

Data-Dependent Complexity of First-Order Methods for Support Vector Machines

Matthew Hough

MHOUGH@UWATERLOO.CA

*Department of Combinatorics and Optimization
University of Waterloo
Waterloo, ON N2L 3G1, Canada*

Stephen Vavasis

VAVASIS@UWATERLOO.CA

*Department of Combinatorics and Optimization
University of Waterloo
Waterloo, ON N2L 3G1, Canada*

Abstract

Large-scale problems in data science are often modeled with optimization and solved with first-order methods that may converge at a sublinear rate. It is of interest to terminate the optimization algorithm as soon as the underlying data science task is accomplished. We analyze the complexity of first-order methods for support vector machines (SVMs) in terms of the accuracy required to obtain a useful classifier rather than solve the underlying optimization problem. For a perturbed SVM formulation, we derive a stopping condition certifying individual points as correctly classified without data assumptions. Under a geometric model consisting of two clusters, one per class, together with noisy observations, certifying one point guarantees that the current classifier separates the clusters given a modest number of samples. We additionally derive a data-dependent accuracy threshold guaranteeing this separation. Computable bounds on distance to the optimum allow evaluation of our stopping condition for FISTA and coordinate descent. For FISTA, we derive data-dependent iteration complexity bounds for certification and separation. We demonstrate that LIBLINEAR’s stopping condition can be overly conservative or terminate prematurely with a poor classifier. Experiments demonstrate robust termination and show that a simple FISTA implementation using our condition is competitive with widely used SVM solvers.

Keywords: Support vector machines, data-dependent complexity, FISTA, stopping criteria, first-order methods

1 Introduction

Binary classification is a central task in machine learning with numerous applications. Given data separated into two classes, the task is to construct a decision boundary called a *classifier* that separates the classes completely, or with minimal misclassifications. Since many binary classification models, including the support vector machine (SVM), can be modeled as convex optimization problems, first-order methods are a popular choice for training these models, especially in large-scale settings. Typically, algorithms for binary classification are guaranteed to yield a classifier under the assumption that numerical convergence of the underlying optimization problem is achieved. In practice, however, high-accuracy optimization is often unnecessary to correctly classify most points: a hyperplane that correctly classifies “well-separated” or “easy to classify” points may emerge after only a small number of

iterations. One would expect that the ability to obtain such an approximate separator in few iterations depends on the *classifiability* of the data, i.e. the extent to which the data contain points that are robustly separable from the opposite class. In such cases, the computational effort required to find a useful classifier is not solely determined by worst-case complexity bounds of the optimization problem, but can instead be characterized in terms of *data-dependent complexity*, where the number of iterations required to obtain a classifier scales with geometric properties of the dataset rather than purely algorithmic constants.

The work herein investigates data-dependent complexity for the soft-margin support vector machine (SVM) introduced in Cortes and Vapnik (1995). Here, a perfect separator need not exist, but a large subset of the data may nevertheless be robustly separable. Our goal is to determine how accurately the underlying SVM optimization problem must be solved before such points can be identified and a hyperplane separating them can be recovered.

We focus primarily on a formulation of the SVM that we call H-SVM (homogeneous SVM), which is obtained by replacing each data point $\mathbf{h} \in \mathbb{R}^d$ with the augmented point $(\mathbf{h}, 1) \in \mathbb{R}^{d+1}$ and regularizing the intercept together with the normal vector. This treatment of the intercept is standard in large-scale linear classification and is used, for example, by LIBLINEAR (Fan et al., 2008). The corresponding dual problem has only a box constraint, making projected gradient steps particularly efficient. However, this dual problem may have multiple optimal solutions, which complicates the use of dual solutions for identifying well-classified points. We therefore introduce a strongly concave perturbed formulation of the H-SVM dual that has a unique optimal solution and, under our data model and parameter choices, has condition number $\kappa = L/\mu < 641$, allowing FISTA to converge at a linear rate independent of the number of training observations.

