

A Novel Frank-Wolfe Algorithm. Analysis and Applications to Large-Scale SVM Training

Héctor Allende¹, Emanuele Frandi², Ricardo Ñanculef¹, and Claudio Sartori³

¹*Department of Informatics, Universidad Técnica Federico Santa María, Chile*
{hallende,jnancu}@inf.utfsm.cl

²*Department of Science and High Technology, University of Insubria, Italy*
emanuele.frandi@uninsubria.it

³*Department of Computer Science and Engineering, University of Bologna, Italy,*
claudio.sartori@unibo.it

Abstract

Recently, there has been a renewed interest in the machine learning community for variants of a sparse greedy approximation procedure for concave optimization known as the Frank-Wolfe (FW) method. In particular, this procedure has been successfully applied to train large-scale instances of non-linear Support Vector Machines (SVMs). Specializing FW to SVM training has allowed to obtain efficient algorithms but also important theoretical results, including convergence analysis of training algorithms and new characterizations of model sparsity.

In this paper, we present and analyze a novel variant of the FW method based on a new way to perform away steps, a classic strategy used to accelerate the convergence of the basic FW procedure. Our formulation and analysis is focused on a general concave maximization problem on the simplex. However, the specialization of our algorithm to quadratic forms is strongly related to some classic methods in computational geometry, namely the Gilbert and MDM algorithms.

On the theoretical side, we demonstrate that the method matches the guarantees in terms of convergence rate and number of iterations obtained by using classic away steps. In particular, the method enjoys a linear rate of convergence, a result that has been recently proved for MDM on quadratic forms.

On the practical side, we provide experiments on several classification datasets, and evaluate the results using statistical tests. Experiments show that our method is faster than the FW method with classic away steps, and works well even in the cases in which classic away steps slow down the algorithm. Furthermore, these improvements are obtained without sacrificing the predictive accuracy of the obtained SVM model.

1 Introduction

In this paper we present a novel variant of the Frank-Wolfe (hereafter FW) method [23, 33], designed to deal with large-scale instances of the following problem:

$$\underset{\alpha}{\text{maximize}} g(\alpha) \quad \text{subject to} \quad \alpha \in S := \left\{ \alpha \in \mathbb{R}^m : \sum_i \alpha_i = 1, \alpha_i \geq 0 \right\}. \quad (1)$$

This problem encompasses several models used in machine learning [13, 28], including hard-margin Support Vector Machines (SVMs) [36] and L_2 -loss SVMs (L_2 -SVMs) for binary classification, regression and novelty detection [58, 59].

FW Methods and Focus of this Paper It has been noted by researchers in different fields that approximate solutions to problem (1) can be obtained using quite simple iterative procedures. In [64], for instance, Yildirim presents two iterative algorithms for the task of approximating the Minimum Enclosing Ball (MEB) of a set of points. In [1], Ahipasaoglu *et al.* propose similar methods to solve Minimum Volume Enclosing Ellipsoid problems. In [66], Zhang studies similar techniques for convex approximation and estimation of mixture models. All these methods are nowadays identified as variants of a general approximation procedure for maximizing a differentiable concave function on the simplex, which traces back to Frank and Wolfe [23, 62, 27] and has been recently analyzed by Clarkson [13] and Jaggi [33] under a modern perspective.

In a nutshell, each iteration of the FW method moves the solution towards the direction along which the linearized objective function increases most rapidly but is still feasible. The procedure is related to the idea of *coreset*, coined in the context of computational geometry and denoting a subset of data C_ε which suffices to obtain an approximation to the solution on the whole dataset up to a given precision ε . Clarkson’s framework unifies diverse results regarding the existence of small coresets for different instances of problem (1). These ideas were used in [28] to characterize the sparsity of SVMs and the convergence properties of training algorithms for geometric formulations of the problem.

The algorithm studied in this paper is obtained by incorporating a new type of *away step* into the basic FW method. Loosely speaking, instead of moving the solution towards a direction along which the linearized objective function increases, an away step moves the solution away from a direction along which the linearized objective function decreases. This strategy was suggested by Wolfe in [62] to improve the convergence rate of the FW method, leading to a variant of the original algorithm called Modified Frank-Wolfe method (hereafter MFW). It has been demonstrated that MFW is linearly convergent under some general assumptions on the properties of problem (1). However, we have found in [20] that classic away steps do not improve significantly the running times of the FW method on machine learning problems. A similar conclusion was obtained by Ouyang and Gray in [44]. In contrast, our approach experimentally improves on other FW methods and shows theoretical guarantees (e.g. convergence rate) at least as good as those of MFW.

Applications to SVM Learning Training non-linear SVMs on large datasets is challenging [16]. Effective Interior Point Methods can be devised under some special circumstances, such as kernels which admit low-rank factorizations [18, 63]. However, these methods are not suitable for large-scale problems in a general scenario, mainly due to memory constraints: a general interior point method needs $\mathcal{O}(m^2)$ memory and $\mathcal{O}(m^3)$ time for matrix inversions, and both are prohibitive even for medium-scale problems. Among the traditional methods devised to cope with this problem, Active Set methods [49, 34, 50] and Sequential Minimal Optimization (SMO) [45, 16] are well-known alternatives among practitioners. Indeed, these are the algorithms of choice in the widely known libraries SVMlight [35] and LIBSVM [12], respectively. For the linear kernel case, Stochastic Gradient Descent (SGD) [7, 8], specialized sub-gradient methods like Pegasos [54] and Stochastic Dual Coordinate Ascent (SDCA) [31, 55] have lately gained popularity in the community as approximate but efficient alternatives to the classic solutions on large-scale problems [65].

In the non-linear case, effective methods to deal with large datasets have been recently devised by focusing on formulations which fit problem (1) and then applying FW methods. The first work to specialize a variant of the FW method to SVM training is probably due to Tsang *et al.* [58]. Given a labelled set of examples $\{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathcal{X}, y_i \in \{+1, -1\}, i \in \mathcal{I}\}$, where \mathcal{X} denotes the input space and $\mathcal{I} = \{1, \dots, m\}$ an index set, they adopt the so-called L_2 -SVM formulation, where the model is built by solving the following optimization problem

$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad g(\boldsymbol{\alpha}) = -\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \quad \text{subject to} \quad \sum_i \alpha_i = 1, \alpha_i \geq 0, \quad (2)$$

where $\mathbf{K}_{i,j} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) + y_i y_j + \delta_{i,j}/C$, $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is the kernel function used in the SVM model and C is the regularization parameter [58, 19, 20]. Problem (2) clearly fits problem (1). This formulation is adopted mainly because of efficiency: by using the functional of Eqn. (2), it is possible to exploit the framework introduced in [58], and further developed in [13], to solve the learning problem more easily¹. Note also that in problem (2) \mathbf{K} is positive definite² for $0 < C < \infty$ and thus $g(\cdot)$ is strictly concave.

Borrowing a coresets-based algorithm from computational geometry [10], the authors obtain that the total number of iterations needed to identify a coresets, i.e. an approximation to the L_2 -SVM model up to an arbitrary precision ϵ , is bounded by $\mathcal{O}(1/\epsilon)$, independently of the size of the dataset. From the iterative structure of the algorithm, it follows easily that the size of the coresets is also bounded by $\mathcal{O}(1/\epsilon)$. A similar result regarding linear SVMs trained with SDCA has been recently demonstrated in [55].

¹Strictly speaking, [58] is a special case of a FW method which does not address the general form of problem (2), as a normalization constraint on the quadratic form is required (see Sections 2.2 and 2.4).

²This is easily seen by writing \mathbf{K} as the sum of two positive semi-definite matrices and a multiple of the identity, $\mathbf{K} = \mathbf{y}\mathbf{y}^T \odot \tilde{\mathbf{K}} + \mathbf{y}\mathbf{y}^T + \frac{1}{C}\mathbf{I}$, where \mathbf{y} is the column vector whose components are the labels y_i , $\tilde{\mathbf{K}}$ is the Gram matrix $\tilde{\mathbf{K}}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ and \odot is the Hadamard or componentwise product.

The latter properties imply in particular that the size of the set of examples required to represent the (approximate) solution, i.e. the number of support vectors in the model, is also independent from the size of the dataset, an improvement on previous lower bounds for the support set size, such as those in [56], where the bound grows linearly in the size of the dataset. The obtained training algorithm also exhibits linear running times in the number of examples. These are remarkable results in the context of non-linear SVM models, where the support set needs to be explicitly stored in memory to implement predictions and determines the cost of a classification decision in terms of testing time. In addition, a combination of this procedure with certain sampling techniques allows to obtain sub-linear time approximation algorithms [58, 19]³. In practice, the method was found to be competitive with most well-known SVM software using non-linear kernels [58, 59].

Several other papers have recently stressed the efficiency of FW and coresets-based methods in machine learning. In [19] and [21] the authors investigate the direct application of the FW method to large-scale non-linear SVM training, demonstrating that running times of [58] can be significantly improved as long a minor loss in accuracy is acceptable. Variations of the algorithm based on geometrical reformulations of the learning problem [38, 28], stochastic variants of the method [44], and applications to SVM training on data streams [61, 47] and structural SVMs [39] have also been proposed.

Contributions We present a FW method endowed with a new type of optimization step devised to overcome the difficulties observed with the classic MFW approach, while preserving the intuition and benefits behind the introduction of away steps. On the theoretical side, we formulate and analyze the algorithm for the general case of problem (1), demonstrating that the method matches the guarantees in terms of convergence rate and number of iterations obtained by using classic away steps. In particular, we show that the method converges linearly to the optimal value of the objective function, and achieves a predetermined accuracy ε (primal-dual gap) in at most $\mathcal{O}(1/\varepsilon)$ iterations. Focusing on quadratic objectives, it turns out that the method is strongly related to the Gilbert and Mitchell-Demnyanov-Malozemov (MDM) algorithms, two classic methods in computational geometry [26, 42]. Such methods are well-known in machine learning and their properties, in particular their rate of convergence, have been the focus of recent research [40, 41].

On the practical side, we specialize the algorithm to SVM training and perform detailed experiments on several classification problems. We conclude that our algorithm improves the running times of existing FW approaches without any statistically significant difference in terms of prediction accuracy. In particular, we show that the method is faster than the FW and MFW methods, while MFW is not statistically faster than FW. In addition, we show that the

³To be rigorous, the probability of identifying a good point in a given iteration depends on the size of the sampling. A sub-linear procedure guaranteeing a constant success probability is studied in [14], though it seems that results on the non-linear case are provided only for some kernels.

method is faster than or equal to the FW method when MFW is significantly slower, i.e. when classic away steps fail. In addition, the method is competitive with MFW when FW is significantly slower, i.e., if classic away steps work, our algorithm works as well. Thus, the method represents a robust alternative to implement away steps, enjoying strong theoretical guarantees and providing significant improvements in practice.

Organization The paper is organized as follows. In Section 2 we give an overview of FW methods and introduce the basic concepts required for their analysis. In Section 3 we present the new method, including a minor variant, and provide some details about its specialization to SVMs. The analysis of convergence is provided in Section 4. In Section 5, we discuss the relation of the proposed method to some classic geometric approaches for a quadratic objective. Experiments on SVM problems are presented in Section 6. Finally, Section 7 closes the paper with some concluding remarks. In addition, some technical results required for the proofs of Section 4 are reported in the Appendix.

Notation An optimal solution for problem (1) is denoted α^* . A sequence of approximations $\alpha_0, \alpha_1, \dots, \alpha_k$ to a solution of problem (1) is abbreviated $\{\alpha_k\}_k$. The set of indices $1, 2, \dots, m$ is denoted $[m]$. The *face* $S_{\mathcal{I}}$ of the unit simplex S corresponding to a set of indices $\mathcal{I} \subset [m]$ is the subset of points $\alpha \in S$ such that $\alpha_j = 0 \forall j \notin \mathcal{I}$. The term *active face* indicates the face corresponding to the non-zero indices, \mathcal{I}_k , of the current solution α_k . The term *optimal face*, denoted by S^* , indicates the face corresponding to an optimal solution α^* . The vector \mathbf{e}_i denotes the i -th vector of the canonical basis.

2 Frank-Wolfe Methods

The FW method computes a sequence of approximations $\{\alpha_k\}_k$ to a solution of problem (1) by iterating until convergence the following steps. First, a linear approximation of $g(\cdot)$ at the current iterate α_k is performed in order to find an ascent direction $\mathbf{d}_k^{\text{FW}} = (\mathbf{u}_k - \alpha_k)$, with

$$\mathbf{u}_k \in \operatorname{argmax}_{\mathbf{u} \in S} \psi_k(\mathbf{u}) := g(\alpha_k) + (\mathbf{u} - \alpha_k)^T \nabla g(\alpha_k). \quad (3)$$

Since \mathbf{u}_k lies in S , it is easy to see that the linear approximation step reduces to $\mathbf{u}_k = \mathbf{e}_{i^*}$ where i^* is the largest coordinate of the gradient, i.e. $i^* \in \operatorname{argmax}_i \nabla g(\alpha_k)_i$. The iterate α_k is then moved towards \mathbf{e}_{i^*} , seeking for the best feasible improvement of the objective function. The procedure is summarized in Algorithm 1. In the rest of this paper we refer to $\mathbf{e}_{i^*} \in S$ as the *ascent vertex* used by the method.

As discussed below, the procedure can be stopped when $g(\alpha_k)$ is “close enough” to the optimum.

Algorithm 1: FW method for problem (1).

- 1 Compute an initial estimate α_0 .
 - 2 Set $\mathcal{I}_0 = \{i : \alpha_{0,i} \neq 0\}$.
 - 3 **for** $k = 0, 1, \dots$ **do**
 - 4 Search for $i^* \in \operatorname{argmax}_i \nabla g(\alpha_k)_i$ and define $\mathbf{d}_k^{FW} = \mathbf{e}_{i^*} - \alpha_k$.
 - 5 Perform a line-search to find $\lambda_k \in \operatorname{argmax}_{\lambda \in [0,1]} g(\alpha_k + \lambda \mathbf{d}_k^{FW})$.
 - 6 Update the iterate by $\alpha_{k+1} = \alpha_k + \lambda_k \mathbf{d}_k^{FW} = (1 - \lambda_k) \alpha_k + \lambda_k \mathbf{e}_{i^*}$.
 - 7 Set $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{i^*\}$.
-

2.1 Optimality Measures and Stopping Condition

It can be shown that the FW method is globally convergent under rather weak assumptions on the properties of the objective function [27, 23], which are guaranteed to hold for the SVM problem (2) [64, 20]. In addition, it can be shown that the iterates of this procedure satisfy

$$\Delta^p(\alpha_k) := g(\alpha^*) - g(\alpha_k) \leq \frac{4C_g}{k+3}, \quad (4)$$

where C_g is a constant related to the second derivative of g [13]. This convergence rate is slow compared to other methods. However, the simplicity of the procedure implies that the amount of computation per iteration is usually very small. This kind of tradeoff can be favorable for large-scale applications, as testified for example by the widespread adoption of the SMO method in the context of SVMs [45, 16].

When $g(\alpha)$ is continuously differentiable, the Wolfe dual of problem (1) is

$$\underset{\alpha}{\text{minimize}} w(\alpha), \text{ with } w(\alpha) = g(\alpha) + \max_i \nabla g(\alpha)_i - \alpha^T \nabla g(\alpha). \quad (5)$$

As shown in [13], the strong duality condition

$$g(\alpha) \leq g(\alpha^*) = w(\alpha^*) \leq w(\alpha) \quad (6)$$

holds for any feasible α . Thus, another reasonable measure of optimality for the Frank-Wolfe iterates is the so-called *primal-dual gap*

$$\Delta^d(\alpha) := w(\alpha) - g(\alpha) = \max_i \nabla g(\alpha)_i - \alpha^T \nabla g(\alpha). \quad (7)$$

Up to a multiplicative constant ($4C_g$), the primal-dual gap in Eqn. (7) and the primal measure of approximation in Eqn. (4) are the metrics employed in [13] to analyze the convergence of Algorithm 1. The advantage of $\Delta^d(\alpha_k)$ with respect to $\Delta^p(\alpha_k)$ is that the former does not depend on the optimal value of the objective function. Therefore, $\Delta^d(\alpha_k)$ can be explicitly monitored during the execution of the algorithm and can be adopted to implement a stopping

condition for Algorithm 1. In this paper, we adopt this measure to stop the FW method and any of its variants. That is, the algorithms are terminated when

$$\Delta^d(\boldsymbol{\alpha}_k) = \max_i \nabla g(\boldsymbol{\alpha}_k)_i - \boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) \leq \varepsilon, \quad (8)$$

where $\varepsilon > 0$ is a given tolerance parameter. Note that the strong duality condition implies $\Delta^p(\boldsymbol{\alpha}_k) \leq \Delta^d(\boldsymbol{\alpha}_k)$. Therefore, if the algorithm stops at iteration k we also have $\Delta^p(\boldsymbol{\alpha}_k) \leq \varepsilon$.

Note also that Eqn. (4) implies that the FW method finds a solution fulfilling $\Delta^p(\boldsymbol{\alpha}_k) \leq \varepsilon$ in at most $K \sim \mathcal{O}(1/\varepsilon)$ iterations. Clarkson has recently shown that we also have $\Delta^d(\boldsymbol{\alpha}) \leq \varepsilon$ after at most $\tilde{K} \sim \mathcal{O}(1/\varepsilon)$ iterates [13]. Thus, the solution found by the FW method using the stopping condition (8) is guaranteed to be “close” to the optimum both primally and dually after $\mathcal{O}(1/\varepsilon)$ iterations.

In the analysis presented in this paper, we make use of the following notion of approximation quality introduced in [1].

Definition 1. A feasible solution $\boldsymbol{\alpha}$ to problem (1) is said a Δ -*approximate solution* if

$$\begin{aligned} \Delta^d(\boldsymbol{\alpha}) &\leq \Delta & (9) \\ \text{and } \Delta_i^s(\boldsymbol{\alpha}) &:= \nabla g(\boldsymbol{\alpha})_i - \boldsymbol{\alpha}^T \nabla g(\boldsymbol{\alpha}) \geq -\Delta, \quad \forall i : \alpha_i > 0. & (10) \end{aligned}$$

The first condition guarantees that a Δ -approximate solution is “close” to the optimum both primally and dually. In addition, the second condition ensures that $-\Delta \leq \Delta_i^s(\boldsymbol{\alpha}) \leq \Delta$ for the active face, that is, the primal-dual gap computed on each active coordinate $i : \alpha_i > 0$ is not far from the largest gap computed among all the coordinates of the gradient. This implies also that the solution $\boldsymbol{\alpha}_k$ is “almost” optimal in the face of the simplex defined by the non-zero indices.

2.2 Sparsity of the FW solutions and Coresets

One of the main points of interest for the FW method is the sparsity of the solutions it finds. It should be observed that, in contrast to other methods such as projected or reduced gradient methods, Algorithm 1 modifies only one coordinate of the previous iterate at each step. If the starting solution has K_0 non-zero coordinates, iterate $\boldsymbol{\alpha}_k$ has at most $K_0 + k$ non-zero entries. Therefore, our previous remarks about the convergence of the FW method show that there exist solutions with space-complexity $K_0 + \mathcal{O}(1/\varepsilon)$ that are good approximations for problem (1), even if m (the dimensionality of the feasible space and the number of data points in SVM problems) is much larger.

The above properties are essential for in the context of training non-linear SVMs. In this case, each non-zero coordinate in $\boldsymbol{\alpha}_k$ represents a support training example (a document, image or protein profile) that needs to be explicitly stored in memory during the execution of the algorithm. In addition, the test complexity of non-linear SVMs is proportional to the number of non-zero coordinates in $\boldsymbol{\alpha}_k$, which determines the cost of each iteration in training time, and the cost of a classification decision in testing time.

