i : x_i \neq 0\}$. A function $h : \mathbb{R}^K \to \mathbb{R}$ is called *M-smooth* if it is continuously differentiable and its gradient is Lipschitz continuous with constant M:

$$\|\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^K.$$

An extended real-valued function $h : \mathbb{R}^K \to (-\infty, \infty]$ is called μ-strongly convex ($\mu > 0$ being a parameter) if $h(\cdot) - \frac{\mu}{2}\|\cdot\|^2$ is convex. Given a function $f : \mathbb{R}^n \to (-\infty, \infty]$, the convex conjugate is the function

$$f^*(\mathbf{y}) = \max_{\mathbf{x}\in\mathbb{R}^n} \{\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})\}.$$

We also use standard notations from convex analysis as in [29].

2 The Predicted Decrease Approximation (PDA) method

Consider the composite model

$$\min_{\mathbf{y}\in\mathbb{R}^d} \{H(\mathbf{y}) \equiv F(\mathbf{y}) + G(\mathbf{y})\}, \tag{2.1}$$

where the following assumption is made throughout this section

Assumption 1.

- $F : \mathbb{R}^d \to \mathbb{R}$ *is L-smooth and convex over* $\mathbb{R}^d$.
- $G : \mathbb{R}^d \to (-\infty, \infty]$ *is a proper closed convex function with a compact domain.*

It follows from Assumption 1 that problem (2.1) consists of minimizing a proper convex closed function over a compact domain and thus has a finite optimal value, denoted by H^*, which is attained. When G is an indicator of a compact convex set, problem (2.1) amounts to minimize a smooth function over a compact set. The optimization model (2.1) allows for more general nonsmooth convex functions G. We will denote the diameter of $\operatorname{dom} G$ by

$$\operatorname{diam}(\operatorname{dom} G) = \max_{\mathbf{x},\mathbf{y} \in \operatorname{dom} G} \|\mathbf{x} - \mathbf{y}\|. \tag{2.2}$$

An important mathematical quantity that will be used throughout the analysis of the method is the optimality measure defined by

$$S(\mathbf{y}) \equiv \max_{\mathbf{p}} \{\langle \nabla F(\mathbf{y}), \mathbf{y} - \mathbf{p} \rangle + G(\mathbf{y}) - G(\mathbf{p})\}.$$

Using the definition of the convex conjugate of a function, the optimality measure can be rewritten as

$$S(\mathbf{y}) = G(\mathbf{y}) + G^*(-\nabla F(\mathbf{y})) + \langle \nabla F(\mathbf{y}), \mathbf{y} \rangle. \tag{2.3}$$

By Fenchel's inequality $S(\mathbf{y}) \geq 0$ for any $\mathbf{y} \in \operatorname{dom} G$, and by the conjugate subgradient theorem [29, Theorem 23.5], we have that $S(\mathbf{y}) = 0$ holds if and only if $-\nabla F(\mathbf{y}) \in \partial G(\mathbf{y})$, that is, if and only if $\mathbf{y}$ is an optimal solution of (2.1). It is also known (see e.g., [4]) that

$$H(\mathbf{y}) - H^* \leq S(\mathbf{y}) \tag{2.4}$$

for any $\mathbf{y} \in \operatorname{dom} G$, hence the name "optimality measure". We will also use the notation

$$\mathbf{p}(\mathbf{y}) \in \operatorname{argmin}_{\mathbf{p}} \{\langle \nabla F(\mathbf{y}), \mathbf{p} \rangle + G(\mathbf{p})\}, \tag{2.5}$$

where we will assume throughout the paper that when the optimum is attained at multiple points, there is an arbitrary but fixed choice $\mathbf{p}(\mathbf{y})$ for each $\mathbf{y}$. With this notation, we can write

$$S(\mathbf{y}) = \langle \nabla F(\mathbf{y}), \mathbf{y} - \mathbf{p}(\mathbf{y}) \rangle + G(\mathbf{y}) - G(\mathbf{p}(\mathbf{y})).$$

We can interpret $S(\mathbf{y}) = \langle \nabla F(\mathbf{y}), \mathbf{y} \rangle + G(\mathbf{y}) - [\langle \nabla F(\mathbf{y}), \mathbf{p}(\mathbf{y}) \rangle + G(\mathbf{p}(\mathbf{y}))]$ as the *predicted decrease* at $\mathbf{y}$ by the approximate function $\mathbf{z} \mapsto \langle \nabla F(\mathbf{y}), \mathbf{z} \rangle + G(\mathbf{z})$. The vector $\mathbf{p}(\mathbf{y})$ is important and is being used for example in the so-called generalized conditional gradient method in which at each iteration k, the vector $\mathbf{p}(\mathbf{y}^k)$ is computed and the next iteration is defined by the update rule $\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)$ for an appropriate stepsize t_k (see [1, 4]). In this section we will define a much broader class of methods that is related to a different vector whose predicted decrease is at least a certain portion of the predicted decrease of $\mathbf{p}(\mathbf{y})$.

Definition 2.1 ($\frac{1}{\gamma}$-predicted decrease approximation)**.** *For* $\gamma \geq 1$ *and* $\bar{\mathbf{y}} \in \operatorname{dom} G$*, we say that a vector* $\mathbf{u}(\bar{\mathbf{y}}) \in \operatorname{dom} G$ *is a* **$\frac{1}{\gamma}$-predicted decrease approximation (PDA) vector** *of* H *at* $\bar{\mathbf{y}}$ *if*

$$\frac{1}{\gamma} S(\bar{\mathbf{y}}) \leq \langle \nabla F(\bar{\mathbf{y}}), \bar{\mathbf{y}} - \mathbf{u}(\bar{\mathbf{y}}) \rangle + G(\bar{\mathbf{y}}) - G(\mathbf{u}(\bar{\mathbf{y}})). \tag{2.6}$$

Note that for any $\gamma' \geq \gamma \geq 1$*, any* $\frac{1}{\gamma}$*-PDA vector is also a* $\frac{1}{\gamma'}$*-PDA vector.*

We will sometimes refer to a $\frac{1}{\gamma}$-PDA vector of H as a $\frac{1}{\gamma}$-PDA vector *of problem (2.1)*. The constant $\frac{1}{\gamma}$ will be called *the approximation factor*. For the classical conditional gradient method (in the special case where G is the indicator function of a compact convex set), this definition appeared under the name "approximate linear oracle" with multiplicative error [20, 21] or "rate certifying methods" in the context of SVM training [7, 15]. Although the definition is a simple generalization of the concept of approximate linear oracles to composite models, the point of view is different – the approximation does not necessarily comes from approximation errors, but from the fact that it allows to ensure additional structure in the form of the update while maintaining desirable convergence properties. For example it might be reasonable to construct $\mathbf{u}(\cdot)$ requiring more computations than $\mathbf{p}(\cdot)$ if it ensures additional structural features. One trivial example of a PDA vector is the choice $\mathbf{u}(\bar{\mathbf{y}}) = \mathbf{p}(\bar{\mathbf{y}})$, which is obviously a 1-PDA vector. However, in cases where additional properties are required from the vector $\mathbf{u}(\bar{\mathbf{y}})$, other choices should be considered. For example, in some applications (such as support vector machines [28, 33]), it is important to choose a vector $\mathbf{u}(\bar{\mathbf{y}})$ which is different from $\bar{\mathbf{y}}$ by only a few coordinates, namely that $\mathbf{u}(\bar{\mathbf{y}}) - \bar{\mathbf{y}}$ is sparse. The next example shows that when G is block-separable, we can always construct a $\frac{1}{m}$-PDA vector (m being the number of blocks) at any given vector $\bar{\mathbf{y}}$, which is different from $\bar{\mathbf{y}}$ by only one coordinate.

Example 2.2 (separable nonsmooth parts)**.** *Consider a partition of the decision variables vector* $\mathbf{y}$ *to* m *blocks:*

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix},$$

where $\mathbf{y}_i \in \mathbb{R}^{d_i}$ *and* $d_1 + d_2 + \cdots + d_m = d$*. We define the matrices* $\mathbf{U}_i \in \mathbb{R}^{d \times d_i}, i = 1, 2, \ldots, m$ *for which*

$$(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_m) = \mathbf{I}_d.$$

Suppose that

$$G(\mathbf{y}) = \sum_{i=1}^{m} G_i(\mathbf{y}_i), \tag{2.7}$$

where by the properties of G *given in Assumption 1, it follows that* $G_1, G_2, \ldots, G_m$ *are closed proper and convex with compact domain. Let us define the following* i*th partial optimality measure for any* $i = 1, 2, \ldots, m$*:*

$$S_i(\mathbf{y}) = \max_{\mathbf{p}_i} \{\langle \nabla_i F(\mathbf{y}), \mathbf{y}_i - \mathbf{p}_i \rangle + G_i(\mathbf{y}_i) - G_i(\mathbf{p}_i)\}$$

with $\nabla_i F$ *being the vector of partial derivatives of* F *corresponding to the* i*th block. Obviously,* $S(\mathbf{y}) = \sum_{i=1}^{m} S_i(\mathbf{y})$ *for any* $\mathbf{y} \in \operatorname{dom} G$*. Now, suppose that* $\bar{\mathbf{y}} \in \operatorname{dom} G$*, and let*

$$\bar{i} \in \operatorname{argmax}_{i=1,2,\ldots,m} S_i(\bar{\mathbf{y}}). \tag{2.8}$$

With this definition we have

$$S(\bar{\mathbf{y}}) = \sum_{i=1}^{m} S_i(\bar{\mathbf{y}}) \leq m S_{\bar{i}}(\bar{\mathbf{y}}). \tag{2.9}$$

Let $\mathbf{z}_{\bar{i}} \in \operatorname{dom}(G_{\bar{i}})$ *be given by*

$$\mathbf{z}_{\bar{i}} \in \operatorname{argmin}_{\mathbf{p}_{\bar{i}}} \{\langle \nabla_{\bar{i}} F(\mathbf{y}), \mathbf{p}_{\bar{i}} \rangle + G_i(\mathbf{p}_{\bar{i}})\},$$

so that in particular

$$S_{\bar{i}}(\bar{\mathbf{y}}) = \langle \nabla_{\bar{i}} F(\bar{\mathbf{y}}), \bar{\mathbf{y}}_{\bar{i}} - \mathbf{z}_{\bar{i}} \rangle + G_{\bar{i}}(\bar{\mathbf{y}}_{\bar{i}}) - G_{\bar{i}}(\mathbf{z}_{\bar{i}}). \tag{2.10}$$

Define $\mathbf{u}(\bar{\mathbf{y}}) = \bar{\mathbf{y}} + \mathbf{U}_{\bar{i}}(\mathbf{z}_{\bar{i}} - \bar{\mathbf{y}}_{\bar{i}})$. *Then* $\mathbf{u}(\bar{\mathbf{y}}) \in \operatorname{dom} G$ *and*

$$\langle \nabla F(\bar{\mathbf{y}}), \bar{\mathbf{y}} - \mathbf{u}(\bar{\mathbf{y}}) \rangle + G(\bar{\mathbf{y}}) - G(\mathbf{u}(\bar{\mathbf{y}})) = \langle \nabla_{\bar{i}} F(\bar{\mathbf{y}}), \bar{\mathbf{y}}_{\bar{i}} - \mathbf{z}_{\bar{i}} \rangle + G_{\bar{i}}(\bar{\mathbf{y}}_{\bar{i}}) - G_{\bar{i}}(\mathbf{z}_{\bar{i}}) = S_{\bar{i}}(\bar{\mathbf{y}}) \geq \frac{1}{m} S(\bar{\mathbf{y}}),$$

where the first equality uses (2.7) and the inequality follows from (2.9), establishing the fact that $\mathbf{u}(\bar{\mathbf{y}})$ *is indeed a* $\frac{1}{m}$*-PDA vector.*

2.1 The method

We will require the following standard notation (see e.g., [5]):

$$Q_L(\mathbf{y}, \mathbf{x}) \equiv F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

Below we describe the general form of the $\frac{1}{\gamma}$-predicted decrease approximation ($\frac{1}{\gamma}$-PDA) method (for some given $\gamma \geq 1$). At each iteration k, the method constructs $\mathbf{y}^{k+1}$ out of the current iterate $\mathbf{y}^k$ by minimization of the exact original objective or the quadratic model Q_L over a set that contains the line segment $[\mathbf{y}^k, \mathbf{u}(\mathbf{y}^k)]$, where $\mathbf{u}(\mathbf{y}^k)$ is a $\frac{1}{\gamma}$-predicted decrease approximation vector of H at $\mathbf{y}^k$.

$\frac{1}{\gamma}$-predicted decrease approximation ($\frac{1}{\gamma}$-PDA) Method:
Initialization. $\mathbf{y}^0 \in \operatorname{dom} G$.
General Step. For $k = 0, 1, \ldots,$

(i) – Choose $\mathbf{u}(\mathbf{y}^k)$- a $\frac{1}{\gamma}$-PDA vector of H at $\mathbf{y}^k$.
– Choose a compact set X^k for which $[\mathbf{y}^k, \mathbf{u}(\mathbf{y}^k)] \subseteq X^k$.