Our analysis establishes computable error bounds on the distance from the current dual iterate to the optimal dual solution for FISTA and coordinate descent. These bounds yield a practical stopping condition which certifies individual training points as correctly classified by both the current classifier and the optimal classifier, without requiring assumptions on the data. Under a geometric data model in which each class contains a cluster of well-classified points, while a bounded fraction of the observations may be noisy, we prove that certifying a single point guarantees that the current classifier separates all well-classified points, provided the number of training samples exceed a modest threshold. The stopping condition itself does not require knowledge of the cluster geometry. We also derive a data-dependent accuracy threshold guaranteeing separation from any sufficiently accurate feasible dual solution, independent of the algorithm used to compute such a solution. For FISTA, we obtain explicit data-dependent iteration bounds both for achieving separation with a prescribed positive margin and for satisfying the stopping condition we propose. We further extend our analysis of FISTA to the classical SVM formulation with an unregularized intercept. This analysis appears in the appendix.

Our numerical experiments compare FISTA applied to both H-SVM and the classical SVM with LIBLINEAR (Fan et al., 2008), LIBSVM (Chang and Lin, 2011), and Pegasos (Shalev-Shwartz et al., 2007) on a range of benchmark datasets. The FISTA implementation applied to H-SVM is competitive with these widely used solvers in both runtime and test accuracy, while the stopping condition we propose provides a more robust connection with the quality of the resulting classifier. We investigate this in detail by exhibiting examples in

which LIBLINEAR’s stopping condition is either overly conservative, continuing long after a good classifier has been obtained, or terminates prematurely with a poor classifier.

Related work. Algorithms for support vector machines have been extensively studied, including decomposition methods (Osuna et al., 1997; Platt, 1998; Joachims, 1998; Fan et al., 2005), stochastic first-order methods such as Pegasos (Shalev-Shwartz et al., 2007), and coordinate descent methods such as LIBLINEAR (Fan et al., 2008).

The idea of stopping SVM training using conditions other than a predefined optimization tolerance has appeared previously. In particular, Bandos et al. (Bandos et al., 2007) propose statistical stopping conditions based on the stabilization of moments of the training errors. Their stopping condition is empirical, whereas our work develops computable certificates of correct classification and establishes, under a geometric data model, conditions under which termination guarantees separation of all well-classified points. We also derive explicit data-dependent bounds on the number of FISTA iterations required to achieve separation and satisfy the stopping condition.

Safe-screening methods (Ogawa et al., 2013; Zimmert et al., 2015) are another related area of work. These methods construct safe regions guaranteed to contain an optimizer and use the KKT conditions to certify that certain dual coordinates are at the upper or lower bounds of their box constraints. This allows those coordinates to be fixed, reducing the optimization problem while preserving the optimal solution. In our work, certificates are used to determine when training can terminate, with guarantees that the returned classifier separates the well-classified points under our geometric data model and modest assumptions on the number of training samples.

Finally, our work is related to research analyzing the data-dependent complexity of optimization algorithms. Recent examples include (Dadush et al., 2023; Applegate et al., 2023; Lu and Yang, 2023a,b; Xiong and Freund, 2024), which build on earlier condition-based analyses such as those of Freund and Vera (1999), and Cucker and Peña (2002). These works characterize the computational difficulty of obtaining an approximate solution of the optimization problem in terms of properties of the particular problem instance. Our perspective is different in that the optimization problem we consider is a model of an underlying classification task, and we are interested in how accurately it must be solved before that task can be accomplished. Rather than asking how quickly an algorithm approaches an optimizer, we ask how long before its iterates contain enough information to identify well-classified points and produce a useful classifier.

Outline. Section 2 introduces the SVM and H-SVM formulations, derives the H-SVM dual, motivates the perturbed formulation used throughout the paper, and introduces the proposed stopping condition. This section also derives the primal problem associated with the perturbed H-SVM dual, and its corresponding unconstrained formulation that underlies much of our analysis.

Section 3 introduces the data model we use throughout the paper, as well as the data-dependent theory for identifying and separating well-classified points from approximate dual solutions. A major result of this section is Theorem 9, which proves that under the data model if one point satisfies the stopping condition, all well-classified points must be separated by the returned hyperplane when there are at least 79 training samples. The

section finishes by deriving bounds on the number of iterations before the stopping condition is satisfied.

Section 4 applies the theory of Section 3 to coordinate descent by deriving computable error bounds that allow the use of the proposed stopping condition. This section also introduces an example dataset for which LIBLINEAR’s stopping condition terminates prematurely with a poor classifier.