Existence of sparse approximate solutions for problem (1) can be linked to the idea of ε -coreset, first described for the MEB and other geometrical problems [64]. For $\varepsilon > 0$, an ε -coreset $P' \subset P$ has the property that if the smallest ball containing P' is expanded by a factor of $1 + \varepsilon$, then the resulting ball contains P . That is, if the problem is solved on P' , the solution is “close” to the solution on P . The existence of ε -coresets of size $\mathcal{O}(1/\varepsilon)$ for the MEB problem was first demonstrated by Bădoiu and Clarkson in [9, 10]. Note that in large-scale applications $1/\varepsilon$ can be much smaller than the cardinality of P .

In [13], Clarkson provides a definition of coreset that applies in the general setting of problem (1). Basically, a ε -coreset for problem (1) is a subset of indices spanning a face of S on which we can compute a good approximate solution. The existence of small ε -coresets implies the existence of sparse solutions which are optimal in their respective active faces. The practical consequence of this result would be the possibility of solving large instances of (1) working with a small set of variables of the original problem.

Definition 2. An ε -coreset for problem (1) is a set of indices $\mathcal{I} \subset [m]$ such that the solution $\alpha_{\mathcal{I}}^*$ to the reduced problem

$$\underset{\alpha}{\text{maximize}} \quad g(\alpha) \quad \text{subject to} \quad \alpha \in S_{\mathcal{I}} := \{\alpha \in S : \alpha_i = 0, \forall i \notin \mathcal{I}\} . \quad (11)$$

satisfies $\Delta^d(\alpha_{\mathcal{I}}^*) \leq \varepsilon$.

As discussed in [13], the FW method is not guaranteed to find a ε -coreset after $\mathcal{O}(1/\varepsilon)$ iterations for problem (1). It has been demonstrated that FW is able to find such a coreset in some special cases, e.g., in polytope distance problems [28]. However, in general, $\mathcal{O}(1/\varepsilon^2)$ iterations may be required. Instead, the computationally intensive modification presented in Algorithm 2, generally known as the *fully corrective* variant of FW, does the job.

Algorithm 2: Fully-corrective FW method for problem (1).

- 1 Compute an initial estimate α_0 .
 - 2 Set $\mathcal{I}_0 = \{i : \alpha_{0,i} \neq 0\}$.
 - 3 **for** $k = 0, 1, \dots$ **do**
 - 4 Search for $i^* \in \operatorname{argmax}_i \nabla g(\alpha_k)_i$.
 - 5 Set $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{i^*\}$.
 - 6 Solve the reduced problem (11) with $\mathcal{I} = \mathcal{I}_k$.
-

Note that Algorithm 2 needs to solve an optimization problem of increasing size at each iteration. This can be considered a generalized version of the well-known Bădoiu-Clarkson (BC) method to compute MEBs in computational geometry and, up to our knowledge, corresponds to the first variant of the FW method applied to SVM problems [58].

2.3 Boosting the Convergence using Away-steps

It is well-known that the FW method often exhibits a tendency to stagnate near the solution α^* , resulting in a slow convergence rate [27]. As discussed in [64, 20], this problem can be explained geometrically. Near the solution, the gradient at α_k has a tendency to become nearly orthogonal to the face of the simplex spanned by \mathcal{I}_k (the non-zero coordinates of α_k). Therefore, very little improvement can be achieved by moving α_k towards the ascent vertex \mathbf{u}_k . However, since the solution is not optimal, it is reasonable to think that the solution can be improved working *on the face* spanned by \mathcal{I}_k . Actually, Algorithm 2 works on \mathcal{I}_k until approximate optimality before exploring the next ascent direction.

It can be shown that the convergence of the FW method can be boosted by introducing a new type of optimization step. In short the idea is that, instead of moving *towards* the point \mathbf{u}_k maximizing the local linear approximation $\psi_k(\cdot)$ of $g(\cdot)$, we can move *away* from the point of the current face \mathbf{v}_k minimizing $\psi_k(\cdot)$. At each iteration, a choice between these two options is made by determining which of the directions (moving towards \mathbf{u}_k or moving away from \mathbf{v}_k) is more promising.

Since the point \mathbf{v}_k must lie in the current active face, it is easy to see that the linear approximation step reduces to $\mathbf{v}_k = \mathbf{e}_{j^*}$, where j^* is the smallest active coordinate of the gradient, i.e., $j^* \in \operatorname{argmin}_{j \in \mathcal{I}_k} \nabla g(\alpha_k)_j$. The whole procedure, known as the *Modified Frank-Wolfe* (MFW) method, is summarized in Algorithm 3. In the rest of this paper, we refer to $\mathbf{e}_{j^*} \in S$ and $\mathbf{d}_k^A = (\alpha_k - \mathbf{e}_{j^*})$ as the *descent vertex* and the *away direction* used by the method respectively.

Algorithm 3: MFW method for problem (1).

```

1 Compute an initial estimate  $\alpha_0$ .
2 Set  $\mathcal{I}_0 = \{i : \alpha_{0,i} \neq 0\}$ .
3 for  $k = 0, 1, \dots$  do
4     Search for  $i^* \in \operatorname{argmax}_i \nabla g(\alpha_k)_i$  and define  $\mathbf{d}_k^{FW} = \mathbf{e}_{i^*} - \alpha_k$ .
5     Search for  $j^* \in \operatorname{argmin}_{j \in \mathcal{I}_k} \nabla g(\alpha_k)_j$  and define  $\mathbf{d}_k^A = \alpha_k - \mathbf{e}_{j^*}$ .
6     if  $\nabla g(\alpha_k)^T \mathbf{d}_k^{FW} \geq \nabla g(\alpha_k)^T \mathbf{d}_k^A$  then
7         Perform a line-search to find  $\lambda_{\text{fw}} \in \operatorname{argmax}_{\lambda \in [0,1]} g(\alpha_k + \lambda \mathbf{d}_k^{FW})$ .
8         Perform the FW step  $\alpha_{k+1} = \alpha_k + \lambda_{\text{fw}}(\mathbf{e}_{i^*} - \alpha_k)$ .
9         Update  $\mathcal{I}_k$  by  $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{i^*\}$ .
10    else
11        Perform a line-search to find  $\lambda_{\text{away}} \in \operatorname{argmax}_{\lambda \in [0,1]} g(\alpha_k + \lambda \mathbf{d}_k^A)$ .
12        Clip the line-search parameter,
13         $\lambda_{\text{away}^*} = \max(\lambda_{\text{away}}, \alpha_{k,j^*} / (1 - \alpha_{k,j^*}))$ 
14        Perform the AWAY step  $\alpha_{k+1} = \alpha_k + \lambda_{\text{away}^*}(\alpha_k - \mathbf{e}_{j^*})$ .
15        Set  $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{j^*\}$ .
16        If  $\lambda_{\text{away}^*} = \alpha_{k,j^*}$ ,  $\mathcal{I}_{k+1} = \mathcal{I}_{k+1} \setminus \{j^*\}$ .

```

In contrast to the FW method, for which only a sub-linear rate of convergence can be expected in general [27, 64], it has been shown that MFW asymptotically exhibits linear convergence to the solution of problem (1) under some assumptions on the form of the objective function and the feasible set [27, 64, 1]. In addition, the MFW algorithm has the potential to compute sparser solutions in practice, since in contrast to the FW method it allows reducing the coordinates of α_k at each step.

2.4 Adaptations to SVMs

In the context of SVM learning, the work of Tsang *et al.* in [58] was arguably the first to point out the properties of the algorithms than can be obtained by applying FW methods to formulations fitting problem (1). Their work relies on the equivalence between the SVM problem (2) and a MEB problem, which holds under a normalization assumption on the kernel function employed in the model [58, 59]. Exploiting this equivalence, and adapting the Bădoiu-Clarkson algorithm for computing a MEB to the problem of training non-linear SVMs, an algorithm called Core Vector Machine (CVM) is obtained, which enjoys remarkable theoretical properties and competitive performance in practice [58].

First, the number of support vectors of the model obtained by the CVM is $K_0 + \mathcal{O}(1/\varepsilon)$ where K_0 is a constant and ε is the tolerance parameter of the method. Therefore, the space complexity of the model is independent of the size and dimensionality of the training set. Second, the number of iterations of the algorithm before termination is also $\mathcal{O}(1/\varepsilon)$, independent of the size and dimensionality of the training set. To determine the overall time complexity of this method, we note that Algorithm 2 requires a search for the point i^* representing the best ascent direction in the current approximation of the objective function, an operation that is also performed by the FW and MFW methods. Searching among all of the m training points requires a number of kernel evaluations of order $\mathcal{O}(q_k^2 + mq_k) = \mathcal{O}(mq_k)$, where q_k is the cardinality of \mathcal{I}_k . Since the cardinality of \mathcal{I}_k is bounded as $\mathcal{O}(1/\varepsilon)$ (the worst-case number of iterations), we obtain that the CVM has an overall time complexity (measured as the total number of kernel evaluations) of $\mathcal{O}(1/\varepsilon) * \mathcal{O}(m/\varepsilon) = \mathcal{O}(m/\varepsilon^2)$, linear in the number of examples, improving on the super-linear time complexity reported empirically for popular methods like SMO to train SVMs [45, 16].

If m is very large, however, the complexity per iteration can still become prohibitive in practice. A sampling technique, called *probabilistic speedup*, was proposed in [52] to overcome this obstacle. This technique was also used to implement the CVM in [58, 57] leading to SVM training algorithms with an overall time complexity which is independent of the number of training examples. In practice, the index i^* is computed just on a random subset $\varphi(S') \subset \varphi(S)$ of coordinates, with $|S'| \ll |S| = \text{constant}$. The overall complexity per iteration is thereby reduced to order $\mathcal{O}(q_k^2 + q_k) = \mathcal{O}(q_k^2)$, a major improvement on the previous estimate, since we generally have $q_k \ll m$. Refer to [51] or [58] for details about this speed-up technique.

More recently, several authors have explored the adaptation of the original

FW methods to the task of training SVMs. The advantage of Algorithms 1 and 3 over Algorithm 2 is that they rely only on analytical steps. As a result, each training iteration becomes significantly cheaper than a CVM iteration and does not depend on any external numerical solver. In practice, the training algorithm might probably require more iterations in order to obtain a solution within the predefined tolerance criterion ε , but the work per iteration is significantly smaller. Such a trade-off has been shown to be worthwhile when dealing with large-scale applications [45, 16, 19].

In [19, 20] the authors show that adopting Algorithms 1 and 3 the running times of [58] can be significantly improved as long a minor loss in accuracy is acceptable. From the analysis presented in [13], it is possible to conclude that this approach enjoys similar theoretical guarantees, namely, linear time in the number of examples and a number of iterations which is independent of the number of examples. The sampling technique to speed-up the computation of i^* introduced above can be used with these methods as well, in order to obtain overall time complexities which are independent of the number of training patterns.

In a closely related work [38], Kumar and Yildirim present a specialization of the MFW method to SVM problems, adopting the geometrical formulation studied in [6]. This approach reformulates the SVM problem as a minimum polytope distance problem. The obtained method and its properties are also strongly related to the work of Gartner and Jaggi [28], in which the authors show (theoretically) that the FW method as well as the core-set framework introduced in [13] can be applied to all the currently most used hard and soft margin SVM variants, with arbitrary kernels, to obtain approximate training algorithms needing a number of iterations independent of the number of attributes and training examples. In [44], Ouyang and Gray propose a stochastic variant of FW methods for online learning of L_2 -SVMs, obtaining comparable and sometimes better accuracies than state-of-the-art batch and online algorithms for training SVMs. A similar technique has recently been proposed in [29] to allow smooth and general online convex optimization with sub-linear regret bounds [53]. Variants of the method proposed in [58] have been introduced in [61] and [47] for training SVMs on data streams. In [39] the authors adapted the FW method to train SVMs with structured outputs like graphs and other combinatorial objects [60, 3], obtaining an algorithm which outperforms competing structural SVM solvers⁴.

3 The SWAP Method

We have described in the previous sections how the basic FW method can be modified in order to avoid stagnation near a solution, in this way obtaining an algorithm with a guaranteed rate of convergence. Our previous remarks about the MFW method suggest that this algorithm should terminate faster and find

⁴To be precise, the block-coordinate FW in [39], when applied on the binary SVM as a special case of the structured SVM, becomes a variant of dual coordinate ascent [31].

sparser solutions. In practice however, the MFW method is not always as fast as one could expect from the theory. For instance, the experimental results reported in [64] and [1] for the MEB and Minimum Volume Enclosing Ellipsoid problems respectively, show that very tight improvements, if any, are obtained using the enhanced method (MFW) with respect to the basic approach. As concerns the problem of training SVMs, results in [20] confirm using statistical tests that MFW is not systematically better than FW. Indeed it may sometimes be slower. Similarly, the authors of [44] argue that the use of away steps does not provide a clear advantage with respect to the standard FW method.

A possible interpretation of these results can be given by looking at the way in which MFW implements the away steps to keep feasibility, i.e., to ensure the constraint $\sum_i \alpha_i = 1$ is satisfied. The basic idea in the MFW approach is to include the alternative of getting away from a descent vertex of the current face \mathbf{e}_{j^*} , decreasing the j^* -th weight in α_k , instead of going toward an ascent vertex \mathbf{e}_{i^*} , which would increase the i^* -th weight in α_k . The choice is mutually exclusive. If the algorithm decides to work around j^* , it may lose the opportunity to explore a promising direction of the feasible space, and vice-versa.

On the other hand, if an away step is performed, the weights of the active vertices $i \in \mathcal{I}_k$ are uniformly scaled by $(1+\lambda)$ to keep feasibility. This scheme not only does considerably perturb the current approximation, since all the weights are modified, but, more importantly, can increase the weights of vertices which do not belong to the optimal face S^* . Away steps in the MFW method are thus prone to increase the need of further away steps to eliminate such “spurious points” ($i \in \mathcal{I}_k$, but $i \notin S^*$).

Here, we introduce a new type of away step devised to circumvent these problems while preserving the advantages of MFW. We discuss two variants of the method, obtained by using first and second order approximations of the objective function at each iteration, respectively.

3.1 Main Construction

Our method is obtained as follows. As in the previous FW methods, we find, at each iteration, an *ascent vertex* \mathbf{e}_{i^*} , as

$$i^* \in \underset{i}{\operatorname{argmax}} \nabla g(\alpha_k)_i, \quad (12)$$

and a *descent vertex* \mathbf{e}_{j^*} on the face spanned by the current solution α_k , as

$$j^* \in \underset{j \in \mathcal{I}_k}{\operatorname{argmax}} -\nabla g(\alpha_k)_j = \underset{j \in \mathcal{I}_k}{\operatorname{argmin}} \nabla g(\alpha_k)_j. \quad (13)$$

However, instead of considering the update $\alpha_{k+1} = \alpha_k + \lambda(\alpha_k - \mathbf{e}_{j^*})$ for the away step, we propose a step of the form

$$\alpha_{k+1} = \alpha_k + \lambda(\mathbf{e}_{i^*} - \mathbf{e}_{j^*}), \quad (14)$$

where λ is determined by a line-search. That is, instead of exploring the away direction $\mathbf{d}_k^{\text{MFW}} = (\alpha_k - \mathbf{e}_{j^*})$, our algorithm explores the direction $\mathbf{d}_k^{\text{SWAP}} =$

$(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})$. A sketch is included in Figure (1). This scheme for implementing away steps provides the following conceptual advantages.

1. This away step perturbs the current solution $\boldsymbol{\alpha}_k$ only locally, in the sense that the weight of any vertex other than \mathbf{e}_{i^*} and \mathbf{e}_{j^*} is preserved.
2. This away step does not increase the weight of vertices \mathbf{e}_j of the active face corresponding to descent vertices. These points may correspond to spurious points that need to be removed from the active face to reach the optimal face of the problem.
3. This away step moves the current solution in the away direction and simultaneously in the direction of a *toward step*. That is, it moves away from the descent vertex \mathbf{e}_{j^*} , but also gets closer to the ascent vertex \mathbf{e}_{i^*} in the same iteration. The step (14) can actually be written as the superposition of two separate steps,

$$\begin{aligned} \boldsymbol{\alpha}_{k+1} = & \frac{1}{2} (\boldsymbol{\alpha}_k + \lambda (\mathbf{e}_{i^*} - \boldsymbol{\alpha}_k)) \quad \text{toward step} \\ & + \frac{1}{2} (\boldsymbol{\alpha}_k + \lambda (\boldsymbol{\alpha}_k - \mathbf{e}_{j^*})) \quad \text{away step ,} \end{aligned} \quad (15)$$

where the first term of the right-hand side $\boldsymbol{\alpha}_k + \lambda (\mathbf{e}_{i^*} - \boldsymbol{\alpha}_k)$ represents the standard toward step in the FW method and the second term, $\boldsymbol{\alpha}_k + \lambda (\boldsymbol{\alpha}_k - \mathbf{e}_{j^*})$, the away step considered in the MFW approach. Note that the term $\lambda \boldsymbol{\alpha}_k$ disappears in the sum, so that only the components corresponding to i^* and j^* are updated, leaving the rest of the current solution unchanged.

The new type of away step is called a *SWAP step* and substitutes the MFW away steps in Algorithm 3. The procedure is summarized in Algorithm 4. Note that we deliberately include some steps which do not represent computational tasks but definitions which simplify the convergence analysis of the next section.

So as to choose the type of step to perform, the MFW criterion cannot be employed in our method. The MFW method employs a first order approximation of $g(\cdot)$ at the current iterate to predict the value of the objective function at the next iterate. That is, if \mathbf{d} denotes the search direction,

$$\psi_k(\boldsymbol{\alpha}_k + \lambda \mathbf{d}) = g(\boldsymbol{\alpha}_k) + \lambda \mathbf{d}^T \nabla g(\boldsymbol{\alpha}_k). \quad (16)$$

is computed. The step which gives the largest value of ψ_k is selected. However, a SWAP step *always* gives a larger value of ψ_k than the value obtained using a toward step. Indeed, the value of ψ_k using a SWAP step is

$$\begin{aligned} \psi_k(\boldsymbol{\alpha}_k + \lambda \mathbf{d}_{swap}) &= g(\boldsymbol{\alpha}_k) + \lambda (\mathbf{e}_{i^*} - \mathbf{e}_{j^*})^T \nabla g(\boldsymbol{\alpha}_k) \\ &= g(\boldsymbol{\alpha}_k) + \lambda \nabla g(\boldsymbol{\alpha}_k)_{i^*} - \lambda \nabla g(\boldsymbol{\alpha}_k)_{j^*}. \end{aligned} \quad (17)$$

Algorithm 4: The SWAP algorithm for problem (1).

```

1 Set  $k = 0$ .
2 Compute an initial estimate  $\alpha_0$ .
3 Set  $\mathcal{I}_0 = \{i : \alpha_{0,i} \neq 0\}$ .
4 for  $k = 0, 1, \dots$  do
5     Search for  $i^* \in \operatorname{argmax}_i \nabla g(\alpha_k)_i$  (ascent direction).
6     Search for  $j^* \in \operatorname{argmin}_{j: \alpha_{k,j} \neq 0} \nabla g(\alpha_k)_j$  (descent direction).
7     Perform a line-search to find
8      $\lambda_{\text{swap}} \in \operatorname{argmax}_{\lambda \in [0,1]} g(\alpha_k + \lambda(\mathbf{e}_{i^*} - \mathbf{e}_{j^*}))$ .
9     Perform a line-search to find  $\lambda_{\text{fw}} \in \operatorname{argmax}_{\lambda \in [0,1]} g(\alpha_k + \lambda(\mathbf{e}_{i^*} - \alpha_k))$ .
10    Compute  $\delta_{\text{swap}} = g(\alpha_k + \lambda_{\text{swap}}(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})) - g(\alpha_k)$  (improvement of a SWAP step).
11    Compute  $\delta_{\text{fw}} = g(\alpha_k + \lambda_{\text{fw}}(\mathbf{e}_{i^*} - \alpha_k)) - g(\alpha_k)$  (improvement of a toward step).
12    Compute  $\delta_k = \max(\delta_{\text{swap}}, \delta_{\text{fw}})$  (the best improvement).
13    if  $\delta_k = \delta_{\text{swap}}$  then
14        Clip the line-search parameter,  $\lambda_{\text{swap}^*} = \max(\lambda_{\text{swap}}, \alpha_{k,j^*})$ 
15        If  $\lambda_{\text{swap}^*} = \alpha_{k,j^*}$  mark the iteration as a SWAP-drop step.
16        If  $\lambda_{\text{swap}^*} = \lambda_{\text{swap}}$  mark the iteration as a SWAP-add step.
17        Perform the SWAP step  $\alpha_{k+1} = \alpha_k + \lambda_{\text{swap}^*}(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})$ .
18        Set  $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{i^*\}$ .
19        If a SWAP-drop step was performed,  $\mathcal{I}_{k+1} = \mathcal{I}_{k+1} \setminus \{j^*\}$ .
20    else
21        Mark the iteration as a FW step.
22        Perform the FW step  $\alpha_{k+1} = \alpha_k + \lambda_{\text{fw}}(\mathbf{e}_{i^*} - \alpha_k)$ .
23        Set  $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{i^*\}$ .