(ii) Perform one of the following:

$$\text{Local model update: } \mathbf{y}^{k+1} = \operatorname{argmin}_{\mathbf{y} \in X^k} Q_{L_k}(\mathbf{y}, \mathbf{y}^k) + G(\mathbf{y}) \tag{2.11}$$

$$\text{Exact update: } \mathbf{y}^{k+1} = \operatorname{argmin}_{\mathbf{y} \in X^k} F(\mathbf{y}) + G(\mathbf{y}) \tag{2.12}$$

Remark 2.3. *The description that has been made so far is formal and highlights the important mechanisms of the PDA framework. Therefore, the steps that have been described in the algorithm may not reflect exactly the computational effort for each specific instance of the method. Two comments are in order.*

- *Only* X^k*, and not* $\mathbf{u}(\mathbf{y}^k)$*, is required for computation in step (ii) and the only important property of this set is the second condition in (i). In some settings (e.g. greedy coordinate descent), the computation of* $\mathbf{u}(\mathbf{y}^k)$ *is also required but this is not necessarily the case and* $\mathbf{u}(\mathbf{y}^k)$ *could be implicit and not computed in practice (e.g. proximal gradient algorithm). See also Section 2.2 for more details.*

- *In general, there is a tradeoff between steps (i) and (ii). Step (i) can be seen as a reduction step which goal is to decrease the complexity of computing step (ii) or increase its efficiency in term of reaching the global minimum. In many PDA-methods, the separation between steps (i) and (ii) is not that clear and both steps can be mixed. The current presentation highlights the different roles of each step, but does not necessarily reflect the practical implementation of the algorithm.*

We will sometimes refer to a $\frac{1}{\gamma}$-PDA method as a PDA method with approximation factor $\frac{1}{\gamma}$. Note that from Definition 2.1 if $\gamma' \geq \gamma \geq 1$, then a $\frac{1}{\gamma}$-PDA method is also a $\frac{1}{\gamma'}$-PDA method. The

local model update step can be equivalently written as a proximal gradient step:

$$\mathbf{y}^{k+1} = \text{prox}_{\frac{1}{L_k}G+\delta_{X^k}}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla F(\mathbf{y}^k)\right), \tag{2.13}$$

where for a given proper closed convex extended real-valued convex function $h : \mathbb{R}^n \to (-\infty, \infty]$, the proximal operator is defined by [26]

$$\text{prox}_h(\mathbf{x}) = \text{argmin}_{\mathbf{u}}\left\{h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\right\}.$$

One element that is missing in the above description of the $\frac{1}{\gamma}$-PDA method when local model updates are chosen is the way the constants L_k are chosen. When the step involves an exact update, only for the purpose of analytical proofs, we will artificially define $L_k = L$. When a local model update is employed, then the underlying assumption on L_k is that

$$F(\mathbf{y}^{k+1}) \le Q_{L_k}(\mathbf{y}^{k+1}, \mathbf{y}^k). \tag{2.14}$$

Two choices of L_k that warrant inequality (2.14) are

1. $L_k \equiv L$, where L is the global Lipschitz constant of ∇F (which exists thanks to Assumption 1).
2. L_k is chosen by a backtracking procedure. Specifically (see [5, FISTA with backtracking]), we take $\eta > 1$ and $\bar{L} > 0$ (initial estimate of L_k) and at each iteration we pick the smallest nonnegative integer i_k for which (2.14) is satisfied with $L_k = \eta^{i_k}\bar{L}$ and $\mathbf{y}^{k+1}$ given by (2.13).

Since (2.14) is satisfied with $L_k \ge L$, we obtain that if the kth step uses the local model update with backtracking, then

$$L_k \le \max\{\eta L, \bar{L}\}. \tag{2.15}$$

2.2 Special cases of the PDA method

The PDA method is actually a very general scheme and different choices of PDA vectors $\mathbf{u}(\cdot)$ and sets X^k can result in quite different methods – some are well known.

2.2.1 Generalized conditional gradient method

Taking $\mathbf{u}(\mathbf{y}) \equiv \mathbf{p}(\mathbf{y}), X^k = [\mathbf{y}^k, \mathbf{u}(\mathbf{y}^k)]$ and using the exact update scheme, we obtain that the PDA method reduces to the generalized conditional gradient method with exact line search [1].

generalized conditional gradient
Initialization: $\mathbf{y}^0 \in \text{dom}\, G$.
General step (k=0,1,...):

- Compute $\mathbf{p}(\mathbf{y}^k) \in \text{argmin}_{\mathbf{p}}\left\{\langle \nabla f(\mathbf{y}^k), \mathbf{p}\rangle + G(\mathbf{p})\right\}$.
- Set $\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)$ where

$$t_k \in \text{argmin}_{t\in[0,1]} H(\mathbf{y}^k + t(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)).$$

The above method is a PDA method with approximation factor 1. Note that if we change the choice of X^k to $X^k = \{\mathbf{y}^k + t(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k) : t \ge 0\}$, then the PDA method amounts to a variation of the generalized conditional gradient method in which larger stepsizes can be taken. Specifically, the stepsize in this setting is given by

$$t_k \in \text{argmin}_{t\ge 0}\{H(\mathbf{y}^k + t(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)) : \mathbf{y}^k + t(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k) \in \text{dom}\, G\}.$$

Obviously, since X^k still contains $\mathbf{p}(\mathbf{y}^k)$, it follows that the approximation factor is still 1.

2.2.2 Proximal gradient method

Taking $\mathbf{u}(\mathbf{y}) \equiv \mathbf{p}(\mathbf{y}), X^k = \mathbb{R}^d$ and using the local model update, we obtain the proximal gradient method [5, 9].

proximal gradient
Initialization: $\mathbf{y}^0 \in \operatorname{dom} G$.
General step (k=0,1,...):

- compute

$$\mathbf{y}^{k+1} = \operatorname{prox}_{\frac{1}{L_k}G}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla F(\mathbf{y}^k)\right),$$

  where L_k satisfies the condition (2.14).

The above description of the proximal gradient method encompasses both a constant stepsize scheme where $L_k = L$ for any k, as well as a backtracking scheme that guarantees the validity of the inequality (2.14). The approximation factor of the method is 1.

2.2.3 Hybrid proximal gradient/generalized conditional gradient

Since the strategy for choosing $\mathbf{u}(\mathbf{y}^k)$ and X^k at each iteration can be different, one can construct a 1-PDA method that chooses at each iteration to either employ a generalized conditional gradient step or a proximal gradient step.

2.2.4 Greedy coordinate descent for separable problems

Back to the setting of Example 2.2, assume that G is separable and has the form (2.7). Let us consider the following two choices for the set X^k:

$$\begin{aligned}\bar{X}^k &= \{\bar{\mathbf{y}}_1\} \times \{\bar{\mathbf{y}}_2\} \times \cdots \{\bar{\mathbf{y}}_{\bar{i}-1}\} \times [\mathbf{y}^k_{\bar{i}}, \mathbf{p}_{\bar{i}}(\mathbf{y}^k)] \times \{\bar{\mathbf{y}}_{\bar{i}+1}\} \times \cdots \times \{\bar{\mathbf{y}}_m\},\\ \tilde{X}^k &= \{\bar{\mathbf{y}}_1\} \times \{\bar{\mathbf{y}}_2\} \times \cdots \{\bar{\mathbf{y}}_{\bar{i}-1}\} \times \operatorname{dom} G_{\bar{i}} \times \{\bar{\mathbf{y}}_{\bar{i}+1}\} \times \cdots \times \{\bar{\mathbf{y}}_m\}.\end{aligned}$$

The general form of the resulting method, which uses a greedy-type index selection strategy is now described.

greedy coordinate descent
Initialization: $\mathbf{y}^0 \in \operatorname{dom} G$.
General step $(k = 0, 1, \ldots)$**:**

- Compute

$$\bar{i} \in \operatorname{argmax}_{i=1,2,\ldots,m} S_i(\mathbf{y}^k),$$

  where

$$S_i(\mathbf{y}^k) = \langle \nabla F_i(\mathbf{y}^k), \mathbf{y}^k_i - \mathbf{p}_i(\mathbf{y}^k)\rangle + G_i(\mathbf{y}^k_i) - G_i(\mathbf{p}_i(\mathbf{y}^k))$$

  with

$$\mathbf{p}_i(\mathbf{y}^k) \in \operatorname{argmin}_{\mathbf{p}_i}\{\langle \nabla F_i(\mathbf{y}^k), \mathbf{p}_i\rangle + G_i(\mathbf{p}_i)\}.$$

- **Core step:** Compute $\mathbf{y}^{k+1}$.

The update formula of $\mathbf{y}^{k+1}$ ("core step") depends on the specific choice of X^k and the type of update rule (exact/local model). Some options are given below.

- *greedy block conditional gradient* ($X^k = \bar{X}^k$, exact update)

$$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k \mathbf{U}_{\bar{i}}(\mathbf{p}_{\bar{i}}(\mathbf{y}^k) - \mathbf{y}^k_{\bar{i}}),$$

  where $t_k \in \operatorname{argmin}_{0 \le t \le 1} H\left(\mathbf{y}^k + t\mathbf{U}_{\bar{i}}(\mathbf{p}_{\bar{i}}(\mathbf{y}^k) - \mathbf{y}^k_{\bar{i}})\right)$.

- *greedy block minimization* ($X^k = \tilde{X}^k$, exact update)

$$\mathbf{y}^{k+1}_i \begin{cases} = \mathbf{y}^k_i, & i \neq \bar{i}, \\ \in \operatorname{argmin}_{\mathbf{y}_{\bar{i}}} \left\{F(\mathbf{y}^k + \mathbf{U}_{\bar{i}}(\mathbf{y}_{\bar{i}} - \mathbf{y}^k_{\bar{i}})) + G_{\bar{i}}(\mathbf{y}_{\bar{i}}) : \mathbf{y}_{\bar{i}} \in \operatorname{dom} G_{\bar{i}}\right\}, & i = \bar{i}. \end{cases} \tag{2.16}$$

- *greedy block projected-gradient* ($X^k = \tilde{X}^k$, local model step)

$$\mathbf{y}^{k+1}_i = \begin{cases} \mathbf{y}^k_i, & i \neq \bar{i}, \\ \operatorname{prox}_{\frac{1}{L_k} G_{\bar{i}}}\left(\mathbf{y}^k_{\bar{i}} - \frac{1}{L_k}\nabla_{\bar{i}} F(\mathbf{y}^k)\right), & i = \bar{i}. \end{cases}$$

As shown in Example 2.2, all these methods are $\frac{1}{m}$-PDA methods.

2.2.5 Block descent method for linearly constrained problems

In this section we consider instances of the general model (2.1) in which the constraints are the intersection of linear equalities and bound constraints, see also [25, 33, 2]. These models admit extensions of the working set methods originally developed for the dual of the SVM training problem ("SMO-type methods") [27, 28, 17], see also [8] and references therein. We present a new working set selection rule based on the $\frac{1}{\gamma}$-PDA framework and then comment on the application to the dual SVM problem.

Model and construction of PDA: Suppose that $G(\mathbf{y}) \equiv \delta_C(\mathbf{y})$, where

$$C = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{D}\mathbf{y} = \mathbf{b}, \boldsymbol{\ell} \le \mathbf{y} \le \mathbf{u}\},$$

with $\mathbf{D} \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^m$ and $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^d$ are two vectors satisfying $\boldsymbol{\ell} \le \mathbf{u}$ (inequalities between vectors are understood coordinatewise). In this case, problem (2.1) takes the form

$$\begin{array}{ll} \min & F(\mathbf{y}) \\ \text{s.t.} & \mathbf{D}\mathbf{y} = \mathbf{b}, \\ & \boldsymbol{\ell} \le \mathbf{y} \le \mathbf{u}. \end{array} \tag{2.17}$$

The vector $\mathbf{p}(\mathbf{y})$ is obviously a 1-PDA vector of the problem at $\mathbf{y}$. The question is whether we can find a PDA vector with an appropriate approximation factor, which is different from $\mathbf{y}$ by only a few components, thus enabling sparse updates. For that, we introduce below a procedure, termed sparseDir, which, for a given point $\bar{\mathbf{y}} \in C$, finds a direction vector $\mathbf{d}_s(\bar{\mathbf{y}}) \in \mathbb{R}^d$ that will be shown in Lemma 2.5 to satisfy that (i) it has at most $m+1$ nonzero elements and (ii) $\bar{\mathbf{y}} + \mathbf{d}_s(\bar{\mathbf{y}})$ is a $\frac{1}{d}$-PDA vector of problem (2.17) at $\bar{\mathbf{y}}$.

sparseDir has three main steps. The first is based on a regular general conditional gradient step which allows to find a dense 1-PDA vector. The second one is a reduction step. It consists in finding basic feasible solution of a certain linear program while not decreasing a given objective function (this is standard in linear programming). The linear program is designed such that the resulting basic feasible solutions are sparse. The last step is the construction of a sparse PDA vector from the solution of the linear program.

spraseDir
Input: $\bar{\mathbf{y}} \in C$.
Output: $\mathbf{d}_s(\bar{\mathbf{y}}) \in \mathbb{R}^d$.
Algorithm:

(i) Set

$$\begin{aligned} \mathbf{r} &= \mathbf{p}(\bar{\mathbf{y}}) - \bar{\mathbf{y}}, \\ \tilde{\mathbf{D}} &= \mathbf{D}\mathrm{diag}(\mathbf{r}), \\ \mathbf{c} &= \mathbf{r} \circ \nabla F(\bar{\mathbf{y}}). \end{aligned} \tag{2.18}$$

(ii) Compute $\bar{\mathbf{v}}$, a basic feasible solution of the linear system

$$\begin{aligned} \tilde{\mathbf{D}}\mathbf{v} &= \mathbf{0}, \\ \langle \mathbf{1}, \mathbf{v} \rangle &\leq \|\mathbf{r}\|_0, \\ \mathbf{v} &\geq \mathbf{0}. \end{aligned} \tag{2.19}$$

such that

$$\langle \mathbf{c}, \bar{\mathbf{v}} \rangle \leq \langle \mathbf{c}, \mathbf{r}^\dagger \circ \mathbf{r} \rangle. \tag{2.20}$$

(iii) If $\|\mathbf{r}\|_0 = 0$, set $\mathbf{d}_s(\bar{\mathbf{y}}) := \mathbf{0}$. Otherwise set

$$\mathbf{d}_s(\bar{\mathbf{y}}) := \frac{1}{\|\mathbf{r}\|_0} \mathbf{r} \circ \bar{\mathbf{v}}. \tag{2.21}$$

Remark 2.4 (Validity of the procedure)**.** *The set of solutions of (2.19) is nonempty and bounded. The boundedness follows from the constraints $\langle \mathbf{1}, \mathbf{v} \rangle \leq \|\mathbf{r}\|_0, \mathbf{v} \geq \mathbf{0}$. The feasibility of (2.19) follows by the fact that the vector $\mathbf{v} = \mathbf{r}^\dagger \circ \mathbf{r}$ is feasible. Indeed, since $\mathbf{D}\mathbf{p}(\bar{\mathbf{y}}) = \mathbf{D}\bar{\mathbf{y}} = \mathbf{b}$, we have*

$$\tilde{\mathbf{D}}(\mathbf{r}^\dagger \circ \mathbf{r}) = \mathbf{D}\mathrm{diag}(\mathbf{r})(\mathbf{r}^\dagger \circ \mathbf{r}) = \mathbf{D}\mathbf{r} = \mathbf{D}(\mathbf{p}(\bar{\mathbf{y}}) - \bar{\mathbf{y}}) = \mathbf{0}.$$

Furthermore, it can be easily checked that $\langle \mathbf{1}, \mathbf{r}^\dagger \circ \mathbf{r} \rangle = \|\mathbf{r}\|_0$ and $\mathbf{r}^\dagger \circ \mathbf{r} \geq \mathbf{0}$, establishing the feasibility of $\mathbf{r}^\dagger \circ \mathbf{r}$. The fundamental theorem of linear programming [18] (with objective $\langle \mathbf{c}, \cdot \rangle$ and constraints (2.19)) ensures that there exists a basic feasible solution of system (2.19) for which (2.20) is satisfied.

Given a vector $\mathbf{y}$, define

$$\mathbf{u}_s(\mathbf{y}) = \mathbf{y} + \mathbf{d}_s(\mathbf{y}). \tag{2.22}$$

We will now show that $\mathbf{u}_s(\mathbf{y})$ is a $\frac{1}{d}$-PDA vector of problem (2.17) at $\mathbf{y}$, which is different from $\mathbf{y}$ by at most $m + 1$ elements.

Lemma 2.5. *Fix $\mathbf{y} \in C$. Then $\mathbf{u}_s(\mathbf{y})$ given by (2.22) satisfies*

(a) $\|\mathbf{u}_s(\mathbf{y}) - \mathbf{y}\|_0 \leq m + 1$.

(b) $\mathbf{u}_s(\mathbf{y})$ *is a $\frac{1}{d}$-PDA vector of problem (2.17) at $\mathbf{y}$.*

Proof. By (2.22), $\mathbf{u}_s(\mathbf{y}) - \mathbf{y} = \mathbf{d}_s(\mathbf{y})$, where $\mathbf{d}_s(\mathbf{y})$ is either $\mathbf{0}$, or given by (2.21):

$$\mathbf{d}_s(\mathbf{y}) = \frac{1}{\|\mathbf{r}\|_0} \bar{\mathbf{v}} \circ \mathbf{r} \tag{2.23}$$

with $\mathbf{r} = \mathbf{p}(\mathbf{y}) - \mathbf{y}$ and $\bar{\mathbf{v}}$ being a basic feasible solution of (2.19) satisfying

$$\langle \mathbf{c}, \bar{\mathbf{v}} \rangle \leq \langle \mathbf{c}, \mathbf{r}^\dagger \circ \mathbf{r} \rangle,$$

where

$$\mathbf{c} = \mathbf{r} \circ \nabla F(\mathbf{y}). \tag{2.24}$$

The case where $\|\mathbf{r}\|_0 = 0$ is trivial since in this case $\mathbf{y}$ is an optimal solution of problem (2.17), and hence $S(\mathbf{y}) = 0$, implying that the PDA condition (2.6) is satisfied. Assume then that $\mathbf{d}_s(\mathbf{y})$ is given by (2.23). Since $\bar{\mathbf{v}}$ is a basic feasible solution of (2.19), it has at most $m+1$ nonzero elements. Hence, $\mathbf{u}_s(\mathbf{y}) - \mathbf{y} = \mathbf{d}_s(\mathbf{y})$ has at most $m+1$ nonzero elements, proving (a).
To prove (b), we begin by establishing the feasibility of $\mathbf{u}_s(\mathbf{y})$ with respect to problem (2.17). We have $\langle \mathbf{1}, \bar{\mathbf{v}} \rangle \leq \|\mathbf{r}\|_0, \bar{\mathbf{v}} \geq \mathbf{0}$, and therefore, $\mathbf{0} \leq \frac{\bar{\mathbf{v}}}{\|\mathbf{r}\|_0} \leq \mathbf{1}$. Combining this with the obvious inequalitites $\boldsymbol{\ell} \leq \mathbf{p}(\mathbf{y}) \leq \mathbf{u}$ and $\boldsymbol{\ell} - \mathbf{y} \leq \mathbf{0} \leq \mathbf{u} - \mathbf{y}$, it follows that

$$\boldsymbol{\ell} - \mathbf{y} \leq \frac{\bar{\mathbf{v}}}{\|\mathbf{r}\|_0} \circ (\boldsymbol{\ell} - \mathbf{y}) \leq \underbrace{\frac{\bar{\mathbf{v}}}{\|\mathbf{r}\|_0} \circ (\mathbf{p}(\mathbf{y}) - \mathbf{y})}_{\mathbf{d}_s(\mathbf{y})} \leq \frac{\bar{\mathbf{v}}}{\|\mathbf{r}\|_0} \circ (\mathbf{u} - \mathbf{y}) \leq \mathbf{u} - \mathbf{y}, \tag{2.25}$$

and thus

$$\boldsymbol{\ell} \leq \mathbf{u}_s(\mathbf{y}) = \mathbf{y} + \mathbf{d}_s(\mathbf{y}) \leq \mathbf{u}. \tag{2.26}$$

In addition,

$$\mathbf{D}\mathbf{u}_s(\mathbf{y}) = \mathbf{D}\mathbf{y} + \mathbf{D}\mathbf{d}_s(\mathbf{y}) = \mathbf{b} + \frac{1}{\|\mathbf{r}\|_0}\mathbf{D}(\bar{\mathbf{v}} \circ \mathbf{r}) = \mathbf{b} + \frac{1}{\|\mathbf{r}\|_0}\mathbf{D}\mathrm{diag}(\mathbf{r})\bar{\mathbf{v}} = \mathbf{b} + \frac{1}{\|\mathbf{r}\|_0}\tilde{\mathbf{D}}\bar{\mathbf{v}} = \mathbf{b} + \mathbf{0} = \mathbf{b},$$

which combiend with (2.26) implies that $\mathbf{u}_s(\mathbf{y}) \in C$. We are left with the task of showing that inequality (2.6) is satisfied with $\gamma = d$ and $G = \delta_C$. For that, note that (recalling (2.24))

$$\langle \mathbf{c}, \bar{\mathbf{v}} \rangle \leq \langle \mathbf{c}, \mathbf{r}^\dagger \circ \mathbf{r} \rangle = \langle \nabla F(\mathbf{y}) \circ \mathbf{r}, \mathbf{r}^\dagger \circ \mathbf{r} \rangle = \langle \nabla F(\mathbf{y}), \mathbf{r} \rangle = \langle \nabla F(\mathbf{y}), \mathbf{p}(\mathbf{y}) - \mathbf{y} \rangle = -S(\mathbf{y}). \tag{2.27}$$

Finally,

$$\begin{array}{lll}
\langle \nabla F(\mathbf{y}), \mathbf{y} - \mathbf{u}_s(\mathbf{y}) \rangle & = -\langle \nabla F(\mathbf{y}), \mathbf{d}_s(\mathbf{y}) \rangle & \text{by (2.22)} \\
& = -\frac{1}{\|\mathbf{r}\|_0}\langle \nabla F(\mathbf{y}), \bar{\mathbf{v}} \circ \mathbf{r} \rangle & \text{by (2.23)} \\
& = -\frac{1}{\|\mathbf{r}\|_0}\langle \mathbf{r} \circ \nabla F(\mathbf{y}), \bar{\mathbf{v}} \rangle & \\
& = -\frac{1}{\|\mathbf{r}\|_0}\langle \mathbf{c}, \bar{\mathbf{v}} \rangle & \text{by (2.24)} \\
& \geq \frac{1}{\|\mathbf{r}\|_0} S(\mathbf{y}) & \text{by (2.27)} \\
& \geq \frac{1}{d} S(\mathbf{y}), & \|\mathbf{r}\|_0 \leq d
\end{array}$$

establishing the fact that $\mathbf{u}_s(\mathbf{y})$ is a $\frac{1}{d}$-PDA vector of problem (2.17) at $\mathbf{y}$. □

PDA-based algorithms: Based on the $\frac{1}{d}$-PDA vector $\mathbf{u}_s(\mathbf{y})$, we can define a variety of $\frac{1}{d}$-PDA methods depending on the choice of (i) the sets X^k and (ii) the update step (exact/local model). Below we describe four options. At iteration k, all the methods begin by computing $\mathbf{u}_s(\mathbf{y}^k)$. The first two possibilities fully exploit $\mathbf{u}_s(\mathbf{y}^k)$, and they actually resort to line search. The last options only use the information on the support of $\mathbf{d}_s(\mathbf{y}^k)$, and utilize the set of indices

$$J_k = \{i : \mathbf{u}_s(\mathbf{y}^k)_i = \mathbf{y}^k_i\}.$$

- *line segment minimization* ($X^k = [\mathbf{y}^k, \mathbf{u}_s(\mathbf{y}^k)]$, exact update)

  $$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k),$$

  where $t_k \in \mathrm{argmin}_{0 \leq t \leq 1} F(\mathbf{y}^k + t(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k))$.

- *ray minimization* ($X^k = \{\mathbf{y}^k + t(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k) : t \geq 0\}$, exact update)
$$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k),$$
where $t_k \in \operatorname{argmin}_{t\geq 0}\{F(\mathbf{y}^k + t(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k)) : \boldsymbol{\ell} \leq \mathbf{y}^k + t(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k) \leq \mathbf{u}\}$.
- *block exact minimization* ($X^k = \{\mathbf{y} \in C : \mathbf{y}_i = \mathbf{y}_i^k, i \in J_k\}$, exact update)
$$\mathbf{y}^{k+1} \in \operatorname{argmin}\{F(\mathbf{y}) : \mathbf{y} \in C, \mathbf{y}_i = \mathbf{y}_i^k, i \in J_k\}.$$
- *block projected gradient* ($X^k = \{\mathbf{y} \in C : \mathbf{y}_i = \mathbf{y}_i^k, i \in J_k\}$, local model update)
$$\mathbf{y}^{k+1} = P_{X^k}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla F(\mathbf{y}^k)\right).$$

Rank reduction. Given $\mathbf{p}(\bar{\mathbf{y}})$, computing $\bar{\mathbf{v}}$ as given by step (ii) of sparseDir can be done by finding a basic feasible optimal solution of the auxiliary linear program (2.19). This may be a prohibitive additional cost in many settings. Alternatively, it is possible to compute $\bar{\mathbf{v}}$ by a classical rank reduction technique. As outlined in Remark 2.4, $\mathbf{v} = \mathbf{r}^\dagger \circ \mathbf{r}$ is always feasible for the auxiliary linear program. Starting with $\mathbf{v}$, it is possible to find $\mathbf{s} \neq \mathbf{0}$ with support included in that of $\mathbf{v}$, such that $\tilde{\mathbf{D}}\mathbf{s} = \mathbf{0}$, $\langle \mathbf{s}, \mathbf{1}\rangle = 0$ and $\langle \mathbf{s}, \mathbf{c}\rangle \leq 0$ using Gaussian elimination in $O(m^3)$ operations. One can then preform a step in the direction $\mathbf{s}$ to remove a coordinate from the support of $\mathbf{v}$. After at most d iterations of this procedure, a basic feasible solution for the auxiliary program is found. The total cost for this rank reduction is $O(dm^3)$ which remains linear in the dimension of the problem.

Application to the dual SVM. One important motivation for this type of working set techniques is that they provide scalable algorithms for solving the SVM dual problem (see Section 4.2). In this case, the dimension of the problem is equal to the number of examples in the training set and we have only one linear equality constraint in C. This motivates the use of SMO-type working set methods which update only pairs of variables at each iterations. The $\frac{1}{d}$-PDA construction that we propose in this section can be used here. The most costly steps of the working set selection rule that we propose here are the computation of $\mathbf{p}(\bar{\mathbf{y}})$ and the computation of a basic feasible solution of the auxiliary linear program. As we have seen in the previous paragraph, the latter can be done, given $\mathbf{p}(\bar{\mathbf{y}})$, in linear time in d, the dimension of the problem. For the dual SVM, the computation of $\mathbf{p}(\bar{\mathbf{y}})$ is a fractional knapsack problem which can be solved in a number of operations which is linear in the dimension (see Section 4.2). This, combined with the previous rank reduction scheme, gives a completely explicit construction of a $\frac{1}{d}$-PDA vector for the dual SVM whose construction has complexity which is linear in the number of examples. A similar but much less explicit construction was proposed in [25] for the same model.