```

The value of ψ_k using a toward step is

$$\begin{aligned}
\psi_k(\alpha_k + \lambda \mathbf{d}_{\text{fw}}) &= g(\alpha_k) + \lambda(\mathbf{e}_{i^*} - \alpha_k)^T \nabla g(\alpha_k) \\
&= g(\alpha_k) + \lambda \nabla g(\alpha_k)_{i^*} - \lambda \alpha_k^T \nabla g(\alpha_k).
\end{aligned} \tag{18}$$

Since $\alpha_k^T \nabla g(\alpha_k)$ is always larger than $\nabla g(\alpha_k)_{j^*}$, a SWAP step would always be preferred using first-order information to predict the objective function value.

To address this problem, we observe that the MFW method computes an exact line-search for the search direction selected using ψ_k . We thus formulate our method computing the line-search before deciding the type of step to perform. This design requires to perform two line-searches instead of one. However, the estimation of the objective function value at the next iterate is more accurate.

As we will discuss in the section regarding the adaptation of the procedure to the SVM problem, this computation is particularly simple for the objective function in problem (2). All the computations are analytical. Furthermore, the exact computation of δ_{fw} and δ_{swap} involve terms already computed in the

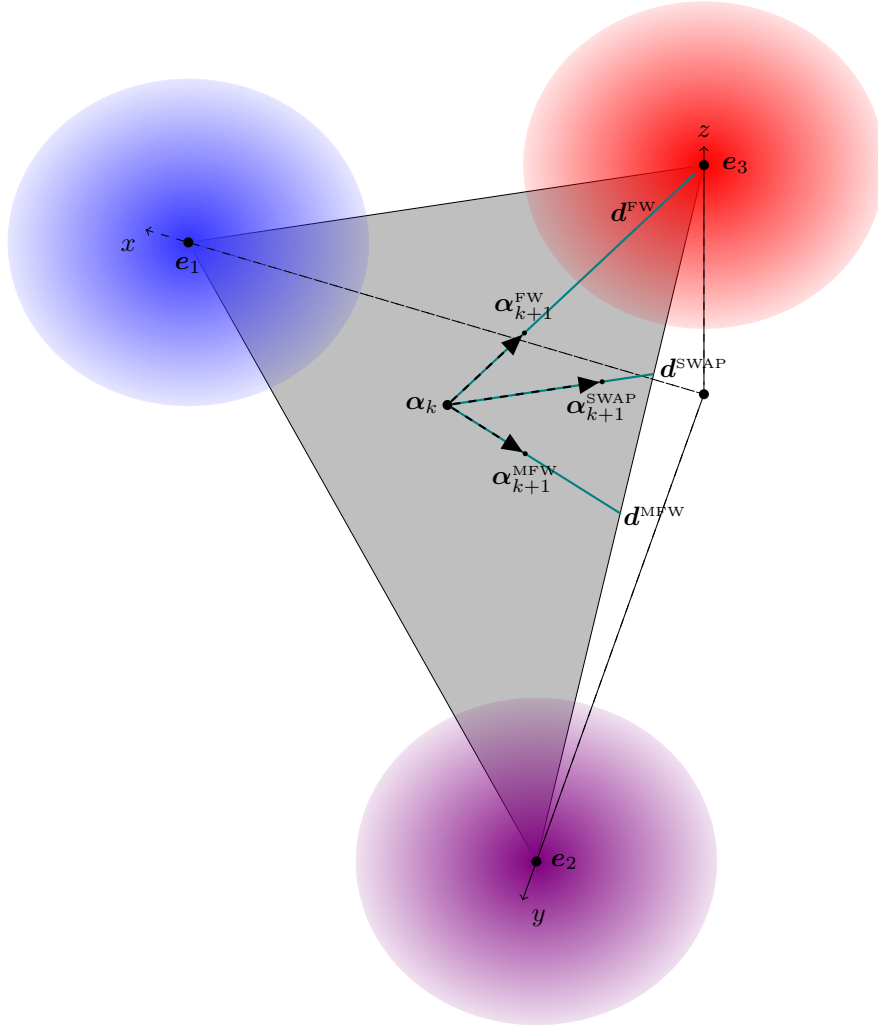


Figure 1: A sketch of the search directions used by FW, MFW and SWAP methods. In this representation, $\mathbf{e}_{i^*} = \mathbf{e}_3$ (ascent vertex wrt the current iterate) and $\mathbf{e}_{j^*} = \mathbf{e}_1$ (descent vertex wrt the current iterate). The search directions explored by the algorithms (solid lines along the updates) are $\mathbf{d}^{\text{FW}} = (\mathbf{e}_3 - \boldsymbol{\alpha}_k)$, $\mathbf{d}^{\text{MFW}} = (\boldsymbol{\alpha}_k - \mathbf{e}_1)$ and $\mathbf{d}^{\text{SWAP}} = (\mathbf{e}_3 - \mathbf{e}_1)$ respectively. Note that MFW and SWAP are more effective than FW in reducing the weight of the descent vertex \mathbf{e}_1 . However, if \mathbf{e}_2 is also a descent vertex, the MFW update has the side effect of increasing the weight of another descent vertex. This is avoided by the SWAP update, which only increases the weight corresponding to the best ascent vertex.

line-searches and therefore does not represent an additional overhead for the algorithm.

3.2 A Variant using Second Order Information

Algorithm 5: The SWAP-2o algorithm for problem (1).

1 Proceed as in Algorithm 4, but modify step 6 as follows:

$$j^* \in \operatorname{argmax}_{j \in I_k} \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_j)^2}{-2(\nabla_{i^*,i^*}^2 - 2\nabla_{i^*,j}^2 + \nabla_{j,j}^2)}. \quad (19)$$

All the FW methods introduced previously make use of first-order approximations of the objective function in order to determine the direction toward which the current iterate should be moved. Here, we consider the possibility of using second-order information. If we assume that the objective function is twice differentiable, the second-order Taylor approximation of $g(\cdot)$ in a neighborhood of $\boldsymbol{\alpha}_k$ is

$$g(\boldsymbol{\alpha}_k + \lambda \mathbf{d}) \approx g(\boldsymbol{\alpha}_k) + \lambda \nabla g(\boldsymbol{\alpha}_k)^T \mathbf{d} + \frac{1}{2} \lambda^2 \mathbf{d}^T \nabla^2 g(\boldsymbol{\alpha}_k) \mathbf{d}, \quad (20)$$

where the Hessian matrix $\nabla^2 g(\boldsymbol{\alpha}_k)$ is negative semi-definite. Finding the best ascent direction would thus require the computation of the quadratic form $\mathbf{d}^T \nabla^2 g(\boldsymbol{\alpha}_k) \mathbf{d}$. Since the matrix $\nabla^2 g(\boldsymbol{\alpha}_k)$ may be highly dense, which is usually the case in SVM applications, employing a first order relaxation as in Frank-Wolfe methods makes sense in order to obtain lighter iterations. However, we note that the search direction for a SWAP step $\mathbf{d}_{\text{swap}} = \mathbf{e}_{i^*} - \mathbf{e}_{j^*}$ yields a particularly simple expression

$$\begin{aligned} & g(\boldsymbol{\alpha}_k + \lambda \mathbf{d}_{\text{swap}}) \\ & \approx g(\boldsymbol{\alpha}_k) + \lambda \nabla g(\boldsymbol{\alpha}_k)^T (\mathbf{e}_{i^*} - \mathbf{e}_{j^*}) + \frac{1}{2} \lambda^2 (\mathbf{e}_{i^*} - \mathbf{e}_{j^*})^T \nabla^2 g(\boldsymbol{\alpha}_k) (\mathbf{e}_{i^*} - \mathbf{e}_{j^*}) \quad (21) \\ & = g(\boldsymbol{\alpha}_k) + \lambda (\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*}) + \frac{1}{2} \lambda^2 (\nabla_{i^*,i^*}^2 - 2\nabla_{i^*,j^*}^2 + \nabla_{j^*,j^*}^2), \end{aligned}$$

where $\nabla_{i,j}^2 = \nabla^2 g(\boldsymbol{\alpha}_k)_{i,j}$.

In order to determine the best pair $\mathbf{e}_{i^*}, \mathbf{e}_{j^*}$ we thus need to evaluate three entries of the Hessian matrix. However this is still a computationally hard task for each iteration, since we would need to consider $m|\mathcal{I}_k|$ pairs of points in order to take a step. We thus adopt the strategy used in the second-order version of SMO proposed in [16]. We fix the ascent index i^* as in the first-order SWAP, and search for the index j^* in the active set which maximizes the improvement of the second order approximation (21). We call the obtained procedure *second-order SWAP* and we denote it as *SWAP-2o* in the next Sections.

It is worth to note that this approximation is exact for quadratic objective functions, which is the case for the SVM problem (2). Note also that in this case the line-search along the ascent direction \mathbf{d}_k defined by i^* and j^* has a closed-form solution. Indeed,

$$\lambda_* = \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*})}{-\left(\nabla_{i^*,i^*}^2 - 2\nabla_{i^*,j^*}^2 + \nabla_{j^*,j^*}^2\right)}. \quad (22)$$

From the negative semi-definiteness of $\nabla^2 g(\boldsymbol{\alpha}_k)$ it follows that λ_* is non-negative. Substituting this step-size in (21), the improvement in the objective function becomes

$$g(\boldsymbol{\alpha}_k + \lambda_* \mathbf{d}_{\text{swap}}) - g(\boldsymbol{\alpha}_k) = \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*})^2}{-2\left(\nabla_{i^*,i^*}^2 - 2\nabla_{i^*,j^*}^2 + \nabla_{j^*,j^*}^2\right)}, \quad (23)$$

which again, from the negative semi-definiteness of $\nabla^2 g(\boldsymbol{\alpha}_k)$, is non-negative. Naturally, we need to restrict the value of λ_* to the interval $[0, 1]$ in order to obtain a feasible solution for the next step. We thus modify Algorithm 4 as specified in Algorithm 5.

3.3 Notes on the Adaptation to SVM Training

Here we provide analytical expressions for all the computations required by Algorithm 4 and Algorithm 5 applied to the the SVM problem (2). Similar expressions follow for any quadratic objective function.

For problem (2), the gradient and Hessian at given iterate $\boldsymbol{\alpha}_k$ take particularly simple expressions:

$$\nabla g(\boldsymbol{\alpha}_k) = -2\mathbf{K}\boldsymbol{\alpha}_k, \quad \nabla^2 g(\boldsymbol{\alpha}_k) = -2\mathbf{K}. \quad (24)$$

Notice that $\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) = 2g(\boldsymbol{\alpha}_k)$. Therefore, the line-searches in Algorithm 4 or Algorithm 5 can be performed analytically as follows. For FW steps,

$$\lambda_{\text{fw}} = \frac{\nabla g(\boldsymbol{\alpha}_k)_{i^*} - 2g(\boldsymbol{\alpha}_k)}{2\left(\mathbf{K}_{i^*,i^*} + \nabla g(\boldsymbol{\alpha}_k)_{i^*} - g(\boldsymbol{\alpha}_k)\right)}. \quad (25)$$

Note that the quantity $\nabla g(\boldsymbol{\alpha}_k)_{i^*}$ has been already computed to find the ascent vertex i^* . For SWAP steps,

$$\lambda_{\text{swap}} = \frac{\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*}}{2\left(\mathbf{K}_{i^*,i^*} - 2\mathbf{K}_{i^*,j^*} + \mathbf{K}_{j^*,j^*}\right)}. \quad (26)$$

Again, the quantity $\nabla g(\boldsymbol{\alpha}_k)_{j^*}$ has been already computed to choose the descent vertex j^* .

The improvement in the objective function can also be calculated analytically. For FW steps,

$$\delta_{\text{fw}} = \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - 2g(\boldsymbol{\alpha}_k))^2}{4\left(\mathbf{K}_{i^*,i^*} + \nabla g(\boldsymbol{\alpha}_k)_{i^*} - g(\boldsymbol{\alpha}_k)\right)}. \quad (27)$$

All the terms involved here have already been computed to perform the line-search. Similarly, for SWAP steps,

$$\delta_{\text{swap}} = \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*})^2}{4(\mathbf{K}_{i^*,i^*} - 2\mathbf{K}_{i^*,j^*} + \mathbf{K}_{j^*,j^*})}. \quad (28)$$

With the exception of the term \mathbf{K}_{i^*,j^*} , all the computations have already been performed to compute δ_{fw} and to choose the descent vertex j^* . We conclude that, compared with MFW procedure, the SWAP method adapted for problem (2) involves the computation of just one additional term, which is an entry of the kernel matrix \mathbf{K} defining the SVM problem.

The objective function value $g(\boldsymbol{\alpha}_k)$ can be computed recursively from the relationship⁵ $g(\boldsymbol{\alpha}_{k+1}) = g(\boldsymbol{\alpha}_k) + \delta_k$. Finally, we observe that the stopping criterion of Eqn. (8) takes the form

$$\Delta^d(\boldsymbol{\alpha}_k) = \nabla g(\boldsymbol{\alpha}_k)_{i^*} - 2g(\boldsymbol{\alpha}_k) \leq \varepsilon, \quad (29)$$

which involves the same already computed terms.

4 Convergence Analysis of the SWAP Method

In this section we study the convergence of the SWAP method on problem (1), of which the L_2 -SVM problem (2) is a particular instance.

We start by demonstrating the global convergence of the SWAP method. Then, we analyze its rate of convergence towards the optimum. For this purpose we will adapt the analysis presented by Ahipasaoglu *et al.* in [1]. Using this framework and using a set of observations concerning the improvement in the objective function after an iteration of the SWAP method, we will be able to prove that the algorithm converges linearly to the optimal value of the objective function. From a theoretical point of view, these results show that the SWAP enjoys the same mathematical properties of the MFW method. Finally, we provide bounds on the number of iterations required to fulfill the stopping condition of Eqn. (8). We demonstrate that the algorithm stops in at most $\mathcal{O}(1/\varepsilon)$ iterations independently of the number of variables m , which coincides with the number of training examples in the SVM problem (2).

Here we only provide proofs for the first-order SWAP method, described in Algorithm 4. However all the convergence results follow easily for the second-order variant as well. The statements and proofs of some of the technical results used in this section can be found in the Appendix.

We develop our analysis under the following assumptions:

- B1. g is twice continuously differentiable;
- B2. There is an optimal solution $\boldsymbol{\alpha}^*$ of the optimization problem satisfying the strong sufficient condition of Robinson in [48].

⁵A similar recursive equation can be derived to handle the case of SWAP-drop steps.

This is the same set of hypotheses imposed by Yildirim in [64] and Ahipasaoglu *et al.* in [1] to study the convergence of Frank-Wolfe methods for the MEB problem and the Minimum Volume Enclosing Ellipsoid problem, respectively.

Remark 1. Robinson’s condition is a general version of the classical second order sufficient condition for a solution α^* to be an isolated local extremum, i.e. locally unique [17]. Referring to the case of a constrained maximization problem for a concave objective $g(\alpha)$, this result requires two conditions to be fulfilled:

- α^* is a KKT point [17].
- The Hessian of the Lagrangian function at α^* behaves as a negative definite matrix (positive definite for minimization problems) along the directions belonging to the *critical cone* of the KKT point [43]. Specialized to a quadratic problem on the simplex, i.e. a problem with the form of (2), this condition assumes the form:

$$\mathbf{y}^T \mathbf{K} \mathbf{y} > 0$$

for all $\mathbf{y} \neq 0$ such that $\alpha_i^* y_i = 0 \forall i$ s.t. $\mathbf{u}_i^* > 0$, $\alpha_i^* y_i \geq 0 \forall i$ s.t. $\alpha_i^* = 0$ and $\mathbf{u}_i^* = 0$, $\sum_i \mathbf{y}_i = 0$ (where \mathbf{u}^* is the vector of Lagrange multipliers at α^* corresponding to inequality constraints, which is unique since the constraints are linear and linearly independent).

The additional analysis in [48], which plays a key role in our convergence analysis, essentially describes the conditions under which the stationary points of a small perturbation of the problem lie in a neighborhood of the solution of the original problem. This is also the key step in the proofs of linear convergence provided in [64] and [1] for the MFW method.

In [27], Guélat and Marcotte analyzed the convergence properties of FW and MFW methods under the following alternative hypotheses:

- A1. ∇g is Lipschitz-continuous on the feasible set;
- A2. $g(\alpha)$ is strongly concave;
- A3. Let α^* be optimal for (1) and T^* be the smallest face of the feasible set containing α^* . Then

$$(\alpha - \alpha^*)^T \nabla g(\alpha^*) = 0 \Leftrightarrow \alpha \in T^* \text{ (strict complementarity).}$$

However, this set of assumptions can be difficult to satisfy in practice. In particular, A3 is a quite strong assumption and cannot be guaranteed in general.

Note that assumption B1 implies A1, as from the mean value theorem it follows that

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \tag{30}$$

for any \mathbf{x}, \mathbf{y} in the unit simplex, where L is the largest eigenvalue (in modulus) of the matrix $\nabla^2 g(\mathbf{z})$ for \mathbf{z} in the unit simplex. In addition, B1 holds most of the times in machine learning problems.

It can also be shown that, if problem (1) is strongly concave, the strong sufficient condition of Robinson holds, i.e. A2 implies B2 [1]. In particular, this is satisfied by the Wolfe dual of the L_2 -SVM problem. This fact has been used by Kumar and Yildirim in [38] to demonstrate the linear convergence of the MFW method on SVM problems which match (1) with quadratic objectives. The convergence of the FW method specialized to a quadratic program (arising from a linear system) has been also studied in [4]. An implicit assumption to demonstrate linear convergence is that the Gram matrix involved in the quadratic form is positive definite. From Remark (1) it is easy to see that, for quadratic objectives and linear constraints, this implies the Robinson condition. Recently, the linear convergence of some variants of the FW method for convex optimization on polytopes has been demonstrated in [25]. The key ingredient in the proof is A2.

It is worth noting that the Robinson condition is not only weaker than A2 in the sense that A2 implies B2, but also in the sense that it is a local property of the objective function at the solution, instead than a global condition on g .

4.1 Global Convergence

Proposition 1. *Suppose hypothesis A1 is satisfied. Starting from any feasible α_0 , Algorithm 4 produces a series of iterates $\{\alpha_k\}_k$ such that $g(\alpha_k)$ converges to $g(\alpha^*)$, where α^* is a solution of problem (1). If α^* is unique, $\{\alpha_k\}_k$ converges to α^* .*

Proof. The key observation is that both FW and SWAP search directions \mathbf{d}_k in Algorithm 4 satisfy

$$g(\alpha^*) - g(\alpha_k) \leq \mathbf{d}_k^T \nabla g(\alpha_k), \quad (31)$$

where $\mathbf{d}_k = (\mathbf{e}_{i^*} - \alpha_k)$ for FW steps and $\mathbf{d}_k = (\mathbf{e}_{i^*} - \mathbf{e}_{j^*})$ for SWAP steps. In the case of FW steps, the result was stated in [27]. However, it is not hard to see that

$$(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})^T \nabla g(\alpha_k) \geq (\mathbf{e}_{i^*} - \alpha_k)^T \nabla g(\alpha_k), \quad (32)$$

and thus Eqn. (31) also holds for SWAP steps. The rest of the proof (see [2]) follows easily by replicating the strategy used to demonstrate Theorem 1 in [27]. \square

Note that this result holds in particular under our set of hypotheses, since, as stated above, B1 implies A1.

4.2 Analysis of the Rate of Convergence

We now prove a linear convergence result for the SWAP algorithm. In the proof, we make use of the following technical Lemma.

Lemma 1. *Suppose B1 holds. After any iteration marked as SWAP-add or FW in Algorithm 4, the iterate α_k is a Δ -approximate solution with $\Delta = \max(2\sqrt{L\delta_k}, 2\delta_k)$.*

Proof. It follows immediately from reordering Eqns. (75) and (88) in the Appendix. \square

Note that this result holds for the SWAP algorithm and not for the FW method, since Eqn. (88) requires the SWAP steps.

Note also that for convergence analysis purposes we can assume that $\delta_k \leq L$ for k sufficiently large. This follows from the fact that Algorithm 4 converges globally and that an iterate α_k generated by the algorithm is always feasible. From the first fact it follows that $g(\alpha^*) - g(\alpha_k)$ becomes arbitrarily small for a sufficiently large k . From the second fact it follows that $g(\alpha^*) \geq g(\alpha_k)$. Since δ_k is the improvement in the objective function at each iteration of Algorithm 4, this quantity will be, from some iteration onwards, lower than any predefined constant, in particular L . Note now that, if $\delta_k < L$, then

$$2\delta_k < 2\sqrt{L\delta_k}. \quad (33)$$

Thus, Lemma 1 states that for sufficiently large k the iterate α_k produced by a SWAP-add or FW step is a Δ -approximate solution with $\Delta = 2\sqrt{L\delta_k}$.

Proposition 2. *Suppose hypotheses B1 and B2 hold. Let α^* be the solution of problem (1). Then, for sufficiently large k , any iteration marked as SWAP-add or FW in Algorithm 4 produces an iterate α_k satisfying the inequality*

$$\frac{g(\alpha^*) - g(\alpha_{k+1})}{g(\alpha^*) - g(\alpha_k)} \leq \left(1 - \frac{1}{M}\right) \quad (34)$$

for some constant $M > 1$.

Proof. Lemma 1 shows that for sufficiently large k the iterate α_k produced by Algorithm 4 after a SWAP-add or FW step is a Δ -approximate solution, with $\Delta = 2\sqrt{L\delta_k}$. In addition, since the SWAP is globally convergent, δ_k can be chosen to be arbitrarily small. Thus, for k large enough, the conditions of Lemma 5 hold with $\Delta_\star = 2\sqrt{L\delta_k}$. From Eqn. (62), we then have that there exists a constant N such that

$$g(\alpha^*) - g(\alpha_k) \leq Nm \left(2\sqrt{L\delta_k}\right)^2 = 4NmL\delta_k, \quad (35)$$

where $m \gg 1$ is the dimensionality of α , N is a Lipschitz constant depending on the problem, and L is the largest eigenvalue (in modulus) of the Hessian matrix of $g(\alpha)$ on the simplex. Now, for a SWAP-add or a FW step we have, by definition of δ_k ,

$$g(\alpha_{k+1}) - g(\alpha_k) = \delta_k. \quad (36)$$

Note that the latter is not true for SWAP-drop steps because the real improvement in the objective function differs from the value computed to decide the type of step to perform. Thus,

$$g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_k) \leq M (g(\boldsymbol{\alpha}_{k+1}) - g(\boldsymbol{\alpha}_k)) , \quad (37)$$

with $M = 4NmL$. Adding and subtracting $Mg(\boldsymbol{\alpha}^*)$ to the right-hand side produces

$$g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_k) \leq M (g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_k)) - M (g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_{k+1})) . \quad (38)$$

Equivalently,

$$\begin{aligned} M (g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_{k+1})) &\leq M (g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_k)) - (g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_k)) \\ g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_{k+1}) &\leq \left(1 - \frac{1}{M}\right) (g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_k)) . \end{aligned} \quad (39)$$

Thus,

$$\frac{g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_{k+1})}{g(\boldsymbol{\alpha}^*) - g(\boldsymbol{\alpha}_k)} \leq \left(1 - \frac{1}{M}\right) . \quad (40)$$

□

This result is analogous to the linear convergence theorems obtained in [1] and [64] for the MFW algorithm.

Proposition 3. *At any iteration of Algorithm 4, the number of SWAP-drop steps does not exceed a half of the total number of steps T made by the algorithm, plus a finite constant.*

Proof. Let F be the number of FW steps, S the number of SWAP-add steps, C the number of SWAP-drop steps and A the number of steps that include points to the coreset \mathcal{I}_k . We have $A \leq F + S$, because only FW steps and SWAP-add steps can add points to the coreset. Sometimes they include new points, sometimes they do not. Clearly, $T = F + S + C$. Thus, from the previous inequality we have $T \geq A + C$. Now, the number of steps C that drop points from the coreset cannot be greater than the number of steps that add points to the coreset plus the number of points I in the coreset just after initialization, that is, $A + I \geq C$. Combining the last two inequalities leads $T + A + I \geq A + 2C$, that is, $T + I \geq 2C$. Therefore $C \leq \frac{T}{2} + \frac{I}{2}$, which concludes the proof. □

Proposition 2 states that there exist a subsequence of the iterates $\{\boldsymbol{\alpha}_k\}_k$ produced by Algorithm 4 such that $\{g(\boldsymbol{\alpha}_k)\}_k$ converges linearly to the optimal value $g(\boldsymbol{\alpha}^*)$ of the objective function in problem (1). This subsequence is obtained by dropping from $\{\boldsymbol{\alpha}_k\}_k$ the iterates corresponding to SWAP-drop steps, for which we can only say that the objective function value does not decrease. Thanks to Proposition 3, we know that these steps do not affect the overall complexity bound on the number of iterations needed to achieve a given accuracy.

4.3 Iteration Complexity Bounds

We start by proving the following lemma.

Lemma 2. *Suppose B1 holds. By using the stopping condition of Eqn. (8), any iteration marked as SWAP-add or FW in Algorithm 4 produces an improvement in the objective function*

$$\delta_k \geq \min\left(\frac{\varepsilon^2}{4L}, \frac{\varepsilon}{2}\right), \quad (41)$$

where ε is the tolerance parameter.

Proof. If the algorithm enters the loop after checking the stopping condition of Eqn. (8),

$$\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) > \varepsilon. \quad (42)$$

From Eqn. (74) we obtain

$$\max\left(2\sqrt{L\delta_k}, 2\delta_k\right) > \varepsilon, \quad (43)$$

which leads to the result ⁶. □

Note that the converse is not true. The algorithm can stop even if the improvement in the objective function in the last iteration was greater than $\min\left(\frac{\varepsilon^2}{4L}, \frac{\varepsilon}{2}\right)$. This happens because the proposed termination criterion fundamentally looks at the possible improvement with standard FW steps.

Proposition 4. *Let K be the number of iterations performed by Algorithm 4 until the stopping condition of Eqn. (8) is fulfilled. Then, under the hypotheses of Lemma 2,*

$$K \leq Q + \frac{M}{\varepsilon}, \quad (44)$$

where Q, M are constants independent of m and ε .

Proof. Let $k(\delta)$ denote the number of iterations of Algorithm 4 from the first iterate such that the primal-dual gap satisfies $\Delta^d \leq \delta$ until the first that satisfies $\Delta^d \leq \delta/2$. Since the total improvement in the objective function cannot be greater than δ and the improvement in the objective function given by a SWAP-add or a FW step is at least that of Lemma 2 with $\varepsilon = \delta/2$, we can bound $k(\delta)$ as follows⁷:

$$k(\delta) \leq 2 \frac{\delta}{\frac{(\delta/2)^2}{4L}} = 2 \frac{16L}{\delta} = \frac{32L}{\delta}, \quad (45)$$

where the multiplying factor 2 comes from the fact that the total number of iterations is at most two times the number of SWAP-add and FW iterations

⁶Note that the previous proof is based on the minimal improvement of a FW step. The results holds in general because a SWAP step is performed if and only if the unconstrained SWAP yields a larger improvement.

⁷We assume, for the sake of simplicity, that $\varepsilon < L/2$. Otherwise the proof can be adapted.

plus a finite constant (see the discussion in the proof of Proposition 3). Now, let $K(\varepsilon)$ the number of iterations from the first iterate such that the primal-dual gap satisfies $\Delta^d \leq 1$ until the first that satisfies $\Delta^d \leq \varepsilon$. Clearly, $K \leq K(\varepsilon)$. Now, it is not hard to see that $\lceil \log_2 1/\varepsilon \rceil - 1$ is the smallest positive integer p such that $1/2^p \leq 2\varepsilon$. Therefore, we can bound $K(\varepsilon)$ as:

$$\begin{aligned} K(\varepsilon) &\leq k(1) + k\left(\frac{1}{2}\right) + k\left(\frac{1}{4}\right) \dots + k\left(\frac{1}{2^{\lceil \log_2 1/\varepsilon \rceil - 1}}\right) \\ K(\varepsilon) &\leq 32L \left(1 + 2 + 4 \dots + 2^{\lceil \log_2 1/\varepsilon \rceil - 1}\right) \\ &\leq 32L \left(2^{\lceil \log_2 1/\varepsilon \rceil} - 1\right) \leq \frac{64L}{\varepsilon}. \end{aligned} \tag{46}$$

Set $M = 64L$ and Q to the number of iterations required to obtain an iterate satisfying $\Delta^d \leq 1$ (which is finite and independent of ε) to obtain the result. \square

It is also possible to provide a logarithmic bound in $1/\varepsilon$. However, in this case, both the multiplicative and additive constants depend on m . Thus, from such result alone we cannot infer the important property that the overall complexity of the algorithm can be bounded independently from the problem size. Furthermore, if m is comparable to or larger than $1/\varepsilon$ (which is often the case in large-scale applications), there is no guarantee that the obtained bound is tighter than the one given by Eqn. (44). The proof of this result, which we state below for completeness, can be found in [2].

Proposition 5. *Suppose B2 holds. Let K be the number of iterations performed by Algorithm 4 until the stopping condition of Eqn. (8) is fulfilled. Then, there exists $\varepsilon_0 > 0$ such that, if $\varepsilon < \varepsilon_0$,*

$$K \leq \tilde{Q} + \tilde{M} \log_2 \left(\frac{1}{\varepsilon}\right), \tag{47}$$

where \tilde{Q} and \tilde{M} are constants independent of ε but dependent on m . In particular, $\tilde{M} \propto m$.

5 Relation of the SWAP to Geometric Algorithms, SMO and Other Results

In the last years, a number of authors have proposed training SVMs by first reducing the task to a computational geometry problem, and then applying a dedicated algorithm to obtain an exact or approximate solution. Some of these approaches are indeed specialized versions of FW methods. For instance, the so-called Gilbert method [26] can be used to train SVMs by approaching the task as a Minimum Norm Problem (MNP) or a Nearest Point Problem (NPP) [36]. It has been noted in [28] that the FW method is equivalent to the Gilbert method on these geometric problems. Thus, up to some implementation details,

specializations of the FW and Gilbert methods to the SVM problem coincide. Similarly, the Bădoiu-Clarkson algorithm [9] can be used to train SVMs which admit a reduction to a Minimum Enclosing Ball (MEB) problem [58]. Nowadays is well-understood that this algorithm is nothing else than the fully corrective FW method (Algorithm 2) applied to the MEB [13]. Finally, the algorithms proposed in [38] are direct applications of the FW and MFW methods to SVM models which admit an interpretation as an NPP, and the algorithms proposed in [20] are the corresponding application to SVM problems which admit an interpretation as an MNP.

Recently, in [40], the Mitchell-Demnyanov-Malozemov (MDM) algorithm [42], another classic geometric algorithm to solve MNPs and NPPs, was found to be essentially equivalent to the SMO algorithm [45, 16] devised specifically for SVMs and similar quadratic programs [37]. In this section, we show that the method proposed in this paper is closely related to the Gilbert and MDM algorithms when applied to MNPs or NPPs. Indeed, on these specific problems, the SWAP can be considered a Gilbert method with the possibility of performing MDM steps.

Polytope Problems and SVMs Problem (2) can be cast as an instance of the MNP, which consists in finding the point in a polytope nearest to the origin: minimize $_{\mathbf{z}} \|\mathbf{z}\|^2$ subject to $\mathbf{z} \in \mathcal{Z}$. In this case, the polytope \mathcal{Z} is the convex hull of a finite set of points $Z = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m\}$ in a dot product space and the MNP admits the following formulation

$$\underset{\alpha \in \mathbb{R}^m}{\text{minimize}} \quad \frac{1}{2} \alpha^T \mathbf{Z}^T \mathbf{Z} \alpha \quad \text{subject to} \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1, \quad (48)$$

where \mathbf{Z} is the matrix with the points \mathbf{z}_i arranged in the columns. Any feasible solution α for problem (48) yields a feasible solution \mathbf{z} to the original problem by setting $\mathbf{z} = \mathbf{Z}\alpha$. The feasible space in problem (48) corresponds to the unit simplex and the objective function is convex. Thus, the MNP is an instance of the more general problem (1) studied in this paper, with $g(\alpha) = -\frac{1}{2} \alpha^T \mathbf{Z}^T \mathbf{Z} \alpha =: g_G(\alpha)$. Furthermore, it is not hard to see that (48) is an instance of the SVM problem (2) with $\mathbf{K} = \mathbf{Z}^T \mathbf{Z}$. Similarly, (2) is an instance of (48) in the space spanned by the points $\mathbf{z}_i = (y_i \phi(\mathbf{x}_i)^T, y_i, \frac{1}{\sqrt{C}} \mathbf{e}_i^T)^T$, where $\phi(\mathbf{x}_i)$ corresponds to any feature map associated to the kernel used by the SVM. Therefore, algorithms devised to solve MNPs may be adapted to train L_2 -SVMs and vice-versa. This equivalence was first pointed out and used to build a training algorithm in [36].

Other SVM formulations admit similar geometric interpretations (see [13] and Table 1 in [28]). For instance, hard-margin SVMs have been shown to be equivalent to an NPP, i.e. the problem of computing the pair of nearest points between two polytopes $\mathcal{Z}_1, \mathcal{Z}_2$: minimize $_{\mathbf{z}_1, \mathbf{z}_2} \|\mathbf{z}_1 - \mathbf{z}_2\|^2$ subject to $\mathbf{z}_1 \in \mathcal{Z}_1, \mathbf{z}_2 \in \mathcal{Z}_2$. Some variants of the L_2 -SVM considered here can be transformed into a classic hard-margin SVM and thus admit an NPP interpretation [36, 24]. Soft margin L_1 -SVMs and ν -SVMs are essentially NPPs with additional constraints of the form $\alpha_i \leq \gamma$ [5, 11]. The authors of [28] use all these equivalences to

characterize the sparsity of the solutions and show the existence of linear time training algorithms for all the currently most used SVM variants.

The Gilbert and MDM Algorithms Two classic iterative methods used to solve MNPs are the Gilbert algorithm [26] and the Mitchell-Demnyanov-Malozemov (MDM) algorithm [42]. These methods can be easily adapted to solve NPPs by means of the so called Minkowski difference trick [28]. An iteration of the Gilbert algorithm for problem (48) can be written as $\mathbf{z}_{k+1} = (1 - \lambda_k)\mathbf{z}_k + \lambda_k\mathbf{z}_{i^*}$, where λ_k is chosen by performing an exact line-search to minimize $\|\mathbf{z}_{k+1}\|^2$, and

$$\mathbf{z}_{i^*} \in \underset{\mathbf{z}_i \in \mathcal{Z}}{\operatorname{argmin}}(\mathbf{z}_i^T \mathbf{z}_k). \quad (49)$$

It can be shown that the specializations of the Gilbert algorithm and the FW method for problem (48) coincide. Note that $\nabla g_G(\boldsymbol{\alpha}_k) = -\mathbf{Z}^T \mathbf{Z} \boldsymbol{\alpha}_k$ and $\nabla g_G(\boldsymbol{\alpha}_k)_i = -\mathbf{z}_i^T \mathbf{Z} \boldsymbol{\alpha}_k$. Therefore, applied to (48), an iteration of the FW method can be written as $\boldsymbol{\alpha}_{k+1} = (1 - \lambda_k)\boldsymbol{\alpha}_k + \lambda_k \mathbf{e}_{i^*}$ with $i^* \in \operatorname{argmax}_i(-\mathbf{z}_i^T \mathbf{Z} \boldsymbol{\alpha}_k)$. By setting $\mathbf{z}_k = \mathbf{Z} \boldsymbol{\alpha}_k$, this iteration translates into $\mathbf{z}_{k+1} = (1 - \lambda_k)\mathbf{z}_k + \lambda_k \mathbf{z}_{i^*}$, where λ_k is obtained by a line search and $i^* \in \operatorname{argmin}_i(\mathbf{z}_i^T \mathbf{z}_k)$, which is exactly what the Gilbert method does.

As the Gilbert method, the MDM algorithm updates \mathbf{z}_k towards the direction given by the point \mathbf{z}_{i^*} . However, the search direction is determined using an additional point of the polytope \mathcal{Z} corresponding to

$$\mathbf{z}_{j^*} \in \underset{\mathbf{z}_j \in \mathcal{Z}}{\operatorname{argmax}}(\mathbf{z}_k^T \mathbf{z}_j). \quad (50)$$

An iteration of the MDM algorithm can be written as $\mathbf{z}_{k+1} = \mathbf{z}_k + \lambda_k (\mathbf{z}_{i^*} - \mathbf{z}_{j^*})$, where λ_k is computed by a line search. It has been recently shown in [40] that the well-known SMO algorithm [45, 16, 37], devised to train SVMs and to solve similar quadratic programs, is essentially equivalent to MDM considering SVMs which reduce to NPPs. The key difference is that, on NPPs, MDM chooses \mathbf{z}_{i^*} and \mathbf{z}_{j^*} on the same polytope, which correspond to a class of the SVM problem. SMO instead can choose \mathbf{z}_{i^*} and \mathbf{z}_{j^*} from different classes. This difference however disappears in applications of MDM to MNPs, since in that case the geometric problem involves only one polytope.