2.3 Sublinear rate of convergence analysis

In this section we prove a sublinear rate of convergence of the $\frac{1}{\gamma}$-PDA method. The analysis is based on the artificial introduction of the diminishing stepsize, $t_k = 2\gamma/(k + 2\gamma)$, which is due to [21]. We begin with the following recursion that characterizes sequences produced by the $\frac{1}{\gamma}$-PDA algorithm. The arguments behind the proof of the lemma have become fairly standard in the analysis of conditional gradient-type methods.

Lemma 2.6. *Let $\{\mathbf{y}^k\}_{k\geq 0}$ be the sequence generated by the $\frac{1}{\gamma}$-PDA method. Fix an arbitrary sequence $\{t_k\}_{k\geq 0}$ such that for any $k \geq 0$, we have $0 \leq t_k \leq 1$. Then for any $k \geq 0$, we have*
$$H(\mathbf{y}^{k+1}) \leq H(\mathbf{y}^k) - \frac{t_k}{\gamma}S(\mathbf{y}^k) + \frac{L_k D^2}{2}t_k^2,$$
where $D = \operatorname{diam}(\operatorname{dom} G)$.

Proof. Using Definition 2.1, for any $k \geq 0$, define $\mathbf{u}^k = \mathbf{u}(\mathbf{y}^k) \in X^k$ as the $\frac{1}{\gamma}$-PDA vector that satisfies

$$\begin{aligned} [\mathbf{y}^k, \mathbf{u}^k] &\subseteq X^k, \\ \frac{1}{\gamma} S(\mathbf{y}^k) &\leq \langle \nabla F(\mathbf{y}^k), \mathbf{y}^k - \mathbf{u}^k \rangle + G(\mathbf{y}^k) - G(\mathbf{u}^k). \end{aligned} \tag{2.28}$$

We have

$$\begin{aligned} &\min_{\mathbf{y} \in X^k} \left\{ Q_{L_k}(\mathbf{y}, \mathbf{y}^k) + G(\mathbf{y}) \right\} \leq \min_{\mathbf{y} \in [\mathbf{y}^k, \mathbf{u}^k]} \left\{ Q_{L_k}(\mathbf{y}, \mathbf{y}^k) + G(\mathbf{y}) \right\} \\ &= \min_{0 \leq t \leq 1} \left\{ Q_{L_k}(t\mathbf{u}^k + (1-t)\mathbf{y}^k, \mathbf{y}^k) + G(t\mathbf{u}^k + (1-t)\mathbf{y}^k) \right\} \\ &\leq Q_{L_k}(t_k \mathbf{u}^k + (1-t_k)\mathbf{y}^k, \mathbf{y}^k) + G(t_k \mathbf{u}^k + (1-t_k)\mathbf{y}^k) \\ &= F(\mathbf{y}^k) + t_k \langle \nabla F(\mathbf{y}^k), \mathbf{u}^k - \mathbf{y}^k \rangle + t_k^2 \frac{L_k}{2} \|\mathbf{y}^k - \mathbf{u}^k\|^2 + G(t_k \mathbf{u}^k + (1-t_k)\mathbf{y}^k) \\ &\leq F(\mathbf{y}^k) + G(\mathbf{y}^k) + t_k \left(\langle \nabla F(\mathbf{y}^k), \mathbf{u}^k - \mathbf{y}^k \rangle + G(\mathbf{u}^k) - G(\mathbf{y}^k) \right) + t_k^2 \frac{L_k \|\mathbf{y}^k - \mathbf{u}^k\|^2}{2} \quad (2.29) \\ &\leq F(\mathbf{y}^k) + G(\mathbf{y}^k) - \frac{t_k}{\gamma} S(\mathbf{y}^k) + t_k^2 \frac{L_k D^2}{2}, \quad (2.30) \end{aligned}$$

where the convexity of G was used in (2.29), and (2.30) follows by (2.28) and the definition of the diameter. Finally, note that for both rules (2.11) or (2.12) we have

$$F(\mathbf{y}^{k+1}) + G(\mathbf{y}^{k+1}) \leq \min_{\mathbf{y} \in X^k} \left[Q_{L_k}(\mathbf{y}, \mathbf{y}^k) + G(\mathbf{y}) \right]. \tag{2.31}$$

Indeed, since in the exact minimization step we have $L_k \equiv L$, it follows that in this case

$$F(\mathbf{y}) \leq Q_{L_k}(\mathbf{y}, \mathbf{y}^k) \text{ for any } \mathbf{y} \in X^k,$$

and hence also that

$$F(\mathbf{y}) + G(\mathbf{y}) \leq Q_{L_k}(\mathbf{y}, \mathbf{y}^k) + G(\mathbf{y}) \text{ for any } \mathbf{y} \in X^k. \tag{2.32}$$

Taking the minimum of both sides over $\mathbf{y} \in X^k$ will yield (2.31). In the local model update setting, using (2.14), we can write

$$F(\mathbf{y}^{k+1}) + G(\mathbf{y}^{k+1}) \leq Q_{L_k}(\mathbf{y}^{k+1}, \mathbf{y}^k) + G(\mathbf{y}^{k+1}) = \min_{\mathbf{y} \in X^k} \left[Q_{L_k}(\mathbf{y}, \mathbf{y}^k) + G(\mathbf{y}) \right],$$

which is the same as (2.31). Finally, combining (2.30) and (2.31), the desired result follows. □

We now need one technical lemma in order to prove the sublinear convergence rate.

Lemma 2.7. *Suppose that $\gamma \geq 1$ and $C \geq 0$. Let $\{a_k\}_{k \geq 0}$ and $\{b_k\}_{k \geq 0}$ be two sequences such that $0 \leq a_k \leq b_k$ for any $k \geq 0$. Set $t_k = \frac{2\gamma}{k+2\gamma}$, and assume in addition that*

$$a_{k+1} \leq a_k - \frac{t_k}{\gamma} b_k + \frac{C}{2} t_k^2. \tag{2.33}$$

Then for any $k \geq 0$,

$$\frac{\sum_{i=0}^k (b_i - a_i)(i + 2\gamma - 1)}{\sum_{i=0}^k (i + 2\gamma - 1)} + a_{k+1} \leq \frac{2\gamma}{k+2\gamma} \left(\frac{2\gamma - 2}{k+1} a_0 + C\gamma \right). \tag{2.34}$$

Proof. From inequality (2.33) and the definition of t_k, we get that for any $i \geq 0$

$$b_i - a_i \leq \left(\frac{\gamma}{t_i} - 1\right) a_i - \frac{\gamma}{t_i} a_{i+1} + \frac{C\gamma}{2} t_i = \frac{i+2\gamma-2}{2} a_i - \frac{i+2\gamma}{2} a_{i+1} + \frac{C\gamma^2}{i+2\gamma}. \tag{2.35}$$

Multiplying inequality (2.35) by $i + 2\gamma - 1 \geq 0$, we get

$$\begin{aligned}(b_i - a_i)(i+2\gamma-1) &\leq \frac{(i+2\gamma-2)(i+2\gamma-1)}{2} a_i - \frac{(i+2\gamma)(i+2\gamma-1)}{2} a_{i+1} + C\gamma^2 \frac{i+2\gamma-1}{i+2\gamma} \\ &\leq \frac{(i+2\gamma-2)(i+2\gamma-1)}{2} a_i - \frac{(i+2\gamma)(i+2\gamma-1)}{2} a_{i+1} + C\gamma^2. \end{aligned} \tag{2.36}$$

Summing inequality (2.36) for $i = 0, 1, \ldots, k$ gives

$$\sum_{i=0}^{k} (b_i - a_i)(i+2\gamma-1) \leq \frac{(2\gamma-2)(2\gamma-1)}{2} a_0 - \frac{(k+2\gamma)(k+2\gamma-1)}{2} a_{k+1} + C\gamma^2(k+1). \tag{2.37}$$

Dividing both sizes of (2.37) by $\frac{k+1}{2}(k+2\gamma) \geq 0$, yields for any $k \geq 0$,

$$\begin{aligned}\frac{\sum_{i=0}^{k}(b_i - a_i)(i+2\gamma-1)}{\frac{k+1}{2}(k+2\gamma)} + \frac{k+2\gamma-1}{k+1} a_{k+1} &\leq \frac{2\gamma}{k+2\gamma}\left(\frac{(2\gamma-2)(2\gamma-1)}{2\gamma(k+1)} a_0 + C\gamma\right) \\ &\leq \frac{2\gamma}{k+2\gamma}\left(\frac{2\gamma-2}{k+1} a_0 + C\gamma\right). \end{aligned} \tag{2.38}$$

Inequality (2.34) now follows by the fact that $\frac{k+2\gamma-1}{k+1} \geq 1$ (since $\gamma \geq 1$), and the relation

$$\sum_{i=0}^{k} (i+2\gamma-1) = \frac{k+1}{2}(k+4\gamma-2) \geq \frac{k+1}{2}(k+2\gamma),$$

where the inequality also follows by the fact that $\gamma \geq 1$. □

We will now utilize Lemma 2.7 to show the sublinear rate of convergence of the sequence of function values generated by the $\frac{1}{\gamma}$-PDA method. By Lemma 2.6 and (2.15) it follows that relation (2.33) holds with $a_k = H(\mathbf{y}^k) - H^*, b_k = S(\mathbf{y}^k)$ and $C = K$, where K is chosen as follows:

$$K = \begin{cases} L \cdot \operatorname{diam}(\operatorname{dom} G)^2, & \text{exact minimization, or local model with constant stepsize,} \\ \max\{\eta L, \bar{L}\} \cdot \operatorname{diam}(\operatorname{dom} G)^2, & \text{local model with backtracking.} \end{cases} \tag{2.39}$$

By (2.4) we also have that $a_k \leq b_k$ for all $k \geq 0$. We can thus invoke Lemma 2.7 and obtain the following result.

Lemma 2.8. *Let $\{\mathbf{y}^k\}_{k\geq 0}$ be the sequence generated by $\frac{1}{\gamma}$-PDA method. Then for any $k \geq 0$*

$$\frac{\sum_{i=0}^{k}\left(S(\mathbf{y}^i) - [H(\mathbf{y}^i) - H^*]\right)(i+2\gamma-1)}{\sum_{i=0}^{k}(i+2\gamma-1)} + H(\mathbf{y}^{k+1}) - H^* \leq \frac{2\gamma}{k+2\gamma}\left(\frac{2\gamma-2}{k+1}(H(\mathbf{y}^0) - H^*) + K\gamma\right), \tag{2.40}$$

where K is given in (2.39).

Since $S(\mathbf{y}^i) \geq H(\mathbf{y}^i) - H^*$, we can deduce the sublinear rate of convergence of the sequence of function values generated by $\frac{1}{\gamma}$-PDA method.

Theorem 2.9. *Let $\{\mathbf{y}^k\}_{k\geq 0}$ be the sequence generated by the $\frac{1}{\gamma}$-PDA method. Then for any $k \geq 0$*

$$H(\mathbf{y}^{k+1}) - H^* \leq \frac{2\gamma}{k+2\gamma}\left(\frac{2\gamma-2}{k+1}(H(\mathbf{y}^0) - H^*) + K\gamma\right),$$

where K is given in (2.39).

Remark 2.10 (Dependancy in γ)**.** *It can be seen that the rate given in Theorem 2.9 is increasing as a function of γ. This is consistent with the fact that a $\frac{1}{\gamma}$-PDA method is also a $\frac{1}{\gamma'}$-PDA method for any $\gamma' \geq \gamma \geq 1$ and highlights the influence of the degree of approximation of the method.*

3 The dual-based γ-PDA method

3.1 Model, duality and basic properties

In this section we will present and analyze a method that employs the $\frac{1}{\gamma}$-PDA framework on a dual problem. The primal optimization model that will be analyzed has the form

$$\bar{p} \equiv \min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{A}\mathbf{x}) + g(\mathbf{B}\mathbf{x})\}, \tag{3.1}$$

where $\mathbf{A} \in \mathbb{R}^{r\times n}$ and $\mathbf{B} \in \mathbb{R}^{q\times n}$. We will make the following standing assumption on the problem's data.

Assumption 2.

(A) $\mathbf{A}$ *has full row rank*

(B) $f : \mathbb{R}^r \to \mathbb{R} \cup (-\infty, \infty]$ *is proper closed and μ-strongly convex.*

(C) $g : \mathbb{R}^q \to \mathbb{R}$ *is closed, convex and has a Lipschitz constant L_g.*

(D) $\operatorname{dom} g^*$ *is closed.*

(E) One of the following holds:

- *(i) g is polyhedral and* $\operatorname{im}(\mathbf{A}^T) \cap \mathbf{B}^T \operatorname{dom}(g^*)$ *is nonempty.*
- *(ii)* $\operatorname{im}(\mathbf{A}^T) \cap \mathbf{B}^T \operatorname{ridom}(g^*)$ *is nonempty, where* $\operatorname{ridom}(g^*)$ *is the relative interior of the domain of g^*.*

Several properties can be readily deduced from Assumption 2:

- $f^* : \mathbb{R}^r \to \mathbb{R}$ is convex and $\frac{1}{\mu}$-smooth (by (B)).
- $g^* : \mathbb{R}^q \to (\infty, \infty]$ is proper closed and convex and its domain is contained in a ball of radius L_g centered at the origin (by (C)).
- If *(E.i)* is satisfied, then g^* is also polyhedral and $\operatorname{dom} g^*$ is a polytope.

Under Assumption 2, problem (3.1) does not fit the general model (2.1) that can be tackled using PDA methods. We will tackle problem (3.1) through duality, and at the same time explore primal-dual properties of PDA methods. The Lagrangian dual of problem (3.1) can be written as

$$\begin{array}{rll} \bar{q} \equiv & \max & -f^*(\mathbf{w}) - g^*(\mathbf{z}) \\ & \text{s.t.} & \mathbf{A}^T\mathbf{w} + \mathbf{B}^T\mathbf{z} = 0, \\ & & \mathbf{w} \in \mathbb{R}^r, \mathbf{z} \in \mathbb{R}^q. \end{array} \tag{3.2}$$

Remark 3.1 (Derivation of the dual)**.** *We artificially introduce additional variables and equality constraints* $\mathbf{x}_1 = \mathbf{A}\mathbf{x}$ *and* $\mathbf{x}_2 = \mathbf{B}\mathbf{x}$ *in problem (3.1). The Lagrangian function then has the form*

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2; \mathbf{w}, \mathbf{z}) &= f(\mathbf{x}_1) + g(\mathbf{x}_2) + \langle \mathbf{w}, \mathbf{A}\mathbf{x} - \mathbf{x}_1 \rangle + \langle \mathbf{z}, \mathbf{B}\mathbf{x} - \mathbf{x}_2 \rangle \\ &= f(\mathbf{x}_1) - \langle \mathbf{w}, \mathbf{x}_1 \rangle + g(\mathbf{x}_2) - \langle \mathbf{z}, \mathbf{x}_2 \rangle + \left\langle \mathbf{A}^T\mathbf{w} + \mathbf{B}^T\mathbf{z}, \mathbf{x} \right\rangle.\end{aligned}$$

Expression (3.2) follows by partial minimization of the Lagrangian with respect to $\mathbf{x}_1$*,* $\mathbf{x}_2$ *and* $\mathbf{x}$*.*

We will denote the feasible set of problem (3.2) by X:

$$X \equiv \left\{ (\mathbf{w},\, \mathbf{z}) : \mathbf{z} \in \operatorname{dom}(g^*),\ \mathbf{A}^T\mathbf{w} + \mathbf{B}^T\mathbf{z} = 0 \right\}.$$

Some elementary arguments can be used to show the compactness of X.

Lemma 3.2. *X is compact.*

Proof. The closedness of X follows by the closedness of $\operatorname{dom} g^*$. Moreover, for any $(\mathbf{w}, \mathbf{z}) \in X$, we have $\mathbf{z} \in \operatorname{dom} g^*$ and hence in particular

$$\|\mathbf{z}\| \leq L_g. \tag{3.3}$$

In addition, by the relation $\mathbf{A}^T\mathbf{w} = -\mathbf{B}^T\mathbf{z}$, it follows that

$$\mathbf{w}^T\mathbf{A}\mathbf{A}^T\mathbf{w} = \|\mathbf{A}^T\mathbf{w}\|^2 = \|\mathbf{B}^T\mathbf{z}\|^2 \leq \|\mathbf{B}\|^2 L_g^2,$$

which implies that

$$\|\mathbf{w}\|^2 \leq \frac{\|\mathbf{B}\|^2 L_g^2}{\lambda_{\min}(\mathbf{A}\mathbf{A}^T)}.$$

Combining this with (3.3) implies that X is bounded, and the compactness is established. □

The next lemma shows, using general duality theory, that the optimal values of the primal-dual pair of problems (3.1) and (3.2) are the same, and the optimal values of both problems are attained.

Lemma 3.3. *The optimal values,* $\bar{p}$ *and* $\bar{q}$*, of problems (3.1) and (3.2) are finite, attained and equal.*

Proof. Problem (3.2) consists of maximizing an upper semicontinuous function over a nonempty compact set (Lemma 3.2), and hence its optimal value is attained. In addition, by duality theory [29], it follows that under the regularity condition (E) in Assumption 2, the optimal value of problem (3.1) is the same as the optimal value of problem (3.2), and that the minimum is attained. □

We will also consider in our analysis the matrix

$$\mathbf{P} \equiv \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}. \tag{3.4}$$

This matrix is associated with the orthogonal projection operator on the row space of $\mathbf{A}$ in the sense that (see [32, Section 3.3])

$$\mathbf{P}\mathbf{x} = \operatorname{argmin}_{\mathbf{y}} \left\{ \|\mathbf{x} - \mathbf{y}\| : \mathbf{y} \in \operatorname{im}(\mathbf{A}^T) \right\}.$$

A useful property of the matrix $\mathbf{P}$ is described in the following lemma.

Lemma 3.4. *For any* $\mathbf{v} \in \mathbb{R}^n$ *and* $\mathbf{w} \in \mathbb{R}^r$*,* $\mathbf{v} = \mathbf{A}^T\mathbf{w}$ *if and only if* $\mathbf{v} = \mathbf{P}\mathbf{v}$ *and* $\mathbf{w} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{v}$*.*

Proof. If $\mathbf{v} = \mathbf{A}^T\mathbf{w}$, then $\mathbf{P}\mathbf{v} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{A}^T\mathbf{w} = \mathbf{A}^T\mathbf{w} = \mathbf{v}$. In addition, since $\mathbf{v} = \mathbf{A}^T\mathbf{w}$, then $\mathbf{w}$ is the solution of the least squares problem $\min_{\mathbf{u}} \|\mathbf{v} - \mathbf{A}^T\mathbf{u}\|^2$, meaning that $\mathbf{w} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{v}$. Conversely, if $\mathbf{v} = \mathbf{P}\mathbf{v}$ and $\mathbf{w} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{v}$, then $\mathbf{A}^T\mathbf{w} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{v} = \mathbf{P}\mathbf{v} = \mathbf{v}$. □

By Lemma 3.4, the equality $\mathbf{B}^T\mathbf{z} + \mathbf{A}^T\mathbf{w} = \mathbf{0}$ holds if and only if $(\mathbf{I} - \mathbf{P})\mathbf{B}^T\mathbf{z} = \mathbf{0}$ and $\mathbf{w} = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}$, and thus the dual problem can be recast as

$$\begin{array}{ll} \max & -f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}) - g^*(\mathbf{z}) \\ \text{s.t.} & (\mathbf{I} - \mathbf{P})\mathbf{B}^T\mathbf{z} = 0, \\ & \mathbf{z} \in \mathbb{R}^q, \end{array} \tag{3.5}$$

or in minimization form:

$$\begin{array}{ll} \min & f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}) + g^*(\mathbf{z}) \\ \text{s.t.} & (\mathbf{I} - \mathbf{P})\mathbf{B}^T\mathbf{z} = 0, \\ & \mathbf{z} \in \mathbb{R}^q. \end{array} \tag{3.6}$$

By the fact that the optimal value of the dual problem (3.2) is finite and attained, it follows that this is also the case for problem (3.6). In addition, since we passed from a maximum to a minimum problem by multiplying the objective function by -1, it follows that the optimal value of problem (3.6) is $-\bar{q}$.

Problem (3.6) fits the general model (2.1) with

$$\begin{array}{rcl} F(\mathbf{z}) & = & F_1(\mathbf{z}) \equiv f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}), \\ G(\mathbf{z}) & = & G_1(\mathbf{z}) \equiv g^*(\mathbf{z}) + \delta_{\{\mathbf{p}:(\mathbf{I}-\mathbf{P})\mathbf{p}=\mathbf{0}\}}(\mathbf{z}). \end{array}$$

Thus, problem (3.6) can be written as

$$\min_{\mathbf{z}\in\mathbb{R}^q}\{H_1(\mathbf{z}) \equiv F_1(\mathbf{z}) + G_1(\mathbf{z})\}. \tag{3.7}$$

The optimality measure associated with (3.7) is given by

$$S_1(\mathbf{z}) = \max_{\mathbf{p}} \left\{ \langle \nabla F_1(\mathbf{z}), \mathbf{z} - \mathbf{p} \rangle + G_1(\mathbf{z}) - G_1(\mathbf{p}) \right\}. \tag{3.8}$$

3.2 The dual-based $\frac{1}{\gamma}$-PDA method

The specific choice $F = F_1$ and $G = G_1$ satisfies the assertions in Assumption 1 with

$$L = \frac{\|(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\|^2}{\mu}, \tag{3.9}$$

and we can thus invoke the $\frac{1}{\gamma}$-PDA method to solve problem (3.6). Below we describe the dual-based $\frac{1}{\gamma}$-PDA method, along with a specification of the primal sequence $\{\mathbf{x}^k\}_{k\geq 0}$.

Dual-Based $\frac{1}{\gamma}$-PDA Method:

Initialization. Pick $\mathbf{z}^0$ satisfying $(\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{z}^0=\mathbf{0}, \mathbf{z}^0\in\operatorname{dom} g^*$.

General Step. For $k=0,1,2\ldots$,

(i) – Choose $\mathbf{u}(\mathbf{z}^k)$- a $\frac{1}{\gamma}$-PDA vector of H_1 at $\mathbf{z}^k$.

– Choose a compact set Z^k for which $[\mathbf{z}^k,\mathbf{u}(\mathbf{z}^k)]\subseteq Z^k$.

(ii) Perform one of the following:

$$\text{Local model update: } \mathbf{z}^{k+1}=\operatorname{prox}_{\frac{1}{L_k}G_1+\delta_{Z^k}}\left(\mathbf{z}^k-\frac{1}{L_k}\nabla F_1(\mathbf{z}^k)\right)$$

$$\text{Exact update: } \mathbf{z}^{k+1}=\operatorname{argmin}_{\mathbf{z}\in Z^k}F_1(\mathbf{z})+G_1(\mathbf{z})$$

(iii) Set $\mathbf{w}^k=-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}^k$ and compute $\mathbf{s}^k$ by one of the following formulas:

$$\text{Averaging: } \mathbf{s}^k=\frac{1}{\sum_{i=0}^k(i+2\gamma-1)}\sum_{i=0}^k(i+2\gamma-1)\nabla f^*(\mathbf{w}^i)$$

$$\text{Best iterate: } \mathbf{s}^k=\nabla f^*(\mathbf{w}^{k_0}),\ k_0\in\operatorname{argmin}_{i=0,1,\ldots k}\{S_1(\mathbf{z}^i)-H_1(\mathbf{z}^i)\}$$

(iv) Compute

$$\mathbf{x}^k\in\operatorname{argmin}_{\mathbf{x}}\left\{g(\mathbf{B}\mathbf{x}):\mathbf{A}\mathbf{x}=\mathbf{s}^k\right\}. \tag{3.10}$$

Remark 3.5 (Primal sequence)**.** *Steps (iii) and (iv) are only required if we are interested in estimating a primal sequence, $\{\mathbf{x}^k\}_{k\geq 0}$. In this case, step (iv) needs to be preformed only at the last iteration. The second option in step (iii) requires the evaluation of S_1 which is given in (3.8). Many examples of PDA methods rely on this evaluation in order to compute predicted decrease directions (generalized conditional gradient, greedy coordinate descent and block descent for linearly constrained problems). Therefore, in these cases the computation of S_1 can be reused in this step.*

Remark 3.6 (Primal feasibility)**.** *Note that since $\nabla f^*(\mathbf{v})\in\operatorname{dom} f$ for any $\mathbf{v}$, it follows by the definition of $\mathbf{s}^k$ and the convexity of f that $\mathbf{s}^k\in\operatorname{dom} f$ for any k. By the definition of $\mathbf{x}^k$, we have $\mathbf{A}\mathbf{x}^k\in\operatorname{dom} f$, implying that the dual-based $\frac{1}{\gamma}$-PDA method is a primal feasible method, which is actually not a common situation in dual-based methods.*

Remark 3.7 (Online computation)**.** *In the case of averaging, $\mathbf{s}^k$ satisfies the recurrence relation $\mathbf{s}^k=(1-w_k)\mathbf{s}^{k-1}+w_k\nabla f^*(\mathbf{w}^k)$ for any $k\geq 1$, where $w_k=\frac{2(k+2\gamma-1)}{(k+1)(k+4\gamma-2)}$ and is therefore amenable to efficient online computation.*

3.3 Convergence analysis

Convergence of the dual objective function evaluated at the sequence of dual variables can be deduced by Theorem 2.9. However, in many cases (like in the examples discussed in Section 4), we are interested in the rate of convergence of the sequence of primal function values, $\{f(\mathbf{A}\mathbf{x}^k)+g(\mathbf{B}\mathbf{x}^k)\}_{k\geq 0}$ to the optimal value $\bar{p}$. To accomplish this task, we will investigate the optimality measure given in (3.8). The following key theorem shows a representation of the optimality measure that will enable us later on to obtain a rate of convergence of the dual-based $\frac{1}{\gamma}$-PDA method.

Theorem 3.8. *For any $\mathbf{z}\in\operatorname{dom} g^*$*

$$S_1(\mathbf{z})=\min_{\mathbf{A}\mathbf{x}=\nabla f^*(\mathbf{w})}g(\mathbf{B}\mathbf{x})+f(\nabla f^*(\mathbf{w}))+g^*(\mathbf{z})+f^*(\mathbf{w})$$

with $\mathbf{w} = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}$.

In order to prove the theorem, we will need the following strong duality result.

Lemma 3.9. *For any* $\mathbf{s} \in \mathbb{R}^r$*, we have*

$$\begin{array}{rcl} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & g(\mathbf{B}\mathbf{x}) & = \max\limits_{\mathbf{p}\in\mathbb{R}^q} \ \left\langle (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{s}, \mathbf{A}\mathbf{B}^T\mathbf{p}\right\rangle - g^*(\mathbf{p}) \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{s} & \text{s.t.} \quad (\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{p} = \mathbf{0} \end{array} \tag{3.11}$$

and both optimal values are finite and attained.