Relation of Gilbert, MDM and SMO to SWAP From the previous discussion we have that the ascent direction explored by the Gilbert method is given by $(\mathbf{z}_{i^*} - \mathbf{z}_k)$, with

$$i^* \in \underset{i}{\operatorname{argmax}} \nabla g_G(\boldsymbol{\alpha}_k)_i \Leftrightarrow i^* \in \underset{i}{\operatorname{argmax}}(-\mathbf{z}_i^T \mathbf{Z} \boldsymbol{\alpha}_k) \Leftrightarrow i^* \in \underset{i}{\operatorname{argmin}}(\mathbf{z}_i^T \mathbf{z}_k). \quad (51)$$

From here, it is straightforward to see that

$$j^* \in \underset{j}{\operatorname{argmin}} \nabla g_G(\boldsymbol{\alpha}_k)_j \Leftrightarrow j^* \in \underset{j}{\operatorname{argmin}}(-\mathbf{z}_j^T \mathbf{Z} \boldsymbol{\alpha}_k) \Leftrightarrow j^* \in \underset{j}{\operatorname{argmax}}(\mathbf{z}_j^T \mathbf{z}_k), \quad (52)$$

which corresponds to the *away* vertex used by the SWAP method. Thus, the MDM algorithm updates the current iterate using the same vertices of the polytope that would be considered by a specialization of SWAP to problem (48) or to the equivalent SVM problem (2): \mathbf{z}_{i^*} and \mathbf{z}_{j^*} . If SWAP goes for a toward step, it is identical to the FW method on that iteration. From the equivalence between the FW and Gilbert methods, we conclude that the SWAP is identical to Gilbert at iteration k if it decides to not explore the away direction. Otherwise, the search direction used by the SWAP on the simplex is $(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})$, which translates to $\mathbf{d}^{\text{SWAP}} = (\mathbf{z}_{i^*} - \mathbf{z}_{j^*})$ for an MNP. The update $\boldsymbol{\alpha}_{k+1} = \boldsymbol{\alpha}_k + \lambda(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})$ for problem (1) corresponds to updating \mathbf{z}_k as $\mathbf{z}_{k+1} = \mathbf{z}_k + \lambda_k(\mathbf{z}_{i^*} - \mathbf{z}_{j^*})$ and computing λ_k by a line search. This is exactly the same direction and procedure to set the step size used by MDM.

We conclude that, applied to polytope problems, the SWAP is equivalent to a Gilbert method with the possibility of performing MDM steps. In this sense, the SWAP is a kind of hybrid Gilbert-MDM which is presented and analyzed for problems beyond the MNP and NPP. Considering the equivalence between MDM and the SMO on SVM problems, we can also state that on these problems the SWAP is a FW method with the possibility of performing SMO steps. However, again, our presentation and analysis is not limited to quadratic forms. The minor variant of the method, using second order information, uses essentially the same criterion proposed in [16] to improve on the original Platt’s SMO [45].

It is well known from its introduction that, in general, the Gilbert method converges sub-linearly, that is, the specialization of the FW method to the quadratic program in (48) does not improve its rate of convergence [26, 41]. However, it has been recently shown that the MDM algorithm converges linearly under some assumptions about the structure of problem (48), which include the positive definiteness of the matrix $\mathbf{Z}^T \mathbf{Z}$ [41]. In this paper, the analysis addresses a general maximization problem on the simplex and the results rely on hypotheses slightly more general than those used in [41], in the sense that asking for a positive definite matrix $\mathbf{Z}^T \mathbf{Z}$ is equivalent to asking for a positive definite Hessian and this in turn implies the Robinson condition used in our analysis⁸.

6 Experiments

In this section, we present several experiments conducted on benchmark classification datasets to evaluate the performance of the proposed methods and related approaches in practice.

Datasets The datasets used in this section are listed in Table 1 and can be found in several public repositories [12, 22]. In order to provide the reader with an idea of the size of each problem, we specify the size m of the training set, the

⁸See Remark 1. As we have only linear constraints, the Hessian of the Lagrangian coincides with the Hessian of the objective.

number of features n , and the number of classes K . We denote by t the number of *test examples*, set aside to evaluate the expected accuracy of the computed classifier.

In the case of multi-category classification problems, we adopt a one-versus-one approach (OVO) [30]⁹. Note that in these cases the number of examples m does not necessarily reflect the complexity of the training problems to be addressed. For example, according to m , the **MNIST** and **Web w8a** datasets have a similar size. However, the **MNIST** problem has 10 classes and the largest binary problem to solve in the OVO scheme has around 13.000 training examples. The **Web w8a** problem is, in contrast, binary, and thus the whole dataset needs to be handled simultaneously. For this reason, we also report in Table 1 the size m_{\max} of the largest binary subproblem and the size m_{\min} of the smallest binary subproblem in the OVO decomposition.

Initialization and Parameters For the initialization of the CVM, FW, MFW and SWAP methods, that is, the computation of a starting solution, we adopted the method proposed for the CVM in [58]. In this approach, the starting solution is obtained by solving problem (2) on a random subset \mathcal{I}_0 of p training patterns. The indices of α_0 corresponding to other data points are set to zero. We used $p = 20$ points for initialization and $\epsilon = 10^{-6}$ for all the algorithms.

In all but the last experiment described in this section, SVMs were trained using a RBF (Gaussian) kernel

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2}\right), \quad (53)$$

with scale parameter σ^2 . For the relatively small datasets **Pendigits** and **USPS**, parameter σ^2 was determined together with parameter C of SVMs using 10-fold cross-validation on the logarithmic grid $[2^{-15}, 2^5] \times [2^{-5}, 2^{15}]$, where the first collection of values corresponds to parameter σ^2 and the second to parameter C .

For the large-scale datasets, σ^2 was determined using the default method employed for CVM in [58], i.e. it was set to the average squared distance among training patterns. Parameter C was determined on the logarithmic grid $[2^0, 2^{12}]$ using a validation set consisting in a randomly computed 30% of the training-set.

We emphasize that the aim of this paper is not to determine optimal parameter values by fine-tuning each algorithm to seek for the best possible accuracy. Our aim is to compare the performance of the presented methods and analyze their behavior in a manner consistent with our theoretical analysis. Therefore, it is necessary to perform the experiments under the same conditions on a given dataset. That is to say, the optimization problem to be solved should be the same for each algorithm. For this reason, we deliberately avoided using different

⁹This was the method used in [58] to extend the CVM beyond binary classification, and according to [32] it usually outperforms other approaches both in terms of accuracy and training time.

training parameters when comparing different methods. Specifically, parameters σ^2 and C were tuned using the CVM method and the obtained values were used for all the algorithms discussed in this paper (CVM, FW, MFW and SWAP methods).

Caching We also adopted the LRR caching strategy designed in [57] for the CVM to avoid the computation of recently used kernel values.

Dataset	m	t	K	m_{\max}	m_{\min}	n
USPS	7291	2007	10	2199	1098	256
Pendigits	7494	3498	10	1560	1438	16
Letter	15000	5000	26	1213	1081	16
Protein	17766	6621	3	13701	9568	357
Shuttle	43500	14500	7	40856	17	9
IJCNN	49990	91701	2	49990	49990	22
MNIST	60000	10000	10	13007	11263	780
USPS-Ext	266079	75383	2	266079	266079	676
KDD-10pc	395216	98805	5	390901	976	127
KDD-Full	4898431	311029	2	4898431	4898431	127
Reuters	7770	3299	2	7770	7770	8315
Adult a1a	1605	30956	2	1605	1605	123
Adult a2a	2265	30296	2	2265	2265	123
Adult a3a	3185	29376	2	3185	3185	123
Adult a4a	4781	27780	2	4781	4781	123
Adult a5a	6414	26147	2	6414	6414	123
Adult a6a	11220	21341	2	11220	11220	123
Adult a7a	16100	16461	2	16100	16100	123
Web w1a	2477	47272	2	2477	2477	300
Web w2a	3470	46279	2	3470	3470	300
Web w3a	4912	44837	2	4912	4912	300
Web w4a	7366	42383	2	7366	7366	300
Web w5a	9888	39861	2	9888	9888	300
Web w6a	17188	32561	2	17188	17188	300
Web w7a	24692	25057	2	24692	24692	300
Web w8a	49749	14951	2	49749	49749	300

Table 1: Features of the selected datasets.

Assessed Algorithms, Notation and Statistics In this paper we have introduced two variants of the FW method: the SWAP, and the second-order SWAP. The acronyms used to denote these algorithms in the figures will be

SW and SW-2o, respectively. We will compare these methods against the CVM algorithm [58], the FW method and the MFW method.

In the next sections we report test accuracies, training times and model sizes obtained on the classification problems of Table 1. By test accuracy we intend the fraction of correctly classified test instances. Training time is the time in seconds required to obtain a model from the training set. When times differ by more than one order of magnitude among the different methods, we use a logarithmic scale to present figures. Model size is the number of training examples with non-zero weights at the end of the training process, that is, the number of support vectors in the model.

To obtain a more detailed comparison, we compute the speed-ups obtained by the Frank-Wolfe based algorithms with respect to the CVM method. The speed-up of the FW method with respect to CVM will be measured as $s_1 = t_0/t_1$ where t_0 is the training time of the CVM algorithm and t_1 is the training time of the FW method, both measured in seconds. Similarly, the speed-up of the MFW, SWAP and SWAP-2o methods with respect to CVM is measured as $s_2 = t_0/t_2$, $s_3 = t_0/t_3$, $s_4 = t_0/t_4$ respectively, where t_2 is the training time of the MFW method, t_3 is that of SWAP, and t_4 that of SWAP-2o. In addition, we quantify the difference in testing performance with respect to the CVM method. If we denote by a_0 the accuracy of CVM and by a_1 the accuracy of the FW method, the relative difference in accuracy incurred by FW will be quantified as $d_1 = (a_0 - a_1)/a_0$. Similarly, differences in testing performance corresponding to the methods MFW, SWAP and SWAP-2o will be measured as $d_2 = (a_0 - a_2)/a_0$, $d_3 = (a_0 - a_3)/a_0$ and $d_4 = (a_0 - a_4)/a_0$, where a_2 , a_3 and a_4 are the testing accuracies of the MFW, SWAP and SWAP-2o methods respectively.

Computational Environment The experiments were conducted on a personal computer with a 2.66GHz Quad Core CPU and 4 GB of RAM, running 64bit GNU/Linux. The algorithms were implemented based on the C++ source code available at [57].

6.1 Experiments on the Web Dataset Collection

The Web Dataset Collection is a series of classification problems extracted from a webpage categorization dataset, first appeared in Platt’s paper on Sequential Minimal Optimization for training SVMs [45]. The number of training patterns in each instance of the collection grows approximately as $m_i = 1.4^i m_0$, $i = 1, \dots, 8$, where m_0 is the number of training patterns in the first dataset. This scheme makes the series amenable for studying performance and scalability of different training algorithms.

Figures 2(a), 3(a) and 4(a) report test accuracies, training times and model sizes (number of support vectors) obtained in this collection. Note that times are depicted in a logarithmic scale. From Figure 2(a) and Figure 3(a) we confirm that all the Frank-Wolfe based methods are slightly less accurate than CVM but exhibit running times that scale considerably better as the number of training

patterns increases. Each of them is faster than CVM on all the 8 datasets of the collection.

Figure 3(a) illustrates one of the main points of this paper: the theoretical advantages of the MFW method over the basic FW routine often do not correspond to an improvement in practical performance. This collection of problems is actually an extreme case, in which MFW is always significantly slower than FW. In contrast the proposed methods are faster than MFW and competitive with the FW method.

From Figure 3(a), we can observe that the speed-ups of the FW method seem to increase monotonically as the number of training patterns increases, ranging from $12.6\times$ faster up to $\sim 106\times$ faster than CVM. Speed-ups corresponding to the MFW method are in contrast significantly more limited. The SWAP algorithm is clearly more competitive than MFW, with a speed-up of $\sim 250\times$ on the largest dataset.

Both MFW and SWAP endow the basic Frank-Wolfe procedure with away-steps, and both, in contrast to FW, offer a guarantee on the rate of convergence. However, the away steps implemented by SWAP and SWAP-2o work significantly better on this collection of datasets. SWAP-2o however does not perform better than SWAP in this series. We argue that standard away steps do not provide any significant advantage on this particular problem, as proved by MFW resulting to be the slowest algorithm. Since SWAP-2o invests more time in finding a good away direction, finding a solution takes more time in comparison with the simpler SWAP, which seems to provide a better compromise between away and toward steps.

As regards accuracy, MFW is slightly more accurate than SWAP, which in turn is slightly more accurate than FW most of the time. SWAP-2o very often outperforms the other three methods, approaching the accuracy of CVM. Note however that all the relative differences in testing accuracy are most of the time below 0.5%. Note finally that FW is the less accurate among the Frank-Wolfe based methods.

As concerns model sizes, note that the additional computational time incurred by the MFW and SWAP-2o methods is not compensated by an improved ability to find smaller models. Figure 4(a) actually shows that the two faster methods, SWAP and FW, obtain most of the time smaller models. Finally, the size of the models found by CVM is significantly larger than that of the proposed methods. In addition, the percentage of training data used by this method to build the model does not seem to decrease significantly as the series progresses.

6.2 Experiments on the Adult Dataset Collection

The Adult Dataset Collection is a series of problems derived from the 1994 US Census database. The goal is to predict whether an individual’s income exceeded 50000US\$/year, based on personal data. Like the **Web** datasets, this collection was designed with the purpose of analyzing the scalability of SVM

methods. The number of training patterns grows approximately with the same rate, i.e. it increases by a factor of ~ 1.4 each time [45].

Figures 2(b), 3(b) and 4(b) depict accuracies, running times and model sizes (number of support vectors) obtained on this collection. Times are depicted in a logarithmic scale. These results confirm that all the Frank-Wolfe based methods tend to be faster than the CVM algorithm as the number of examples becomes larger. Figure 3(b) shows that SWAP, MFW and SWAP-2o always run faster than CVM, reaching speed-ups of $27\times$, $20\times$ and $15\times$ respectively. Figure 3(a) shows in addition that most of the times the Frank-Wolfe based methods achieve a testing performance greater or equal than CVM.

Note that the speed-ups obtained by the FW method in this experiment are significantly smaller than those obtained in the **Web** collection. The largest speed-up achieved by the algorithm is $3.6\times$ on the sixth dataset of the collection. In contrast, the methods investigated in this paper, SWAP and SWAP-2o, always show speed-ups larger than $10\times$, running faster than FW in all cases. If we compute the median speed-up among all the datasets of this collection, the results for SWAP and SWAP-2o are $15.5\times$ and $20.5\times$ respectively. In contrast, the FW method achieves a median of just $1.45\times$. We conclude that the proposed methods are one order of magnitude faster than the basic FW method in this experiment.

The previous remark suggests that away steps are very useful to speed up the algorithm towards an optimal face in this problem. We confirm this observation by examining the performance of the MFW method in this experiment. Figure 3(b) shows that the MFW method is always faster than FW. This result contrasts with our previous experiment in which MFW was always slower than FW. We conclude that in this experiment all the algorithms incorporating away steps are significantly faster than the algorithms which do not. Note that the proposed methods SWAP and SWAP-2o always run faster than MFW.

As regards testing accuracy, the CVM is most of the time slightly less accurate than Frank-Wolfe methods in this experiment. SWAP always obtains an accuracy greater or equal than FW and in all but one case an accuracy greater or equal than MFW. SWAP-2o is most of the time as accurate as MFW. We conclude that the additional running time incurred by the CVM and FW methods is not compensated by a better accuracy in this series of datasets.

Figure 4(b) shows that the model sizes obtained by the different methods are quite similar.

6.3 Experiments on Other Medium-scale and Large-scale Datasets

Results of Figures 5 to 9 show the accuracies, times, speed-ups and model sizes obtained in the other datasets of Table 1. A detailed description of these datasets can be found in [20] or in the public repositories [12] and [22].

To simplify the presentation and further analysis, datasets were separated into two groups: medium-scale and large-scale datasets. A dataset was included in the first group if the largest binary subproblem (see column m_{\max} of

Table 1) to be addressed was lower than 15,000 training examples, and was included in the second group otherwise. According to this criterion, datasets **Letter**, **Pendigits**, **USPS**, **Reuters** and **MNIST** were put together in the first group and datasets **Shuttle**, **IJCNN**, **USPS-Ext**, **KDD-10pc** and **KDD-Full** were included in the second group. Results for dataset **Protein** were presented/analyzed independently because accuracies and training times were significantly different from other results in the medium-scale group. Note again that most of the problems using in this experiment have been already used to compare CVM against other algorithms to train SVMs [58]. Times and model sizes are depicted in a logarithmic scale.

By examining Figure 5 we again observe a slight advantage of CVM in terms of testing accuracy. In addition, we confirm that the accuracy of the SWAP and SWAP-2o methods tends to be the closest to the best observed performance. The FW method is very often the least accurate among the Frank-Wolfe based algorithms. Note that if we compute the difference in accuracy with respect to CVM we always obtain results lower than 2%.

Results in Figure 9 show that the FW, MFW, SWAP and SWAP-2o methods are most of the time faster than CVM. The speed-up achieved by these methods becomes more significant as the size of the training set grows, with peaks of around $100\times$ and $25\times$ on the largest datasets. Differences among the Frank-Wolfe methods depend on the size of the problem. Among the medium-scale datasets all the methods achieve running times of the same order of magnitude. Speed-ups in the large-scale group are clearly more significant, with medians of $27.3\times$, $15.0\times$, $30.7\times$, $29.5\times$ for FW, MFW, SWAP and SWAP-2o respectively.

The advantage of the methods explored in this paper against standard FW routines can be summarized as follows. The FW and MFW methods can sometimes be faster than SWAP and SWAP-2o, but in that case the advantage is very tight. Often, however, our methods can improve on FW and MFW with more significant speed-ups. MFW in particular tends to be significantly outperformed in the cases where the FW works better. In those cases the performance of our methods tends to be competitive or better. On medium-scale problems all the methods are evenly matched in performance, with a slight advantage for SWAP-2o and MFW. In the large-scale group, SWAP and SWAP-2o tend to outperform FW and MFW more significantly.

Results on the **Protein** dataset deserve a particular comment. This is a dataset of around 18,000 examples distributed into 3 classes, which leads to binary subproblems of around 10,000 examples. According to this size, the problem should be included in the group of medium-scale datasets on which we have seen that the Frank-Wolfe algorithms obtain fairly similar and small speed-ups. On the **Protein** problem however the methods obtain peculiar results. The FW method achieves here a speed-up of $20.8\times$ against CVM. However the standard MFW runs here $123.5\times$ faster than the CVM and $5.95\times$ faster than FW. This suggests that in this problem, away steps significantly help the algorithm to find the solution to the SVM problem more quickly. Since our methods tend to be better when away steps work, we should observe important improvements on the CVM using the proposed methods. Indeed, the respective speed-ups for the

SWAP and SWAP-2o methods on this datasets are $157.3\times$ and $358.0\times$. This means that SWAP runs $17.25\times$ faster than FW and $1.27\times$ faster than MFW. SWAP-2o runs $7.58\times$ faster than FW and $2.90\times$ faster than MFW.

Note finally that Figure 7 suggests that there are no significant differences among the sizes of the models built by the different methods.

6.4 Statistical Tests

In this section, we perform some statistical tests to assess the significance of the experimental results reported in this paper. To this end we adopt the guidelines suggested in [15]. We first conduct a multiple test to determine whether the hypothesis that all the algorithms perform equally can be rejected or not. Then, we conduct separate binary tests to compare the performances of each algorithm against each other. For the binary tests we adopt the *Wilcoxon Signed-Ranks Test* method. For the multiple test we use the non-parametric *Friedman Test*. In [15], Demsar recommends these tests as safe alternatives to the classical parametric t-tests to compare classifiers over multiple datasets.

From the multiple test, we conclude that there is indeed a statistically significant difference among the running times and accuracies of all the algorithms (p -values were lower than 0.001 in both cases).

We then conduct a binary test on each pair of algorithms. The main hypothesis of this paper is that the SWAP method outperforms the MFW and FW methods in terms of training time without significant differences in terms of predictive accuracy. In contrast, we claim that no significant differences between the MFW and FW methods are observed in practice (although MFW seems to be slightly more accurate). We have also observed that the SWAP method significantly outperforms CVM, sometimes at the expense of a little test accuracy. Finally, we have observed that the SWAP-2o usually exhibits larger running times than the SWAP method but outperforms the other FW based methods in terms of predictive power. As regards the comparison of the proposed methods, there is no apparent advantage in terms of running time of one against the other. We thus conduct a two-tailed test for the running times but adopt a one-tailed test for testing accuracy. Considering all the observations above, our design for the binary tests is that of Table 2.