Proof. We begin by rewriting the left-hand problem in (3.11) as

$$\min_{\mathbf{u}\in\mathbb{R}^q, \mathbf{x}\in\mathbb{R}^n} \{g(\mathbf{u}) : \mathbf{B}\mathbf{x} = \mathbf{u}, \mathbf{A}\mathbf{x} = \mathbf{s}\}. \tag{3.12}$$

The Lagrangian of the problem is

$$\mathcal{L}(\mathbf{u}, \mathbf{x}; \mathbf{p}, \mathbf{w}) = g(\mathbf{u}) + \langle \mathbf{p}, \mathbf{B}\mathbf{x} - \mathbf{u}\rangle + \langle \mathbf{w}, \mathbf{s} - \mathbf{A}\mathbf{x}\rangle.$$

Minimizing with respect to $\mathbf{u}$ and $\mathbf{x}$, we obtain the following dual problem:

$$\max_{\mathbf{p}} \{-g^*(\mathbf{p}) + \langle \mathbf{w}, \mathbf{s}\rangle : \mathbf{B}^T\mathbf{p} - \mathbf{A}^T\mathbf{w} = \mathbf{0}\}. \tag{3.13}$$

The feasible set of problem (3.13) is compact since $\operatorname{dom} g^*$ is compact and the fact that the matrix $\mathbf{A}$ has full row rank (see also the argument in the proof of Lemma 3.2). Therefore, since $-g^*$ is upper semicontinuous, it follows that the maximum in problem (3.13) is attained. By the regularity condition (E) in Assumption 2, it follows that strong duality holds meaning that the optimal values of problems (3.12) and (3.13) are equal and the optimal value of (3.12) is attained. Invoking Lemma 3.4 with $\mathbf{v} = \mathbf{B}^T\mathbf{p}$, we obtain that the equality $\mathbf{B}^T\mathbf{p} - \mathbf{A}^T\mathbf{w} = \mathbf{0}$ is equivalent to $(\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{p} = \mathbf{0}$ and $\mathbf{w} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{p}$, which readily implies that problem (3.13) can be reduced to

$$\max_{\mathbf{p}} \{-g^*(\mathbf{p}) + \langle \mathbf{A}\mathbf{B}^T\mathbf{p}, (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{s}\rangle : (\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{p} = \mathbf{0}\},$$

which proves the desired result. □

Equipped with Lemma 3.9, we can now prove Theorem 3.8.

Proof of Theorem 3.8. Since

$$\nabla F_1(\mathbf{z}) = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\nabla f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}),$$

it follows that S_1 given by (3.8) can rewritten as

$$\begin{aligned} S_1(\mathbf{z}) &= \max_{\mathbf{p}:(\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{p}=\mathbf{0}} \left\{\left\langle -\mathbf{B}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\nabla f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}), \mathbf{z}-\mathbf{p}\right\rangle + g^*(\mathbf{z}) - g^*(\mathbf{p})\right\} \\ &= \max_{\mathbf{p}:(\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{p}=\mathbf{0}} \left\{\left\langle (\mathbf{A}\mathbf{A}^T)^{-1}\nabla f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}), \mathbf{A}\mathbf{B}^T(\mathbf{p}-\mathbf{z})\right\rangle + g^*(\mathbf{z}) - g^*(\mathbf{p})\right\}. \end{aligned}$$

Invoking Lemma 3.9 with $\mathbf{s} = \nabla f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z})$, we obtain that S_1 can be written as

$$\begin{aligned} S_1(\mathbf{z}) &= \left[\min_{\mathbf{A}\mathbf{x}=\nabla f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z})} g(\mathbf{B}\mathbf{x})\right] + g^*(\mathbf{z}) - \left\langle (\mathbf{A}\mathbf{A}^T)^{-1}\nabla f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}), \mathbf{A}\mathbf{B}^T\mathbf{z}\right\rangle \\ &= \min_{\mathbf{A}\mathbf{x}=\nabla f^*(\mathbf{w})} g(\mathbf{B}\mathbf{x}) + g^*(\mathbf{z}) + \left\langle \nabla f^*(\mathbf{w}), \mathbf{w}\right\rangle, \end{aligned}$$

where $\mathbf{w} = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}$. Since f is proper closed and convex,

$$\langle \nabla f^*(\mathbf{w}), \mathbf{w}\rangle = f(\nabla f^*(\mathbf{w})) + f^*(\mathbf{w}).$$

Thus,

$$S_1(\mathbf{z}) = \min_{\mathbf{A}\mathbf{x}=\nabla f^*(\mathbf{w})} g(\mathbf{B}\mathbf{x}) + f(\nabla f^*(\mathbf{w})) + g^*(\mathbf{z}) + f^*(\mathbf{w}),$$

as asserted. □

Theorem 3.10 (primal-dual convergence)**.** *Let $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{z}^k\}_{k\geq 0}$ be the sequences generated by the $\frac{1}{\gamma}$-PDA method employed on problem (3.6). Then for any k, $\mathbf{z}^k$ is dual feasible, $\mathbf{x}^k$ is primal feasible and*

$$f(\mathbf{A}\mathbf{x}^k) + g(\mathbf{B}\mathbf{x}^k) + H_1(\mathbf{z}^{k+1}) \leq \frac{2\gamma}{k+2\gamma}\left(\frac{2\gamma-2}{k+1}(H_1(\mathbf{z}^0)+\bar{p}) + 4\tilde{K}\gamma\right), \tag{3.14}$$

where

$$\tilde{K} = \begin{cases} \frac{\|(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\|^2}{\mu} L_g^2, & \text{exact minimization or local model with constant stepsize,} \\ \max\left\{\eta\frac{\|(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\|^2}{\mu}, \bar{L}\right\} L_g^2, & \text{local model with backtracking.} \end{cases} \tag{3.15}$$

Proof. For any $k \geq 0$ the vector $\mathbf{z}^k$ is dual feasible by its construction and $\mathbf{x}^k$ is primal feasible as stated in Remark 3.6. Invoking Lemma 2.8, using the expression for L given in (3.9), the fact that $\operatorname{diam}(\operatorname{dom}(g^*)) \leq 2L_g$ and the fact that $H^* = -\bar{q} = -\bar{p}$ (where H^* is the optimal value of (3.7)), we obtain for any $k \geq 0$ (after cancelation of the constant term $\bar{p}$, or H^* in Lemma 2.8)

$$\frac{\sum_{i=0}^k \left(S_1(\mathbf{z}^i) - H_1(\mathbf{z}^i)\right)(i+2\gamma-1)}{\sum_{i=0}^k (i+2\gamma-1)} + H_1(\mathbf{z}^{k+1}) \leq \frac{2\gamma}{k+2\gamma}\left(\frac{(2\gamma-2)}{k+1}(H_1(\mathbf{z}^0)+\bar{p}) + 4\tilde{K}\gamma\right). \tag{3.16}$$

Using Theorem 3.8, setting $\mathbf{w}^i = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}^i$, we have for any $i = 0, 1, \ldots, k$

$$\begin{aligned} S_1(\mathbf{z}^i) - H_1(\mathbf{z}^i) &= S_1(\mathbf{z}^i) - \left(f^*(\mathbf{w}^i) + g^*(\mathbf{z}^i)\right) \\ &= \min_{\mathbf{A}\mathbf{x}=\nabla f^*(\mathbf{w}^i)} g(\mathbf{B}\mathbf{x}) + f(\nabla f^*(\mathbf{w}^i)) + f^*(\mathbf{w}^i) + g^*(\mathbf{z}^i) - \left[f^*(\mathbf{w}^i) + g^*(\mathbf{z}^i)\right] \\ &= \min_{\mathbf{A}\mathbf{x}=\nabla f^*(\mathbf{w}^i)} g(\mathbf{B}\mathbf{x}) + f(\nabla f^*(\mathbf{w}^i)). \end{aligned} \tag{3.17}$$

We now split the proof into two cases according to the construction of $\mathbf{s}^k$. First assume that we use the averaging construction. In this case, we have $\mathbf{s}^k = \frac{1}{\sum_{i=0}^k (i+2\gamma-1)} \sum_{i=0}^k (i+2\gamma-1)\nabla f^*(\mathbf{w}^i)$. We note that the function $\mathbf{s} \mapsto \min_{\mathbf{A}\mathbf{x}=\mathbf{s}} g(\mathbf{B}\mathbf{x}) + f(\mathbf{s})$ is convex and hence, using (3.17),

$$\begin{aligned} \frac{\sum_{i=0}^k \left(S_1(\mathbf{z}^i) - H_1(\mathbf{z}^i)\right)(i+2\gamma-1)}{\sum_{i=0}^k (i+2\gamma-1)} &= \frac{\sum_{i=0}^k \left(\min_{\mathbf{A}\mathbf{x}=\nabla f^*(\mathbf{w}^i)} g(\mathbf{B}\mathbf{x}) + f(\nabla f^*(\mathbf{w}^i))\right)(i+2\gamma-1)}{\sum_{i=0}^k (i+2\gamma-1)} \\ &\geq \min_{\mathbf{A}\mathbf{x}=\mathbf{s}^k} g(\mathbf{B}\mathbf{x}) + f(\mathbf{s}^k) \\ &= f(\mathbf{A}\mathbf{x}^k) + g(\mathbf{B}\mathbf{x}^k). \end{aligned}$$

This concludes the proof for the case where $\mathbf{s}^k$ is given by averaging. Suppose now that $\mathbf{s}^k$ is given by keeping the best iterate, that is $\mathbf{s}^k = \nabla f^*(\mathbf{w}^{k_0})$ where $k_0 \in \operatorname{argmin}_{i=0,1,\ldots k}\left\{S_1(\mathbf{z}^i) - H_1(\mathbf{z}^i)\right\}$. In this case, using again (3.17),

$$\begin{aligned} \frac{\sum_{i=0}^k \left(S_1(\mathbf{z}^i) - H_1(\mathbf{z}^i)\right)(i+2\gamma-1)}{\sum_{i=0}^k (i+2\gamma-1)} &\geq S_1(\mathbf{z}^{k_0}) - H_1(\mathbf{z}^{k_0}) \\ &= \min_{\mathbf{A}\mathbf{x}=\mathbf{s}^k} g(\mathbf{B}\mathbf{x}) + f(\mathbf{s}^k) \\ &= f(\mathbf{A}\mathbf{x}^k) + g(\mathbf{B}\mathbf{x}^k), \end{aligned}$$

and the proof is complete. □

Remark 3.11. *Recall that H_1 is the opposite of the dual objective of problem (3.2). Thus, the left-hand side of (3.14) is the difference between the objective of problem (3.1) evaluated at $\mathbf{x}^k$ and the objective of its dual problem (3.2) evaluated at $(\mathbf{w}^k, \mathbf{z}^k)$, and can thus be considered as a duality gap. In addition, the term $H_1(\mathbf{z}^0) + \bar{p}$ appearing in the right-hand side of (3.14) is the initial dual suboptimality in (3.2).*

Remark 3.12 (Constant refinement)**.** *The constants μ, L_g and matrix norms that appear in (3.15) can be refined in specific instances. Indeed, the proof of Theorem 3.10 requires to consider the smoothness modulus of f^* and the diameter of* $\operatorname{dom}(g^*)$ *only restricted to sets of the form $G_1 + \delta_{Z^k}$, which for some specific choices of Z^k can yield much better constants than μ and L_g (as for example in Section 4.2). It is possible to propose even finer refinements using curvature constants that take into account the geometry of the problem [16].*

4 Applications and numerical illustration

We illustrate the relevance of the primal model (3.1) with two examples. The first one is a toy one-dimensional inpainting problem, for which we would like to recover a piecewise constant signal from partial noisy measurements. In the second example we consider binary classification with offset and binary SVM with offset. For each of the problems we explicitly write the corresponding dual $\frac{1}{\gamma}$-PDA method and show numerical results.

4.1 1D inpainting

4.1.1 Description of the problem

In the 1D inpainting problem, we assume that we are given noisy measurements of a subset of components of a vector $\tilde{\mathbf{x}} \in \mathbb{R}^n$. Specifically, we are given a function $I : \{1, 2, \ldots, p\} \to \{1, 2, \ldots, n\}$ satisfying

$$1 = I(1) < I(2) < \cdots < I(p) = n.$$

Note that we consider that the first and last entries of $\mathbf{x}_0$ are measured. Indeed, we will use the canonical order on coordinates and focus on interpolation, and in particular on entries between the two extreme measurements which we denote by 1 and n. The indices $I(1), I(2), \ldots, I(p)$ are exactly the indices for which the noisy measurements of $\tilde{\mathbf{x}}$ are given:

$$y_j = \tilde{x}_{I(j)} + \epsilon_j, j = 1, 2, \ldots, p, \tag{4.1}$$

where ϵ_j can be viewed as noise or errors. The vector $\mathbf{y} \in \mathbb{R}^p$ is given and we would like to recover $\tilde{\mathbf{x}}$ based on additional prior structure. We will denote the set of known indices by $\mathcal{I} = \{I(1), I(2), \ldots, I(p)\}$. A different way to represent (4.1) is by defining a matrix $\mathbf{A} \in \mathbb{R}^{p \times n}$ by

$$A_{i,j} = \begin{cases} 1, & j = I(i), \\ 0, & \text{else.} \end{cases}, i \in \{1, 2, \ldots, p\}, j \in \{1, 2, \ldots, n\}.$$

Using the matrix $\mathbf{A}$, (4.1) becomes $\mathbf{y} \approx \mathbf{A}\tilde{\mathbf{x}}$. In order to recover the lost measurements, we assume that the original vector $\tilde{\mathbf{x}}$ is piecewise constant. We can use the total variation norm as a structure inducing prior to recover $\tilde{\mathbf{x}}$. We consider the following penalized least-squares problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \lambda \sum_{i=1}^{n-1} |x_i - x_{i+1}|, \tag{4.2}$$

which can be rewritten as

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2 + \lambda\|\mathbf{B}\mathbf{x}\|_1, \tag{4.3}$$

where $\mathbf{B}\in\mathbb{R}^{(n-1)\times n}$, is such that for all $i=1,2,\ldots,n-1$, $B_{i,i}=1$, $B_{i,i+1}=-1$, and all other entries are zeros. Problem (4.3) is of the general form of the main model (3.1) with $f(\cdot)=\frac{1}{2}\|\cdot-\mathbf{y}\|^2, g(\cdot)=\lambda\|\cdot\|_1$. Thus, the dual problem as given in (3.6) takes the form

$$\begin{array}{ll}\min & \frac{1}{2}\|\mathbf{A}\mathbf{B}^T\mathbf{z}+\mathbf{y}\|^2 \\ \text{s.t.} & (\mathbf{I}-\mathbf{A}^T\mathbf{A})\mathbf{B}^T\mathbf{z}=\mathbf{0}, \\ & \|\mathbf{z}\|_\infty\le\lambda,\end{array} \tag{4.4}$$

where here we used the fact that $\mathbf{A}\mathbf{A}^T=\mathbf{I}$ (since the rows of $\mathbf{A}$ are different unit vectors). Furthermore, $\mathbf{A}^T\mathbf{A}$ is a diagonal matrix whose ith diagonal entry is 1 if $i\in\mathcal{I}$ (hence including 1 and n) and 0 otherwise. In addition, $\mathbf{B}^T$ is of size $n\times(n-1)$ with $\mathbf{B}^T_{i,i-1}=-1$ and $\mathbf{B}^T_{i,i}=1$ for $i=2,3,\ldots,(n-1)$. Combining these two facts, we have that the system of equality constraints in (4.4) is equivalent to the system

$$z_i = z_{i-1},\ \forall i\notin\mathcal{I}. \tag{4.5}$$

The specific form of these constraints makes it easy to construct a basis for the null space. We assume that all elements of this basis are given by the columns of a matrix $\mathbf{U}$ and perform the change of variables $\mathbf{z}=\mathbf{U}\tilde{\mathbf{z}}$. The matrix $\mathbf{U}$ can be chosen to be of the following form

$$\mathbf{U}=\begin{pmatrix} 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&1&0&0&0\\ 0&0&1&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1 \end{pmatrix},$$

where columns that contain several ones account for constraints of the form of (4.5) for several consecutive indices not in I. The special example given here corresponds to $\mathcal{I}=\{1,2,3,7,8,9,10\}$ and $n=10$. The construction of $\mathbf{U}$ ensures that its columns form a basis of the null space of $(\mathbf{I}-\mathbf{A}^T\mathbf{A})\mathbf{B}^T$ and that for any $\tilde{\mathbf{z}}$, $\|\mathbf{U}\tilde{\mathbf{z}}\|_\infty=\|\tilde{\mathbf{z}}\|_\infty$. Therefore, problem (4.4) is equivalent to the problem

$$\begin{array}{ll}\min\limits_{\tilde{\mathbf{z}}} & \frac{1}{2}\left\|\mathbf{A}\mathbf{B}^T\mathbf{U}\tilde{\mathbf{z}}+\mathbf{y}\right\|^2 \\ \text{s.t.} & \|\tilde{\mathbf{z}}\|_\infty\le\lambda,\end{array} \tag{4.6}$$

which is a box constrained problem that can be solved by various methods such as the conditional gradient method or the proximal gradient method (which are 1-PDA methods) or one of the variants of greedy coordinate descent method as explained in Section 2.2.4. We focus on methods which yields primal convergence rates as described in Section 3.

4.1.2 Numerical simulation

We compare the conditional gradient, greedy block conditional gradient and projected gradient on problem (4.6). All these methods can be viewed as $\frac{1}{\gamma}$-PDA methods. The numerical criterion of

interest is the duality gap. Theorem 3.10 provides a primal dual convergence rate estimate of $O(1/k)$ for the three methods. The analysis allows to reconstruct sequences of estimates for the primal problem (4.2). We simulate a piecewise constant signal, remove some of its entries and add gaussian noise. The simulation setting is illustrated in Figure 1. The comparative primal-dual convergence is given in Figure 2. The sparse version of the conditional gradient method performs significantly better than the traditional conditional gradient and slightly better than the traditional projected gradient method. Furthermore, the averaging rule to reconstruct the primal sequence seems to help a bit for the traditional conditional gradient while it tends to degrade performances for both greedy block conditional gradient and projected gradient.

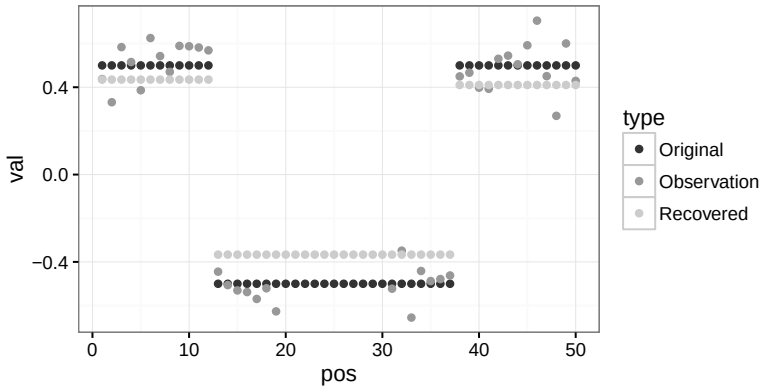

Figure 1: Numerical simulation setting. The original one dimensional signal is in red. The observations consist in removing some part of the signal and adding noise (in green). The signal recovered by the greedy block conditional gradient method is given in blue. Note that the fact that the gaps in the recovered signal are smaller than in the original signal is an unavoidable effect of total variation regularization.

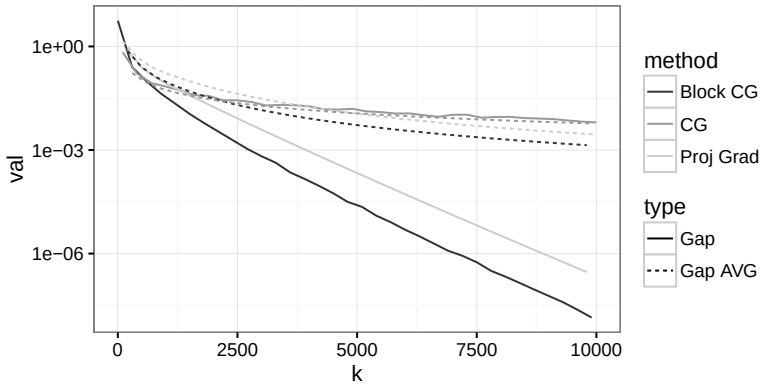

Figure 2: Comparative performances of the full conditional gradient, its greedy block version and the projected gradient algorithm on the simplified dual problem (4.6). The comparison is in terms of duality gap with (Gap AVG) and without (Gap) using the averaging rule for primal reconstruction.

4.2 Binary classification with offset

4.2.1 Setting

Structural risk minimization is the process of estimating a decision function by minimizing a risk term evaluated on an empirical dataset with a capacity control term [34]. We will focus on binary classification with affine predictors. We have q datapoints, that is, for each $i = 1, 2, \ldots, q$, we have a vector of features $\mathbf{s}_i \in \mathbb{R}^n$ and a binary output $t_i \in \{-1, 1\}$. We are looking for a decision boundary given by a pair $(\mathbf{x}, b) \in \mathbb{R}^n \times \mathbb{R}$ of the form $\{\mathbf{a} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{a} \rangle = b\}$. This is done by minimizing a penalized empirical risk.

$$\min_{\mathbf{x},b} \frac{1}{2}\|\mathbf{x}\|^2 + \frac{C}{q}\sum_{i=1}^{q} l(t_i(\langle \mathbf{x}, \mathbf{s}_i \rangle - b)), \tag{4.7}$$

where $C > 0$ is a given regularization parameter and l is a convex Lipschitz continuous and nonincreasing loss function from $\mathbb{R}$ to $\mathbb{R}$ with a nonzero derivative at the origin. Let $\mathbf{S}$ be the matrix whose columns are the vectors $t_i\mathbf{s}_i$, $i = 1, 2, \ldots, q$, and $\mathbf{t}$ be the vector whose entries are t_i, $i = 1, 2, \ldots, q$. It is clear that problem (4.7) fits model (3.1) with $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^q \to \mathbb{R}$ defined by

$$f(\mathbf{w}) \equiv \frac{1}{2}\|\mathbf{w}\|^2, g(\mathbf{u}) \equiv \frac{C}{q}\sum_{i=1}^{q} l(u_i),$$

and $\mathbf{A} \in \mathbb{R}^{n\times(n+1)}, \mathbf{B} \in \mathbb{R}^{q\times(n+1)}$ given by

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1} \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{S}^T & -\mathbf{t} \end{pmatrix}, \tag{4.8}$$

where $\mathbf{0}_{a\times b}$ is the $a \times b$ zeros matrix for $a, b \in \mathbb{N}$. To write explicitly the dual problem (3.6), we will first compute the matrix $\mathbf{P}$

$$\mathbf{P} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & 0 \end{pmatrix}.$$

Therefore, we have the following equivalence:

$$(\mathbf{I} - \mathbf{P})\mathbf{B}^T\mathbf{z} = \mathbf{0} \text{ iff } \mathbf{t}^T\mathbf{z} = 0. \tag{4.9}$$

Also, since $\mathbf{A}\mathbf{A}^T = \mathbf{I}$, we have

$$-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z} = -\mathbf{A}\mathbf{B}^T\mathbf{z} = -\begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1} \end{pmatrix}\begin{pmatrix} \mathbf{S} \\ -\mathbf{t}^T \end{pmatrix}\mathbf{z} = -\mathbf{S}\mathbf{z}. \tag{4.10}$$

The conjugates of f and g are

$$f^*(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|^2, g^*(\mathbf{r}) = \frac{C}{q}\sum_{i=1}^{q} l^*\left(\frac{qr_i}{C}\right). \tag{4.11}$$

Therefore, plugging (4.9), (4.10) and (4.11) into the general form of the dual problem (3.6), we obtain that a dual of problem (4.7) in minimization form is

$$\begin{array}{ll} \min & \frac{1}{2}\mathbf{z}^T\mathbf{S}^T\mathbf{S}\mathbf{z} + \frac{C}{q}\sum_{i=1}^{q} l^*\left(\frac{qz_i}{C}\right) \\ \text{s.t.} & \mathbf{t}^T\mathbf{z} = 0. \end{array} \tag{4.12}$$

4.2.2 Support vector machine and SMO type algorithms

We will be particularly interested in the case of the SVM [10] for which l is the hinge loss, meaning that $l\colon z \to \max\{1-z, 0\}$. In this case, $l^*(z) \equiv z + \iota_{-1\leq z\leq 0}$, and thus (4.12) can be written as

$$\begin{array}{ll} \min & \frac{1}{2}\mathbf{z}^T\mathbf{S}^T\mathbf{S}\mathbf{z} + \mathbf{1}^T\mathbf{z}, \\ \text{s.t.} & -\frac{C}{q} \leq z_i \leq 0, i = 1, 2, \ldots, q \\ & \mathbf{t}^T\mathbf{z} = 0, \end{array} \tag{4.13}$$

which corresponds to the usual expression the dual SVM problem (after making the change of variables $\mathbf{z} \leftarrow -\mathbf{z}$) that consists of minimizing smooth objective function over a box with one additional linear equality constraint. Active set methods rely on updates of pairs of variables in the dual [27, 28, 17]. These algorithms are among the most popular for SVM training with offset because the updates are very cheap and allow to consider large numbers of training points [8]. Since the method we propose falls in this categorty of approaches, we restrict ourselves to this class of methods in the theoretical discussion (Section 4.2.3) and numerical experiments (Section 4.2.4).

Since there is a single linear constraint in the dual, we can use the construction of Section 2.2.5 to build such a working set method that updates only pairs of dual variables at each iteration. The interesting additional property here is that this constitutes a $\frac{1}{q}$-PDA method and our theory applies. Solving the linear oracle for the SVM can be viewed as a fractional knapsack problem. A naive solution requires to sort a q dimensional vector and perform an exhaustive search (linear in q). This problem can also be solved in $O(q)$ operations with a weighted medians algorithm [19, Section 17.1].

To compute the primal sequence $\{(\mathbf{x}^k, b_k)\}_{k\geq 0}$ from the dual sequence $\{\mathbf{z}^k\}_{k\geq 0}$, we use the formula (3.10). We will consider the following two possibilities:

$$\begin{array}{rcl} \mathbf{s}^k & = & \begin{cases} \text{(averaging)} & -\frac{1}{\sum_{i=0}^k (i+2q-1)} \sum_{i=0}^k (i+2q-1)\mathbf{S}\mathbf{z}^i, \\ \text{(last iterate)} & -\mathbf{S}\mathbf{z}^k, \end{cases} \\ \mathbf{x}^k & = & \mathbf{s}^k, \\ b^k & \in & \operatorname{argmin}_{b\in\mathbb{R}} \sum_{i=1}^q l(t_i(\langle s_i^k, \mathbf{s}_i\rangle - b)). \end{array}$$

Note that, as outlined in Remark 3.7, the averaged sequence can be computed online without storing the whole sequence of iterates. We can invoke Theorem 3.10 and obtain $O(1/k)$ rates of convergence for the averaging rule. For the other rule, we have the same rate for the best primal point estimated so far. Note that we do not have primal convergence guaranties concerning the last iterate. We still consider it here in order to investigate the effect of averaging.