In Table 2, we also report the p -values corresponding to each test¹⁰. For reproducibility concerns, p -values were computed using the statistical software R [46]. For the Wilcoxon Signed-Ranks Test, the exact p -values were preferred to the asymptotic ones. The Pratt method to handle ties is employed by default. In the case of the Friedman test, the Iman and Davenport’s correction was adopted, as suggested in [15].

We now point out some of the conclusions which can be obtained from Table 2. At commonly used significance levels (10%, 5%, 1% or lower), the hypothesis

¹⁰In some cases we implement one-sided alternative hypotheses, and in others two-sided tests. If a two-sided test is preferred to a one-sided alternative, it’s enough to double the p -value reported here. Vice versa, if a one-sided test is preferred to a two-sided test, it’s enough to halve the p -value reported here.

Time SWAP vs. FW	p-value	Accuracy SWAP vs. FW	p-value
H_0 : Both equally fast	0.03757	H_0 : Both equally accurate	0.2389
H_1 : SWAP faster		H_1 : Different accuracies	
Time SWAP vs. MFW	p-value	Accuracy SWAP vs. MFW	p-value
H_0 : Both equally fast	1.526e-05	H_0 : Both equally accurate	0.1019
H_1 : SWAP faster		H_1 : Different accuracies	
Time SWAP vs. CVM	p-value	Accuracy SWAP vs. CVM	p-value
H_0 : Both equally fast	5.528e-06	H_0 : Both equally accurate	1.873e-04
H_1 : SWAP faster		H_1 : CVM more accurate	
Time FW vs. MFW	p-value	Accuracy FW vs. MFW	p-value
H_0 : Both equally fast	0.6893	H_0 : Both equally accurate	0.1118
H_1 : Different times		H_1 : Different accuracies	
Time SWAP-2o vs. FW	p-value	Accuracy SWAP-2o vs. FW	p-value
H_0 : Both equally fast	0.3403	H_0 : Both equally accurate	0.01071
H_1 : Different times		H_1 : SWAP-2o more accurate	
Time SWAP-2o vs. MFW	p-value	Accuracy SWAP-2o vs. MFW	p-value
H_0 : Both equally fast	1.087e-04	H_0 : Both equally accurate	0.01634
H_1 : SWAP-2o faster		H_1 : SWAP-2o more accurate	
Time SWAP-2o vs. CVM	p-value	Accuracy SWAP-2o vs. CVM	p-value
H_0 : Both equally fast	4.47e-08	H_0 : Both equally accurate	2.25e-04
H_1 : SWAP-2o faster		H_1 : CVM more accurate	
Time SWAP-2o vs. SWAP	p-value	Accuracy SWAP-2o vs. SWAP	p-value
H_0 : Both equally fast	0.1901	H_0 : Both equally accurate	1.418e-03
H_1 : SWAP faster		H_1 : SWAP-2o more accurate	

Table 2: Null hypotheses, alternative hypotheses and p-values for the binary statistical tests. The conclusion of the test adopting a significance level of 5% is highlighted in blue.

that FW and MFW are equally fast cannot be rejected. Adopting a significance level of 5%, the running times of SWAP method are found to be significantly different from those of all the baseline methods (FW, MFW and CVM), so the null hypothesis is rejected in favor of the alternative hypothesis that the SWAP method is faster. At the same significance level, or better, the hypotheses that the SWAP-2o method is as fast as MFW or CVM are rejected in favor of the conclusion that the SWAP-2o method is faster. Empirical data is however insufficient to reject the hypothesis that the SWAP-2o method is as fast as the FW or the SWAP methods. As regards the testing accuracy, FW, MFW and SWAP are found to be equally as accurate at reasonable significance levels (10%, 5%, 1% or lower). In contrast, the hypothesis that the SWAP-2o method has similar accuracies to FW, MFW and SWAP is rejected in favor of the conclusion that SWAP-2o is more accurate.

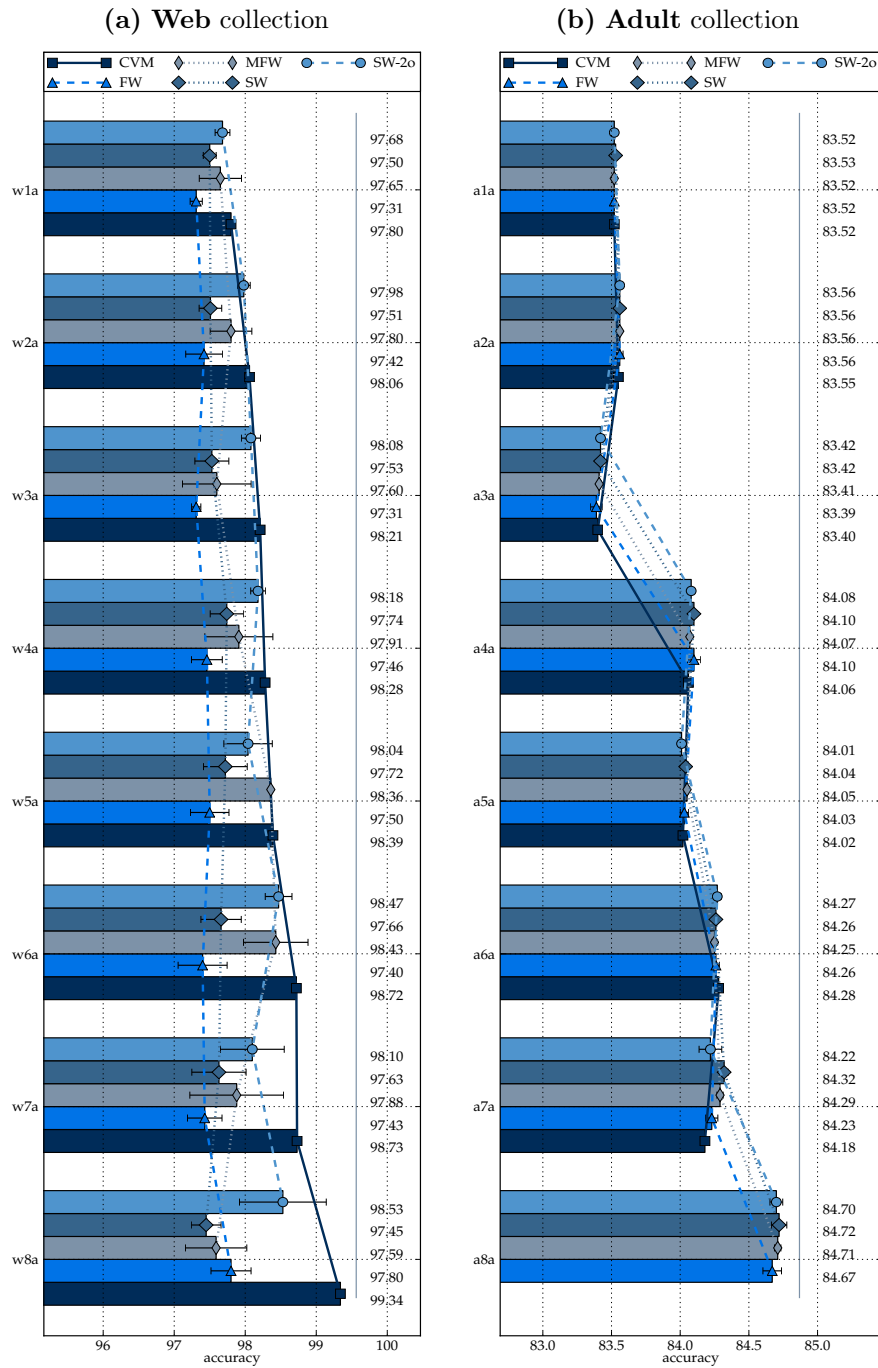
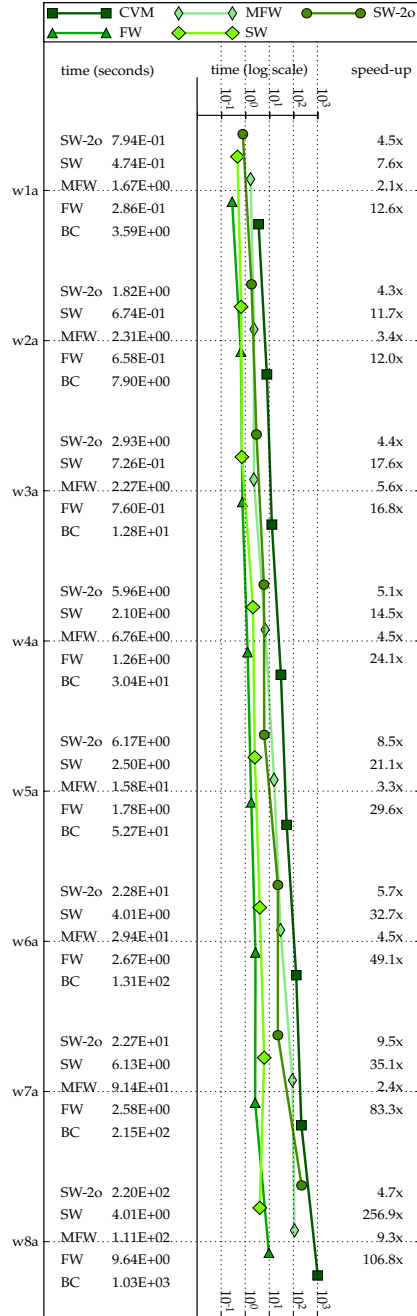


Figure 2: Testing accuracies in the **Web** and **Adult** collections.

(a) Web collection



(b) Adult collection

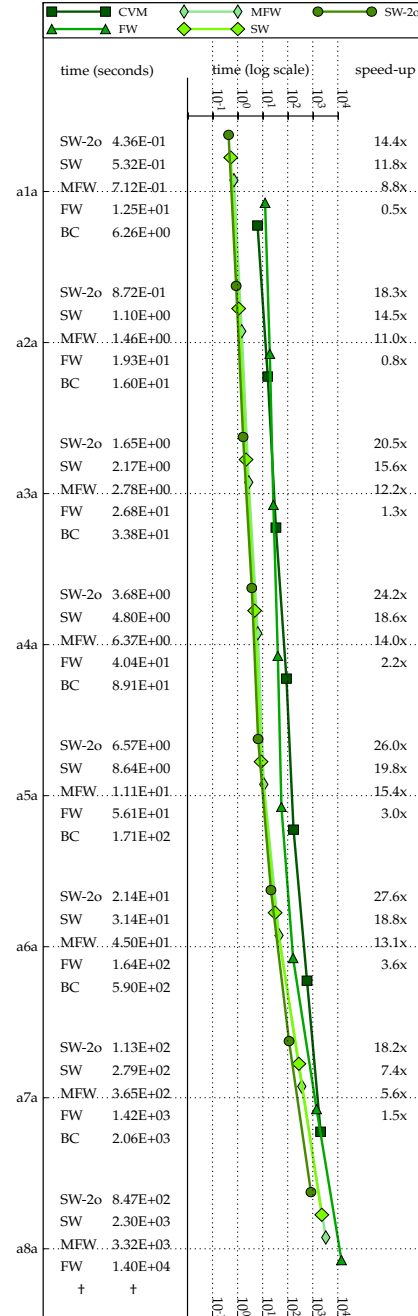
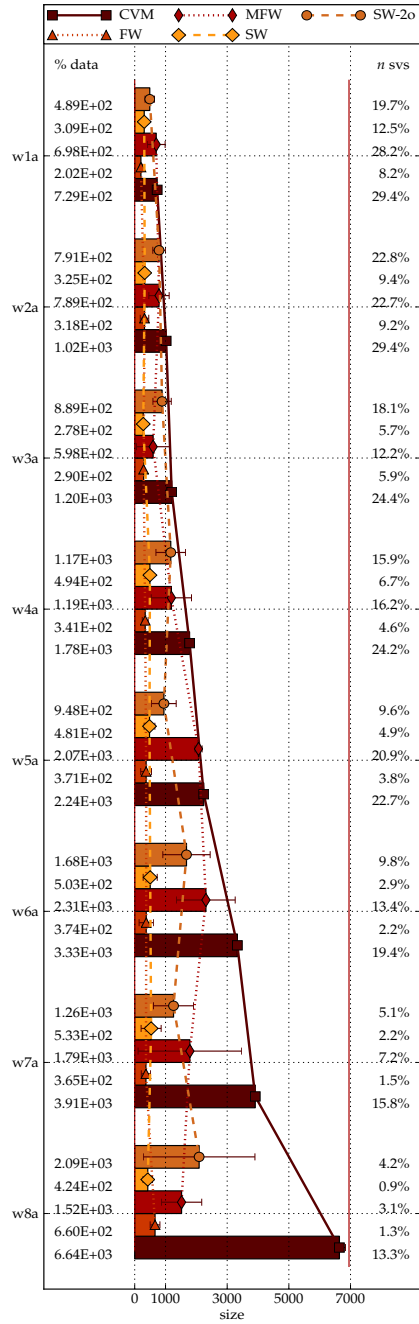


Figure 3: Running times in the **Web** and **Adult** collections. The column on the left shows speed-ups with respect to CVM.

(a) Web collection



(b) Adult collection

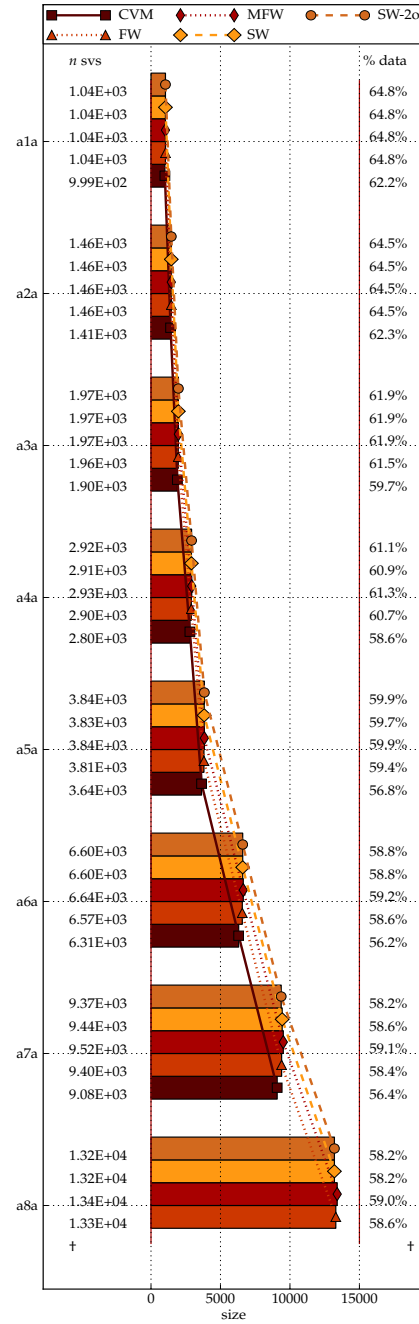
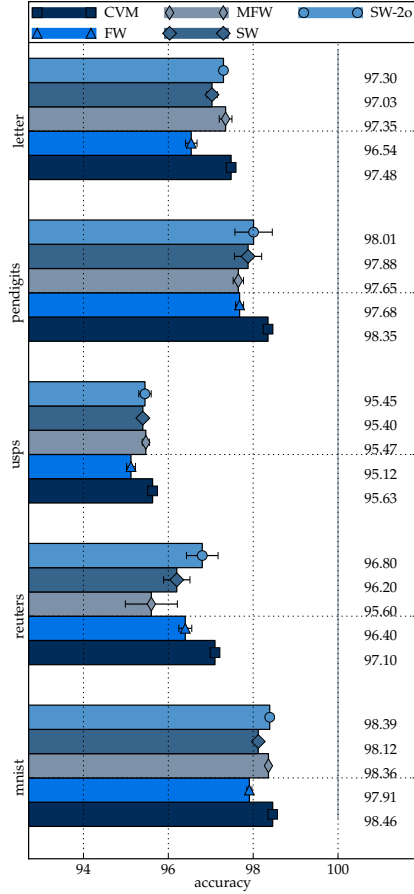


Figure 4: Model sizes in the **Web** and **Adult** collections. The column on the left shows the percentage of the total number of examples.

(a) Medium-scale datasets



(b) Large-scale datasets

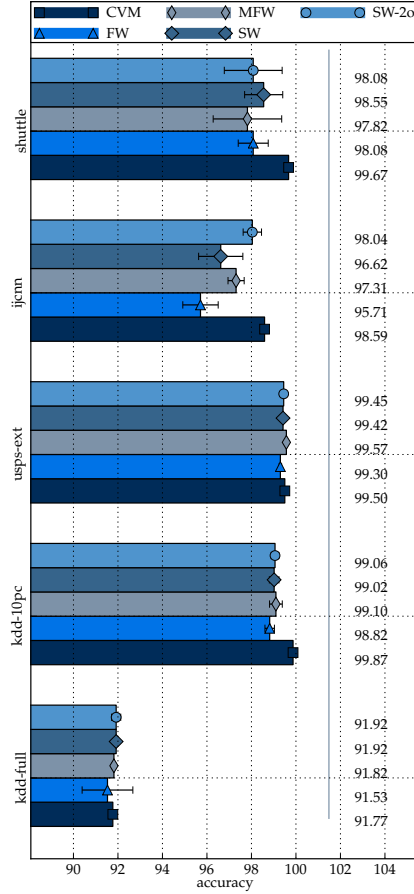


Figure 5: On the left, testing accuracies in the medium-scale datasets: **Letter**, **Pendigits**, **USPS**, **Reuters**, **MNIST**. On the right, testing accuracies in the large dataset collection: **Shuttle**, **IJCNN**, **USPS-Ext**, **KDD-10pc**, **KDD-Full**.

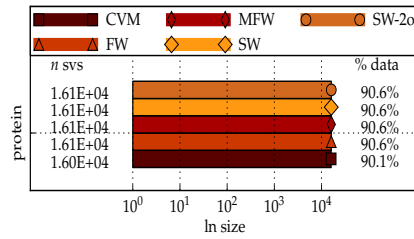
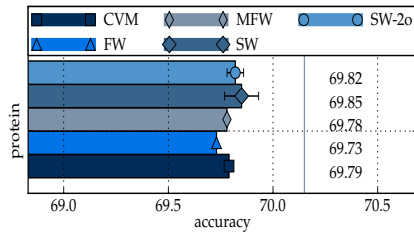
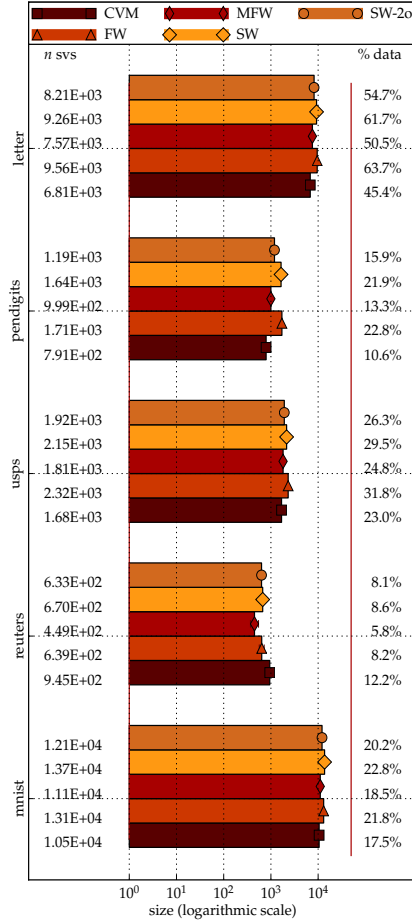


Figure 6: On the left, testing accuracies in the **Protein** dataset. On the right, model sizes in the **Protein** dataset.