4.2.3 Implications of Theorem 3.10 for binary SVM and relation to the literature

Theorem 3.10 ensures that after k iterations of any dual $\frac{1}{q}$-PDA method, we find a set of primal variables $(\mathbf{x}^k, b^k)$ and dual variables $\mathbf{z}^k$ such that the difference between the primal objective evaluated at $(\mathbf{x}^k, b^k)$ with the dual objective evaluated at $\mathbf{z}^{k+1}$ is of order $O(1/k)$. In particular, $(\mathbf{x}^k, b^k)$ achieves a training accuracy of order $O(1/k)$. We emphasize that this training accuracy (or primal suboptimality) is very relevant from a machine learning perspective. For the sake of clariy, we restrict the discussion to active set methods which are widely used for SVM training [8]. In this context, the closest results we could find in the literature are [15, Theorem 2] and [24, Corollary 3] which we comment below.

- [15, Theorem 2] ensures that if $\mathbf{z}^k$ is ϵ-suboptimal for problem (4.13), then the corresponding primal variable as given by the last iterate rule is $\sqrt{\epsilon}$-suboptimal for problem (4.12). They show in addition that $\mathbf{z}^k$ has dual suboptimality of order $O(1/k)$ resulting in a $O(1/\sqrt{k})$ convergence

rate for the primal variable in (4.7). Theorem 3.10 improves this result by showing a rate of $O(1/k)$ for both (4.7) and (4.13).

- [24, Corollary 3] ensures that if the duality gap is less than ϵ, then, the primal variables corresponding to the last iterate are ϵ-optimal for the primal problem (4.12). This is relevant since this quantity is often used as a stoping criterion. However, this analysis is somehow incomplete since there is no explicit condition on k ensuring that the duality gap is less than ϵ. Therefore the result cannot be directly translated in a convergence rate for the primal variable sequence. Theorem 3.10 is stronger because it gives an explicit global rate of $O(1/k)$ for both (4.7) and (4.13).

The approximation factor of the method is $\frac{1}{q}$ where q is the number of datapoints. This translates into a multiplicative constant of order of q^2 in the rate of Theorem 3.10 which is a bit disapoiting for large datasets. Note however that the squared diameter of the feasible domain in problem (4.13) is not more than $\frac{C^2}{q}$. Reading (3.14) with this in mind, we obtain a bound of the form

$$\frac{2}{2+\frac{k}{q}}\left(\frac{2d_0}{\frac{k}{q}}+C^2K_1\right),$$

where d_0 is the dual suboptimality at iteration 0 and K_1 only depends on the singular values of $\mathbf{B}$. This shows that the global suboptimality is roughly inversely proportional to the ratio $\frac{k}{q}$ which is quite reasonable. Furthermore, as explained in Remark 3.12, it is possible to further refine the constants appearing in (3.14), a process that will require further discussions on finding tighter estimates on the scaling of C and singular values of $\mathbf{S}$ with increasing values of q. These specific considerations are beyond the scope of this paper.

4.2.4 Numerical simulations

We consider training a linear SVM as given by (4.7) on a randomly generated dataset. The setting is as follows

- The ambiant dimension is $p = 20$.
- We consider two classes sampled from unit Gaussian random variables with a shift in mean of magnitude 2 (in euclidean norm).
- We vary the number of datapoints $q \in \{100, 200\}$, evenly spread in the two classes.
- We vary the regularization parameter $C \in \{10, 100, 1000\}$.

The main purpose is to illustrate the behaviour of the PDA framework in view of primal and dual suboptimality. We will use a coordinate selection rule that we call WSS1, implemented in LIBSVM [8], one of the most widely used SVM solvers, as a baseline. The only difference with PDA is the coordinate selection rule, the rest of the algorithm being the same. Both selection rules are combined with an exact block minimization step, which is a simple two-dimensional problem here. Note that in this case, the cost of computing the PDA point of Section 2.2.5 is linear in q, the number of training examples. Furthermore, Theorem 3.10 provides convergence guaranties for both problems (4.7) and (4.13). For the WSS1 rule, the number of operations required is of the order of q. Numerical results in term of evolution of the primal and the dual objective values as a function of the iteration counter k are presented in Figure 3, which illustrate convergence of primal (with and without averaging) and dual objective to the global optimum value. The main comment is that the working set selection of the $\frac{1}{q}$-PDA rule is competitive and at times superior to the WSS1

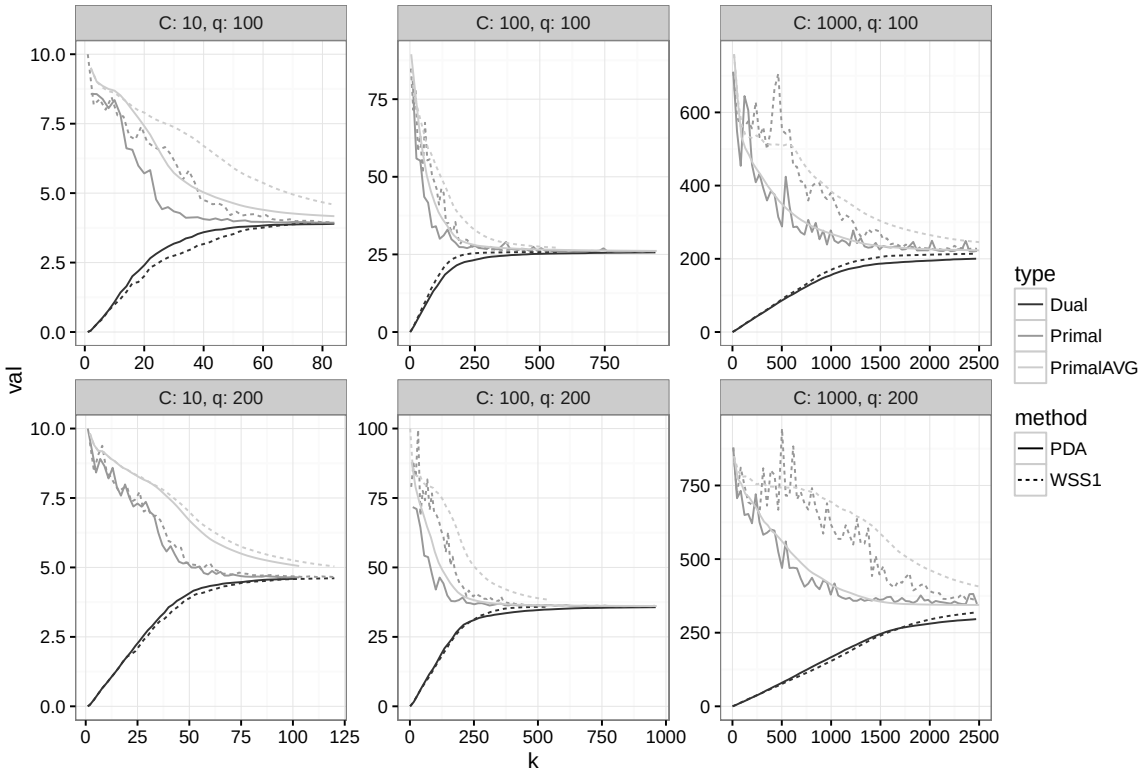

Figure 3: Evolution of the primal and dual objectives with the number of iterations for the training of the SVM on the toy dataset (see Section 4.2.4). PDA implements the update rule described in section 2.2.5 and WSS1 is the coordinate selection rule of LIBSVM [8]. q is the total number of datapoints and C is the regularization parameter of the SVM. For both coordinate selection rules, we plot the value of the dual objective, the primal objective and the primal objective with the averaging rule.

update rule in terms of primal suboptimality on this specific problem. Another important comment is that the absolute performances depend on the parameters of the problem. The averaging rule for the primal sequence reconstruction does not seem to bring a systematic practical advantage beyond smoother primal convergence. An interesting remark is that the WSS1 rule provides in some cases a better dual convergence while the primal convergence is worse compared to our PDA method. This highlights the idea that better convergence in terms of dual objective function does not necessarily translates into faster convergence in the primal.

5 Conclusion

This work builds upon the idea of predicted decrease approximation to provide a unified convergence analysis for various existing decomposition algorithms for constrained convex optimization. We have shown that a single result allows to treat as special cases the generalized conditional gradient method, the proximal gradient method, and greedy coordinate descent method and working set method for smooth problems with linear equality as well as bound constraints. Furthermore, we have shown that the dual application of this approach leads to primal-dual convergence guaranties that hold even if the primal model is only partially strongly convex. This lead to better convergence analysis of SMO-type methods for the training of the SVM in terms of primal sequence suboptimality. To conclude, we comment on the following aspects of the proposed analysis which relates our work to broader considerations in optimization.

- The overall algorithmic recipe leads to block decomposition methods for models involving non-separable constraints. The price to pay is the requirement to consider larger blocks (larger subsets of coordinates), but the convergence remains. Many questions are open in this respect. Could this benefit to parallel computing architectures and distributed data settings? What would be the practical and theoretical impact of introducing randomness in the block selection process? Can we extend these results to more general non-separable settings.

- A general rule of thumb for nonsmooth convex optimization is that the optimal convergence speed of subgradient methods is $O(1/\sqrt{k})$ for convex models and $O(1/k)$ for strongly convex models. The algorithmic framework we proposed takes advantage of partial strong convexity to retain the convergence speed of strongly convex models while being only partially strongly convex.

- The main mechanism in the proposed primal-dual analysis is to build a primal estimate based on the knowledge of a dual feasible point. A property of the proposed approach is that both primal and dual sequences are feasible. This is a difference in comparison to Lagrangian based methods for which feasibility usually holds in an asymptotic and ergodic sense. In our work, going from the dual to the primal requires an additional optimization step in order to ensure primal dual convergence. This occurs because there is a certain level of undetermination in the process of going back to the primal which requires special care. The level of undetermination can be interpreted to be the same as the level of "non-strong convexity" in the primal model. This draws an interesting connection between partial strong convexity in the primal and easiness of switching from the dual to the primal.

References

[1] F. Bach. Duality between subgradient and conditional gradient methods. *SIAM J. Optim.*, 25(1):115–129, 2015.

[2] A. Beck. The 2-coordinate descent method for solving double-sided simplex constrained minimization problems. *J. Optim. Theory Appl.*, 162(3):892–919, 2014.

[3] A. Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemess. *SIAM J. Optim.*, 25(1):185–209, 2015.

[4] A. Beck, E. Pauwels, and S. Sabach. The cyclic block conditional gradient methods for convex optimization problems. to appear in SIAM J. Optim.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[6] M. D. Canon and C. D. Cullum. A tight upper bound on the rate of convergence of frank-wolfe algorithm. *SIAM J. Control*, 6(4):509–516, 1968.

[7] C.-C. Chang, C.-W. Hsu, and C.-J. Lin. The analysis of decomposition methods for support vector machines. *Neural Networks, IEEE Transactions on*, 11(4):1003–1008, 2000.

[8] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[9] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[11] V. F. Dem'yanov and A. M. Rubinov. The minimization of a smooth convex functional on a convex set. *SIAM J. Control*, 5(2):280–294, 1967.

[12] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[13] J. C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *J. Math. Anal. Appl.*, 62(2):432–444, 1978.

[14] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, 3(1-2):95–110, 1956.

[15] D. Hush, P. Kelly, C. Scovel, and I. Steinwart. QP algorithms with guaranteed accuracy and run time for support vector machines. *The Journal of Machine Learning Research*, 7:733–769, 2006.

[16] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 427–435, 2013.

[17] T. Joachims. Making large-scale support vector machine learning practical. In *Advances in kernel methods*, pages 169–184. MIT Press, 1999.

[18] H. Karloff. *Linear programming*. Progress in Theoretical Computer Science. Birkhäuser Boston, Inc., Boston, MA, 1991.

[19] B. Korte and J. Vygen. *Combinatorial optimization*. Springer, 2002.

[20] S. Lacoste-Julien and M. Jaggi. An affine invariant linear convergence analysis for Frank-Wolfe algorithms. *NIPS 2013 Workshop on Greedy Algorithms, Frank-Wolfe and Friends*, 2014.

[21] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate frank-wolfe optimization for structural svms. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 53–61, 2013.

[22] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

[23] E. S. Levitin and B. T. Poljak. Minimization methods in the presence of constraints. *USSR Comput. Math. and Math. Phys.*, 6(5):787–823, 1966.

[24] N. List, D. Hush, C. Scovel, and I. Steinwart. Gaps in support vector optimization. In *Learning Theory*, pages 336–348. Springer, 2007.

[25] N. List and H. U. Simon. General polynomial time decomposition algorithms. *The Journal of Machine Learning Research*, 8:303–321, 2007.

[26] J. J. Moreau. Proximitéet dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[27] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 276–285. IEEE, 1997.

[28] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods-support vector learning*, 3, 1999.

[29] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[30] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[31] H. U. Simon. On the complexity of working set selection. *Theoretical Computer Science*, 382(3):262–279, 2007.

[32] G. Strang. *Linear algebra and its applications*. Academic Press, New York-London, second edition, 1980.

[33] P. Tseng and S. Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Computational Optimization and Applications*, 47(2):179–206, 2010.

[34] V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, 1995.

[35] J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May, 1998.