(a) Medium-scale datasets



(b) Large-scale datasets

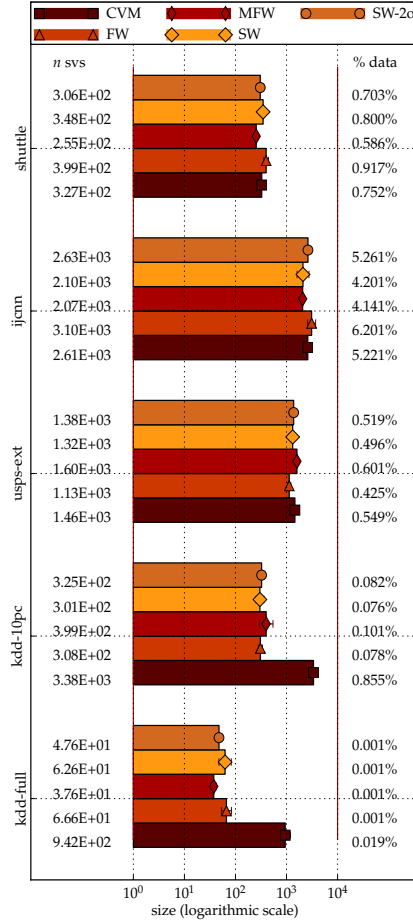


Figure 7: On the left, model sizes in the medium-scale datasets **Letter**, **Pendigits**, **USPS**, **Reuters**, **MNIST**. On the right, model sizes in the large dataset collection: **Shuttle**, **IJCNN**, **USPS-Ext**, **KDD-10pc**, **KDD-Full**.

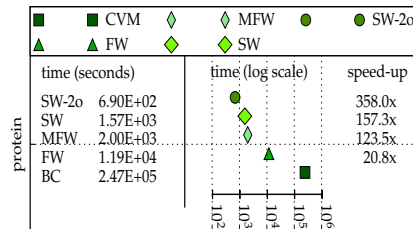
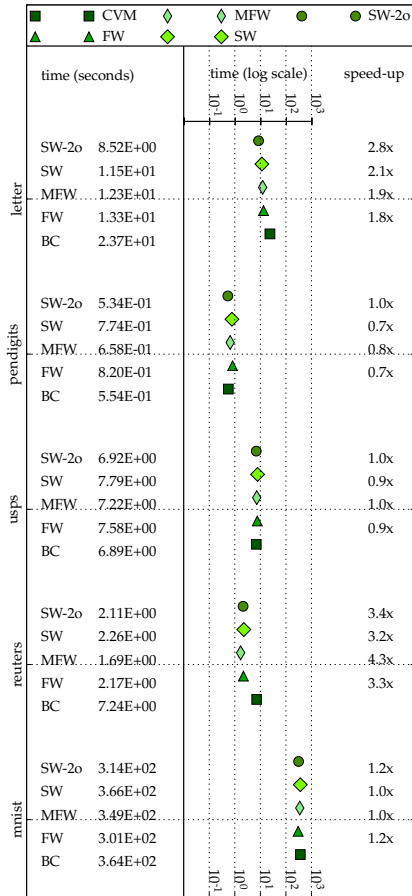


Figure 8: Running times in the **Protein** dataset.

(a) Medium-scale datasets



(b) Large-scale datasets

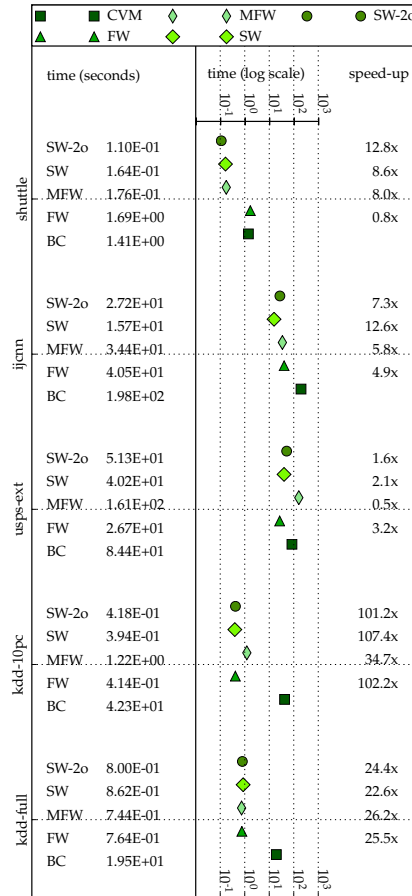


Figure 9: On the left, running times in the medium-scale datasets: **Letter**, **Pendigits**, **USPS**, **Reuters**, **MNIST**. On the right, running times in the large dataset collection: **Shuttle**, **IJCNN**, **USPS-Ext**, **KDD-10pc**, **KDD-Full**.

6.5 Experiments with Non-Normalized Kernels

Solving a classification problem using SVMs requires to select a kernel function. Since the optimal kernel for a given application cannot be specified *a priori*, the capability of a training method to work with any (or the widest possible) family of kernels is an important feature.

In order to illustrate that the proposed methods can obtain effective models even if the kernel does not satisfy the conditions required by CVM, we conduct experiments using the homogenous second order polynomial kernel $k(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j)^2$. Here, parameter γ is estimated as the inverse of the average squared distance among training patterns [58].

Figures 10 and 11 summarize the results obtained in some of the datasets used in this section. We can see that both test accuracies and training times are comparable to those obtained using the Gaussian kernel. It should be noted that the CVM algorithm cannot be used to train an SVM using the kernel selected for this experiment, thus we only incorporate the Frank-Wolfe based methods in the figures. These results demonstrate the capability of our methods to be used with kernels other than those satisfying the normalization condition imposed by CVM.

7 Conclusions

The main contribution of this paper is twofold. On the theoretical side, we proposed a new variant of the FW method for the general problem of maximizing a concave function on the unit simplex, introducing a novel way to perform away steps in the FW method devised to boost its convergence. On the practical side, we demonstrated that our approach is very effective in improving the performance of state-of-the-art SVM learners for large datasets, further expanding on the research about FW methods for Machine Learning problems.

We presented two variants of the procedure, SWAP and SWAP-2o, for which we provided a thorough theoretical analysis. First, we demonstrated that they converge globally. Second, we showed that SWAP and SWAP-2o asymptotically exhibit a linear rate of convergence, which is, as in the case of the MFW method, the main additional property with respect to the standard FW method. Finally, we proved that they achieve a primal-dual gap lower than a given tolerance ε in $\mathcal{O}(1/\varepsilon)$ iterations, independently of m , the dimensionality of the feasible space and the number of examples in SVM problems.

We then carried out an extensive set of performance evaluation experiments for both variants of the algorithm. The obtained results demonstrated that, in contrast to the MFW method, our approach provides a useful and robust alternative to the FW method for training SVMs.

Most often, the proposed methods SWAP and SWAP-2o improved on the performance of MFW. The SWAP method was faster than MFW on all the datasets of the **Adult** collection, the **Web** collection and the **Protein** problem. In the large-scale group of Figure 9(b) SWAP outperformed MFW on 4 (out of

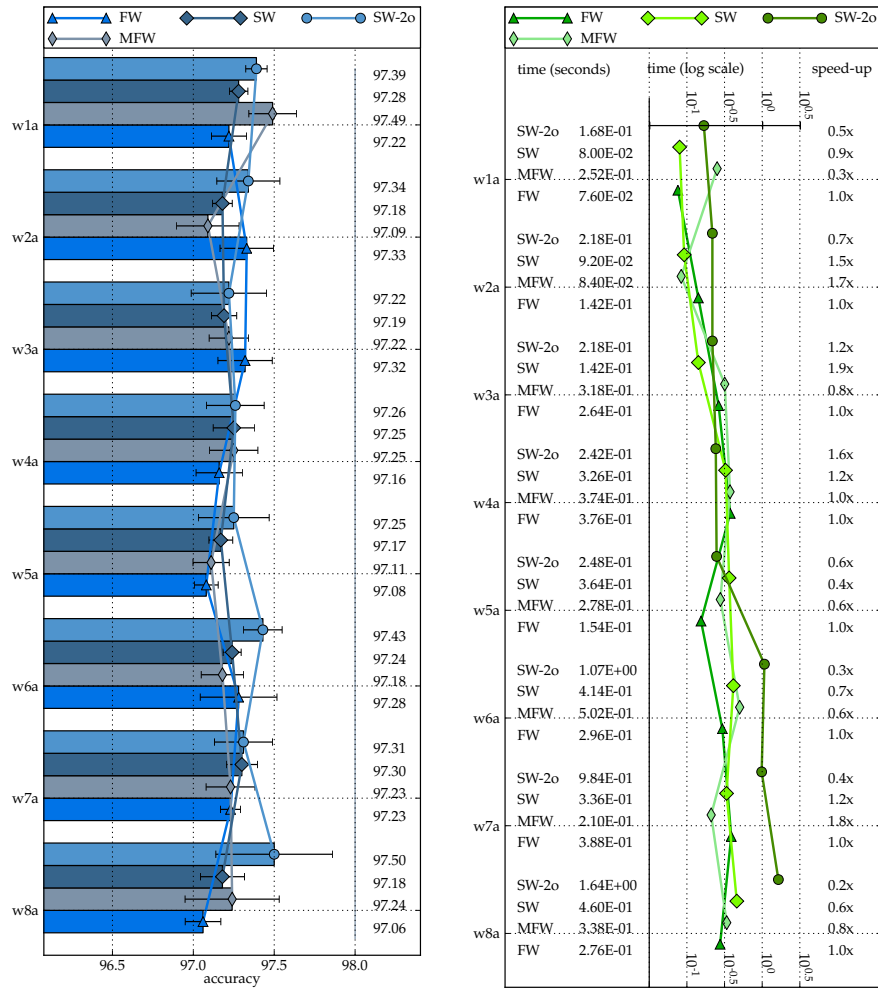


Figure 10: On the left, testing accuracies obtained with the polynomial kernel in the datasets of the **Web** collection, *w1a*, *w2a*, *w3a*, *w4a*, *w5a*, *w6a*, *w7a* and *w8a*. On the right, the corresponding running times.

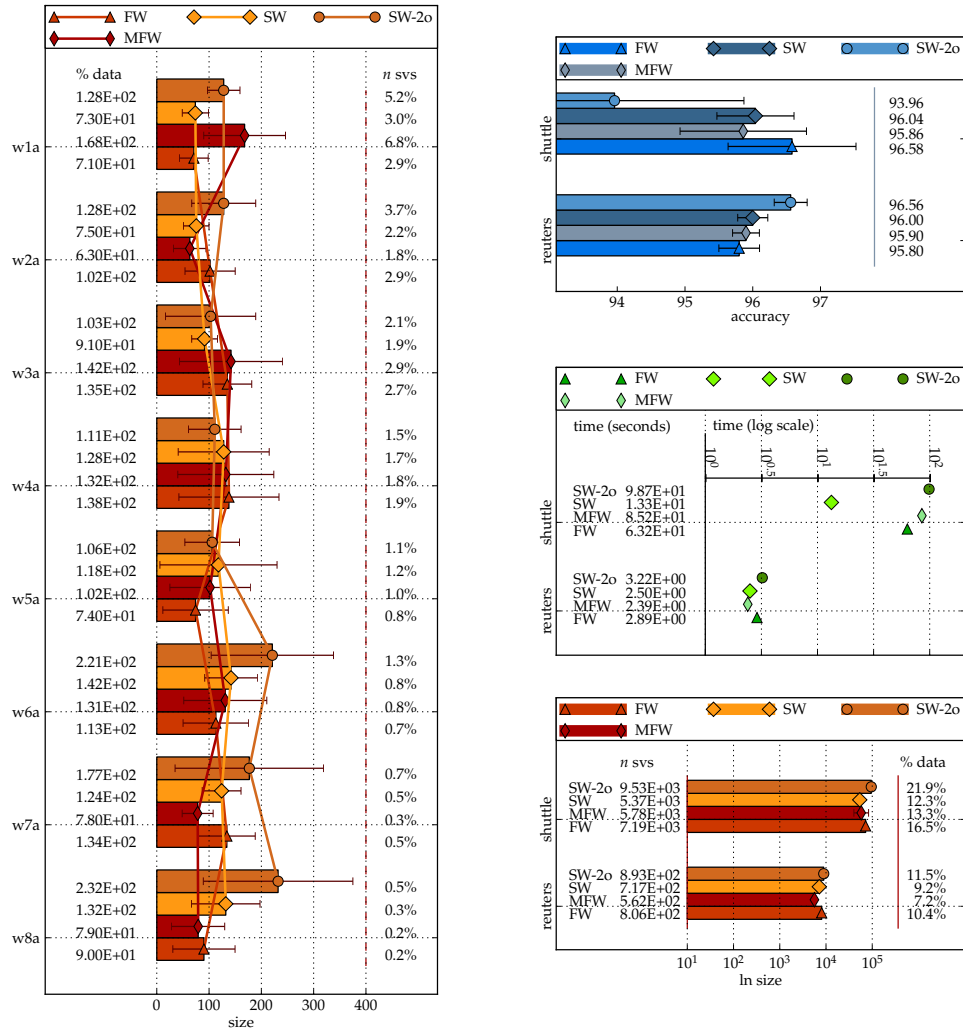


Figure 11: On the left, model sizes obtained with the polynomial kernel in the datasets of the **Web** collection. On the right, testing accuracies, running times and coreset sizes obtained in the *Shuttle* and *Reuters* datasets.

5) datasets. In the medium-scale problems of Figure 9(a) SWAP was slightly slower.

The SWAP-2o method was faster than MFW on all the datasets of the **Adult** collection, 6 (out of 8) datasets in the **Web** collection and 4 (out of 5) datasets in the large-scale group of Figure 9(b). SWAP-2o was also faster in the **Protein** problem and slightly faster on the medium-scale problems of Figure 9(a).

The conclusion that SWAP and SWAP-2o are faster than MFW was found statistically significant at significance levels of 1% or better. Often, the SWAP method improved on MFW by one order of magnitude and sometimes by two orders of magnitude. In addition, in the cases in which MFW was faster, the advantage was less significant than the improvements of our techniques on MFW.

The proposed methods were also faster than the basic FW method several times. For example, SWAP ran in median 15 times faster than FW in the **Adult** collection and SWAP-2o ran 20 times faster. Similar results were observed in the **Shuttle** and **Protein** datasets. We found that the conclusion that SWAP is faster than FW is statistically significant at a critical value of around 4%. In contrast, we were not able to reject the hypothesis that MFW and FW lead to similar training times. Similarly, we cannot conclude that FW and SWAP-2o have different running times.

Another important conclusion of our experimental results arises after an analysis of the cases in which either FW or MFW *fail* in improving running times of CVM by a significant amount.

- In some cases, away steps of MFW significantly speed-up the FW method. Some examples were the **Adult** collection, the **Shuttle** and **Protein** datasets. In those cases, the SWAP method is competitive with or faster than MFW and significantly faster than FW.
- In some other cases, classic away steps fail. MFW achieves in those cases noticeably worse running times. For instance, we observed this behavior in the **Web** collection, the **USPS-Ext** and **KDD-10pc** datasets. In those cases, the SWAP method is clearly faster than MFW. In addition, it is competitive with the fastest algorithm (FW).

We conclude that the SWAP method can be expected to be faster than MFW in those cases in which classic away steps effectively boost the convergence of the FW method but also very competitive against FW when away steps fail. Thus, SWAP is a robust alternative to FW, MFW or CVM. From this point of view, the SWAP-2o method is less appealing. Even if SWAP-2o outperforms more significantly the standard FW when away steps are useful, this technique seems to fail very often in the same cases in which MFW fails. If we knew that away steps were going to be useful for a given problem, SWAP-2o would be the algorithm of choice. However, since we cannot predict that in advance, MFW and SWAP-2o are less reliable in practice.

Finally, our experiments have demonstrated that the improvements in running time that we obtain on FW or MFW do not come at the expense at the

expense of testing accuracy. Most of the time SWAP is slightly more accurate than FW and as accurate as MFW.

A Technical Results

Here we report statements and proofs of a number of technical results, which are used in the theoretical analysis of Section 4.

A.1 Perturbation Analysis

We follow the analysis presented in [1], which is in turn based on the perturbation method of Robinson [48]. Consider the following *perturbed* variant of (1),

$$\begin{aligned} & \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad w(\boldsymbol{\alpha}) = g(\boldsymbol{\alpha}) - \mathbf{z}^T \boldsymbol{\alpha} \\ & \text{subject to} \quad \mathbf{1}^T \boldsymbol{\alpha} = 1, \quad \boldsymbol{\alpha} \geq 0, \end{aligned} \quad (54)$$

where $\mathbf{z} \in \mathbb{R}^m$ is perturbation vector.

Now, suppose we have a Δ_* -approximate solution $\boldsymbol{\alpha}_* \in \mathbb{R}^m$. We are aimed to show that $\boldsymbol{\alpha}_*$ is the solution of a perturbed problem with a certain \mathbf{z} . We define $\mathbf{z} = \mathbf{z}(\boldsymbol{\alpha}_*, \Delta_*)$ by

$$z_i = \begin{cases} \Delta_* & \text{if } \alpha_{*i} = 0, \\ \nabla g(\boldsymbol{\alpha}_*)_i - \boldsymbol{\alpha}_*^T \nabla g(\boldsymbol{\alpha}_*) & \text{if } \alpha_{*i} > 0. \end{cases} \quad (55)$$

Note first that if $\alpha_{*i} \neq 0$

$$\begin{aligned} z_i &= \nabla g(\boldsymbol{\alpha}_*)_i - \boldsymbol{\alpha}_*^T \nabla g(\boldsymbol{\alpha}_*) \geq -\Delta_* \\ z_i &= \nabla g(\boldsymbol{\alpha}_*)_i - \boldsymbol{\alpha}_*^T \nabla g(\boldsymbol{\alpha}_*) \leq +\Delta_*, \end{aligned} \quad (56)$$

because $\boldsymbol{\alpha}_*$ is a Δ_* -approximate solution. If $\alpha_{*i} = 0$, $z_i = \Delta_*$ by construction. Then,

$$\|\mathbf{z}\|^2 = \sum_i |z_i|^2 \leq m\Delta_*^2. \quad (57)$$

Note in addition that

$$\boldsymbol{\alpha}_*^T \mathbf{z} = \sum_i \alpha_{*i} z_i = \sum_{i:\alpha_{*i} \neq 0} \alpha_{*i} z_i = \boldsymbol{\alpha}_*^T \nabla g(\boldsymbol{\alpha}_*) - (\boldsymbol{\alpha}_*^T \nabla g(\boldsymbol{\alpha}_*)) \boldsymbol{\alpha}_*^T \mathbf{1} = 0, \quad (58)$$

because $\boldsymbol{\alpha}_*$ is feasible for (1). Note finally that

$$\nabla w(\boldsymbol{\alpha}) = \nabla g(\boldsymbol{\alpha}) - \mathbf{z}. \quad (59)$$

Thus, from Eqn. (58) we obtain that the following stationarity condition is fulfilled

$$\boldsymbol{\alpha}_*^T \nabla w(\boldsymbol{\alpha}_*) = \boldsymbol{\alpha}_*^T \nabla g(\boldsymbol{\alpha}_*) - \boldsymbol{\alpha}_*^T \mathbf{z} = \boldsymbol{\alpha}_*^T \nabla g(\boldsymbol{\alpha}_*). \quad (60)$$

The following lemma follows easily from the previous remarks.

Lemma 3. *If α_* is a Δ_* -approximate solution, then α_* is optimal for problem (54) with $\mathbf{z} = \mathbf{z}(\alpha_*, \Delta_*)$ as defined in Eqn. (55).*

Proof. It follows from the concavity of problem (54) and the remarks above. See [1], Lemma 3.3 for details. \square

The next lemma is the basis of the analysis of the rate of convergence for the modified Frank-Wolfe methods.

Lemma 4. *Let α^* be the solution of problem (1) and α_* a Δ_* -approximate solution. Then,*

$$g(\alpha^*) - g(\alpha_*) \leq \|\mathbf{z}\| \|\alpha^* - \alpha_*\| \leq \sqrt{m} \Delta_* \|\alpha^* - \alpha_*\| . \quad (61)$$

Proof. The vector α^* is feasible for the perturbed problem (54) with $\mathbf{z} = \mathbf{z}(\alpha_*, \Delta_*)$. Since α_* is optimal for this problem, we have $g(\alpha^*) - \mathbf{z}^T \alpha^* \leq g(\alpha_*) - \mathbf{z}^T \alpha_*$. This demonstrates the first inequality. The other follows from Eqn. (57). \square

From here, the following lemma follows easily.

Lemma 5. *Suppose condition B2 holds. Let α^* be the solution of problem (1) and α_* a Δ_* -approximate solution, where Δ_* is sufficiently small. Then,*

$$g(\alpha^*) - g(\alpha_*) \leq Nm \Delta_*^2 \quad (62)$$

for some Lipschitz constant N .

Proof. See [1] to see how from the Robinson condition it follows that there exists a Lipschitz constant N such that, for sufficiently small Δ_* , $\|\alpha^* - \alpha_*\| \leq N \|\mathbf{z}\| \leq N \sqrt{m} \Delta_*$. Combining this result with the previous lemma yields the result. \square

Now, since g is twice differentiable, the Taylor expansion for $g(\alpha_k + \lambda \mathbf{d})$ as a function of λ is

$$g(\alpha_k + \lambda \mathbf{d}) = g(\alpha_k) + \lambda \nabla g(\alpha_k)^T \mathbf{d} + \frac{1}{2} \lambda^2 \mathbf{d}^T \nabla^2 g(\tilde{\alpha}) \mathbf{d} , \quad (63)$$

where $\tilde{\alpha}$ is some point on the line between $\alpha_k + \lambda \mathbf{d}$ and α_k . Since g is concave, the Hessian matrix $\nabla^2 g(\tilde{\alpha})$ is negative semi-definite, so the last term is always non-positive. To obtain a bound for $g(\alpha_k + \lambda \mathbf{d}) - g(\alpha_k)$, we need a bound L on the norm of $\nabla^2 g(\alpha)$ over the simplex. We can set L to the largest absolute value of an eigenvalue of this matrix. We therefore obtain the following bound:

$$g(\alpha_k + \lambda \mathbf{d}) - g(\alpha_k) \geq \lambda \nabla g(\alpha_k)^T \mathbf{d} - \frac{1}{2} \lambda^2 L \|\mathbf{d}\|^2 . \quad (64)$$

We now exploit the previous bound to analyze the improvement in the objective function δ_{fw} after a standard FW step, and the improvement δ_{swap} after a SWAP step in Algorithm 4.

A.2 Objective Function Improvement after Frank-Wolfe Steps

Under hypothesis B1, we now derive a lower bound for the improvement in the objective function $g(\boldsymbol{\alpha}_k + \lambda d_k) - g(\boldsymbol{\alpha}_k)$ in the case a FW step is performed.

For a FW step we have $\mathbf{d}_k = \mathbf{e}_{i^*} - \boldsymbol{\alpha}_k$. Thus,

$$\begin{aligned} \delta_{\text{fw}} &= g(\boldsymbol{\alpha}_k + \lambda(\mathbf{e}_{i^*} - \boldsymbol{\alpha}_k)) - g(\boldsymbol{\alpha}_k) \\ &\geq \lambda \left(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k \right) - \frac{L}{2} \lambda^2 \|\mathbf{e}_{i^*} - \boldsymbol{\alpha}_k\|^2 . \end{aligned} \quad (65)$$

But both \mathbf{e}_{i^*} and $\boldsymbol{\alpha}_k$ lie in the simplex. Hence $\|\mathbf{e}_{i^*} - \boldsymbol{\alpha}_k\|^2 \leq 2$. This leads to

$$\begin{aligned} \delta_{\text{fw}} &\geq g(\boldsymbol{\alpha}_k + \lambda(\mathbf{e}_{i^*} - \boldsymbol{\alpha}_k)) - g(\boldsymbol{\alpha}_k) \\ &\geq \lambda \left(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k \right) - L\lambda^2 . \end{aligned} \quad (66)$$

The maximum of the right-hand side is obtained for

$$\lambda_{\text{fw}}^* = \frac{\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k}{2L} . \quad (67)$$

If $\lambda_{\text{fw}}^* \leq 1$, the improvement in the objective function after an iteration marked as a FW step in Algorithm 4 is bounded by

$$\delta_{\text{fw}} \geq \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k)^2}{4L} , \quad (68)$$

and by reordering we obtain

$$\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k \leq 2\sqrt{L\delta_{\text{fw}}} \leq 2\sqrt{L\delta_k} , \quad (69)$$

where the latter inequality follows from the definition of $\delta_k = \max(\delta_{\text{fw}}, \delta_{\text{swap}})$. Now, if $\lambda_{\text{fw}}^* > 1$, we cannot use this step-size. In that case we use the step-size $\lambda = 1$. But $\lambda_{\text{fw}}^* > 1$ implies

$$\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k \geq 2L . \quad (70)$$

Thus, using Eqn. (66) with $\lambda = 1$ and exploiting the inequality above, the improvement in the objective function for a FW step can be bounded in this case as

$$\delta_{\text{fw}} \geq \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k)}{2} , \quad (71)$$

which leads to

$$\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k \leq 2\delta_{\text{fw}} \leq 2\delta_k . \quad (72)$$

In any case, we have the following bound for the improvement of the objective function:

$$\delta_{\text{fw}} \geq \min \left(\frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k)^2}{4L} , \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k)}{2} \right) , \quad (73)$$

which guarantees that for any k

$$\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)^T \boldsymbol{\alpha}_k \leq \max\left(2\sqrt{L\delta_{\text{fw}}}, 2\delta_{\text{fw}}\right) \leq \max\left(2\sqrt{L\delta_k}, 2\delta_k\right). \quad (74)$$

Now, since $\nabla g(\boldsymbol{\alpha}_k)_i \leq \nabla g(\boldsymbol{\alpha}_k)_{i^*} \forall i$, the following inequality is guaranteed at each iteration of Algorithm 4 for any i :

$$\nabla g(\boldsymbol{\alpha}_k)_i - \boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) \leq \max\left(2\sqrt{L\delta_k}, 2\delta_k\right). \quad (75)$$

A.3 Objective Function Improvement after SWAP Steps

We now bound the improvement obtained by SWAP steps. In this case, $\mathbf{d} = \mathbf{e}_{i^*} - \mathbf{e}_{j^*}$. Thus,

$$\begin{aligned} \delta_{\text{swap}} &= g(\boldsymbol{\alpha}_k + \lambda(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})) - g(\boldsymbol{\alpha}_k) \\ &\geq \lambda(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*}) - \frac{L}{2}\lambda^2 \|\mathbf{e}_{i^*} - \mathbf{e}_{j^*}\|^2. \end{aligned} \quad (76)$$

But $\|\mathbf{e}_{i^*} - \mathbf{e}_{j^*}\|^2 = 2$. Thus,

$$\delta_{\text{swap}} \geq \lambda(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*}) - \lambda^2 L. \quad (77)$$

The maximum of the right-hand side is obtained for

$$\lambda_{\text{swap}}^* = \frac{\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*}}{2L}. \quad (78)$$

If $\lambda_{\text{swap}}^* \leq 1$, the improvement in the objective function for an unconstrained SWAP step, that is, an iteration marked as SWAP-add in Algorithm 4, is bounded as

$$\delta_{\text{swap}} \geq \frac{(\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*})^2}{4L}. \quad (79)$$

Note now that $\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) \leq \nabla g(\boldsymbol{\alpha}_k)_{i^*}$ because $\nabla g(\boldsymbol{\alpha}_k)_{i^*} = \max_i \nabla g(\boldsymbol{\alpha}_k)_i$ and $\boldsymbol{\alpha}_k^T \mathbf{1} = 1$. This observation leads to

$$\delta_{\text{swap}} \geq \frac{(\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_{j^*})^2}{4L}. \quad (80)$$

By reordering, we obtain the following inequality,

$$\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_{j^*}^T \leq 2\sqrt{L\delta_{\text{swap}}} \leq 2\sqrt{L\delta_k}, \quad (81)$$

where we have used the definition of δ_k . Now, if $\lambda_{\text{swap}}^* > 1$, we cannot use this step-size. In that case we use the step-size $\lambda = 1$. Recall that we are supposing to be performing a SWAP-add step. In a way analogous to the Frank-Wolfe case, $\lambda_{\text{swap}}^* > 1$ implies

$$\nabla g(\boldsymbol{\alpha}_k)_{i^*} - \nabla g(\boldsymbol{\alpha}_k)_{j^*} \geq 2L. \quad (82)$$

Thus, the improvement in the objective function is bounded in this case by

$$\delta_{\text{swap}} \geq \frac{\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_{j^*}}{2}, \quad (83)$$

which leads to

$$\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_{j^*} \leq 2\delta_{\text{swap}} \leq 2\delta_k . \quad (84)$$

In any case, we have the following bounds for the SWAP case

$$\delta_{\text{swap}} \geq \min \left(\frac{(\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_{j^*})^2}{4L}, \frac{(\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_{j^*})}{2} \right), \quad (85)$$

which leads to

$$\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_{j^*} \leq \max \left(2\sqrt{L\delta_{\text{swap}}}, 2\delta_{\text{swap}} \right) \leq \max \left(2\sqrt{L\delta_k}, 2\delta_k \right). \quad (86)$$

Note now that the definition of j^* can be rearranged as

$$j^* \in \underset{j \in \mathcal{I}_k}{\operatorname{argmin}} \nabla g(\boldsymbol{\alpha}_k)_j - \boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) = \underset{j \in \mathcal{I}_k}{\operatorname{argmax}} \boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_j . \quad (87)$$

Thus, we obtain that the following inequality is guaranteed at each iteration of Algorithm 4 $\forall i$ such that $\alpha_{k,i} > 0$:

$$\boldsymbol{\alpha}_k^T \nabla g(\boldsymbol{\alpha}_k) - \nabla g(\boldsymbol{\alpha}_k)_i \leq \max \left(2\sqrt{L\delta_k}, 2\delta_k \right). \quad (88)$$

Remark 2. After a swap-drop step in Algorithm 4 we cannot bound the improvement in the objective function, because the clipped value of the step-size λ_{swap^*} may be arbitrarily small. However, it is not hard to show that the objective function value does not decrease.

References

- [1] S. Damla Ahipasaoglu, Sun Peng, and Michael Todd. Linear convergence of a modified Frank-Wolfe algorithm for computing minimum volume enclosing ellipsoids. *Optimization Methods and Software*, 23(1):5–19, 2008.
- [2] Héctor Allende, Emanuele Frandi, Ricardo Nanculef, and Claudio Sartori. Novel Frank-Wolfe methods for SVM learning. *Technical report*, <http://arxiv.org/abs/1304.1014>, 2013.
- [3] Gükhan Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander Smola, Ben Taskar, and S. V. N. Vishwanathan, editors. *Predicting Structured Data (Neural Information Processing)*. The MIT Press, 2007.
- [4] Amir Beck and Marc Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.
- [5] Kristin P. Bennett and Erin J. Bredensteiner. Geometry in learning. In *Geometry at Work*, 1997.
- [6] Kristin P. Bennett and Erin J. Bredensteiner. Duality and geometry in SVM classifiers. In *In Proc. 17th International Conf. on Machine Learning*, pages 57–64. Morgan Kaufmann, 2000.

- [7] Léon Bottou. Stochastic learning. In Olivier Bousquet and Ulrike von Luxburg, editors, *Advanced Lectures on Machine Learning*, Lecture Notes in Artificial Intelligence, LNAI 3176, pages 146–168. Springer Verlag, 2004.
- [8] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *NIPS*, 2007.
- [9] Mihai Bădoiu and Kenneth Clarkson. Smaller core-sets for balls. In *Proceedings of the SODA '03*, pages 801–802. SIAM, 2003.
- [10] Mihai Bădoiu and Kenneth Clarkson. Optimal core-sets for balls. *Computational Geometry*, 40(1):14–22, 2008.
- [11] Christopher J.C. Burges and David J. Crisp. A geometric interpretation of nu-SVM classifiers. *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, pages 244–250, 2000.
- [12] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2011.
- [13] Kenneth Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. In *Proceedings of SODA '08*, pages 922–931. SIAM, 2008.
- [14] Kenneth Clarkson, Elad Hazan, and David Woodruff. Sublinear optimization for machine learning. *Journal of the ACM*, 59(5), 2012.
- [15] Janez Demsar. Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.
- [16] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 6:1889–1918, 2005.
- [17] Anthony V. Fiacco and Garth P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. SIAM, 1968 (reprinted 1990).
- [18] Shai Fine and Katya Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2002.
- [19] Emanuele Frandi, Maria Grazia Gasparo, Stefano Lodi, Ricardo Ñanculef, and Claudio Sartori. A new algorithm for training SVMs using approximate minimal enclosing balls. In *Proceedings of the 15th Iberoamerican Congress on Pattern Recognition, Lecture Notes in Computer Science*, pages 87–95. Springer, 2010.
- [20] Emanuele Frandi, Maria Grazia Gasparo, Stefano Lodi, Ricardo Ñanculef, and Claudio Sartori. Training support vector machines using Frank-Wolfe methods. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(3), 2011.

- [21] Emanuele Frandi, Maria Grazia Gasparo, Ricardo Ñanculef, and Alessandra Papini. Solution of classification problems via computational geometry methods. *Recent Advances in Nonlinear Optimization and Equilibrium Problems: a Tribute to Marco D’Apuzzo, Quaderni di Matematica*, 27:201–226, 2012.
- [22] Andrew Frank and Arthur Asuncion. *The UCI KDD Archive*. <http://kdd.ics.uci.edu>, 2010.
- [23] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 1:95–110, 1956.
- [24] Thilo-Thomas Friess, Nello Cristianini, and Colin Campbell. The kernel-adatron algorithm: A fast and simple learning procedure for support vector machines. In *Proceedings of the Fifteenth International Conference on Machine Learning, ICML ’98*, pages 188–196, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.
- [25] Dan Garber and Elad Hazan. A polynomial time conditional gradient algorithm with applications to online and stochastic optimization. *CoRR*, abs/1301.4666, 2013.
- [26] Elmer Gilbert. An iterative procedure for computing the minimum of a quadratic form on a convex set. *SIAM Journal on Control*, 4(1):61–80, 1966.
- [27] Jacques Guélat and Patrice Marcotte. Some comments on Wolfe’s “away step”. *Mathematical Programming*, 35:110–119, 1986.
- [28] Bernd Gärtner and Martin Jaggi. Coresets for polytope distance. In John Hershberger and Efi Fogel, editors, *Symposium on Computational Geometry*, pages 33–42. ACM, 2009.
- [29] Elad Hazan and Satyen Kale. Projection-free online learning. In *International Conference on Machine Learning*, 2012.
- [30] Thomas Hofmann, Bernhard Schölkopf, and Alexander Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.
- [31] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and Sellanickam Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the 25th international conference on Machine learning*, pages 408–415. ACM, 2008.
- [32] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.
- [33] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning (to appear)*, 2013.

- [34] Thorsten Joachims. Making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*, pages 169–184. MIT Press, 1999.
- [35] Thorsten Joachims. *SVM-light Support Vector Machine*, 2011.
- [36] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R.K. Murthy. A fast iterative nearest point algorithm for support vector machine classifier design. *IEEE Transactions Neural Networks*, 11(1):124–136, January 2000.
- [37] Sathiya Keerthi and Elmer Gilbert. Convergence of a generalized SMO algorithm for SVM classifier design. *Machine Learning*, 46(1-3):351–360, 2002.
- [38] Piyush Kumar and Alper Yildirim. A linearly convergent linear-time first-order algorithm for support vector classification with a core set result. *INFORMS Journal on Computing*, 23(3):377–391, 2011.
- [39] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *International Conference on Machine Learning (to appear)*, 2013.
- [40] Jorge López, Álvaro Barbero, and José R. Dorronsoro. On the equivalence of the SMO and MDM algorithms for SVM training. In *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I, ECML PKDD '08*, pages 288–300, Berlin, Heidelberg, 2008. Springer-Verlag.
- [41] Jorge Lopez and José R Dorronsoro. The convergence rate of the MDM algorithm. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–7. IEEE, 2012.
- [42] B.F. Mitchell, V.F. Dem'yanov, and V.N. Malozemov. Finding the point of a polyhedron closest to the origin. *SIAM Journal on Control*, 12(1):19–26, 1974.
- [43] Jorge Nocedal and Stephen Wright. *Numerical optimization (2nd edition)*. Springer, 2006.
- [44] Hua Ouyang and Alexander Gray. Fast stochastic Frank-Wolfe algorithms for nonlinear SVMs. In *SDM*, pages 245–256, 2010.
- [45] John Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: support vector learning*, pages 185–208. MIT Press, 1999.
- [46] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

- [47] Piyush Rai, Hal Daumé, and Suresh Venkatasubramanian. Streamed learning: one-pass SVMs. In *IJCAI'09: Proceedings of the 21st international joint conference on Artificial intelligence*, pages 1211–1216. Morgan Kaufmann Publishers, 2009.
- [48] Stephen Robinson. Generalized equations and their solutions, part II: Applications to nonlinear programming. In *Optimality and Stability in Mathematical Programming*, volume 19 of *Mathematical Programming Studies*, pages 200–221. Springer Berlin Heidelberg, 1982.
- [49] Katya Scheinberg. An efficient implementation of an active set method for SVMs. *Journal of Machine Learning Research*, 7:2237–2257, 2006.
- [50] Bernard Schölkopf, Christopher Burges, and Alexander Smola, editors. *Advances in kernel methods: support vector learning*. MIT Press, 1999.
- [51] Bernard Schölkopf and Alexander Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.
- [52] Bernhard Schölkopf and Alexander J. Smola. Sparse greedy matrix approximation for machine learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 911–918, 2001.
- [53] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.
- [54] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011.
- [55] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599, 2013.
- [56] Ingo Steinwart. Sparseness of support vector machines. *The Journal of Machine Learning Research*, 4:1071–1105, 2003.
- [57] Ivor Tsang, Andras Kocsor, and James Kwok. *LibCVM Toolkit*. www.c2i.ntu.edu.sg/ivor/cvm.html, 2011.
- [58] Ivor Tsang, James Kwok, and Pak-Ming Cheung. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.
- [59] Ivor Tsang, James Kwok, and Jacek Zurada. Generalized core vector machines. *IEEE Transactions on Neural Networks*, 17(5):1126–1140, 2006.

- [60] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.
- [61] Di Wang, Bo Zhang, Peng Zhang, and Hong Qiao. An online core vector machine with adaptive MEB adjustment. *Pattern Recognition*, 43:3468–3482, 2010.
- [62] Philip Wolfe. Convergence theory in nonlinear programming. In J. Abadie, editor, *Integer and Nonlinear Programming*, pages 1–36. North-Holland, Amsterdam, 1970.
- [63] Kristian Woodsend and Jacek Gondzio. Exploiting separability in large scale linear support vector machine training. *Computational Optimization and Applications*, 29:241–269, 2011.
- [64] Emre Alper Yildirim. Two algorithms for the minimum enclosing ball problem. *SIAM Journal on Optimization*, 19(3):1368–1391, 2008.
- [65] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. Recent advances of large-scale linear classification. *Proceedings of the IEEE*, 100(9):2584–2603, 2012.
- [66] Tong Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682 – 691, mar 2003.