Section 5 gives implementation details and numerical results, including comparisons with widely used SVM solvers, an investigation of the effect of intercept regularization in H-SVM, and examples illustrating both overly conservative and premature termination of LIBLINEAR.

Details not essential to the readability of the paper are left to the appendix.

Contributions. The contributions of this paper are both theoretical and practical. We derive computable error bounds for FISTA and coordinate descent that yield a practical stopping condition for a perturbed SVM formulation. This condition certifies individual points as correctly classified by both the current classifier and the optimal classifier, without requiring assumptions on the data. Under a geometric data model and a modest assumption on the number of training samples, we show that certifying a single point guarantees that the current classifier separates all well-classified points. We also derive a data-dependent accuracy threshold such that any approximate dual solution within this threshold yields a hyperplane that separates the well-classified points. For FISTA, we additionally obtain explicit data-dependent iteration complexity bounds for satisfying the stopping condition and separating well-classified points by a prescribed margin. Our analysis also reveals limitations of the LIBLINEAR stopping conditions, and we provide examples in which they are either overly conservative or terminate prematurely with a poor classifier. Finally, we describe a FISTA-based algorithm that is simple to implement and whose numerical performance is competitive with widely used SVM solvers across a range of benchmark problems. In Appendix C, we extend our analysis to the classical SVM formulation, which does not regularize the intercept.

2 Support Vector Machines

Consider two classes of points in $\mathbb{R}^d$:

$$\{\mathbf{c}_1, \dots, \mathbf{c}_j\}, \quad \text{and} \quad \{\mathbf{d}_1, \dots, \mathbf{d}_l\}$$

which may or may not be linearly separable. A linear classifier is determined by a normal vector $\mathbf{w} \in \mathbb{R}^d$ and an intercept $t \in \mathbb{R}$, and assigns a point $\mathbf{z}$ according to the sign of $\mathbf{w}^T \mathbf{z} + t$. This problem is traditionally modeled as a convex quadratic programming problem known as the soft-margin support vector machine (SVM):

$$\begin{aligned} \min_{\mathbf{s}, t, \mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{j+l} s_i \\ \text{s.t.} \quad & \mathbf{c}_i^T \mathbf{w} + t \geq 1 - s_i, \quad i = 1, \dots, j, \\ & \mathbf{d}_i^T \mathbf{w} + t \leq -1 + s_{i+j}, \quad i = 1, \dots, l, \\ & \mathbf{s} \geq \mathbf{0}, \end{aligned} \tag{1}$$

where $\gamma > 0$ is a hyperparameter that penalizes violations of $\mathbf{c}_i^T \mathbf{w} + t \geq 1$ and $\mathbf{d}_i^T \mathbf{w} + t \leq -1$. Observe that (1) is bounded below by zero and always feasible, since one may always choose

$\mathbf{s}$ to ensure the constraints are satisfied. In fact, we can rewrite (1) as unconstrained nonsmooth convex optimization by letting $\phi(\theta) := \gamma \max\{1 - \theta, 0\}$ and writing

$$\min_{\mathbf{w}, t} \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^j \phi(\mathbf{c}_i^T \mathbf{w} + t) + \sum_{i=1}^l \phi(-\mathbf{d}_i^T \mathbf{w} - t). \quad (2)$$

An analysis of the classical formulation (1) is provided in Appendix C. In the main text, we focus on the following augmented formulation obtained by writing the classifier $\mathbf{z} \mapsto \mathbf{w}^T \mathbf{z} + t$ as a homogeneous linear classifier in one additional dimension. Namely, we define

$$\tilde{\mathbf{c}}_i = (\mathbf{c}_i, 1), \quad \tilde{\mathbf{d}}_i = (\mathbf{d}_i, 1), \quad \tilde{\mathbf{w}} = (\mathbf{w}, t).$$

Then

$$\tilde{\mathbf{c}}_i^T \tilde{\mathbf{w}} = \mathbf{c}_i^T \mathbf{w} + t, \quad \tilde{\mathbf{d}}_i^T \tilde{\mathbf{w}} = \mathbf{d}_i^T \mathbf{w} + t,$$

and the primal becomes

$$\begin{aligned} \min_{\mathbf{s}, t, \mathbf{w}} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \gamma \sum_{i=1}^{j+l} s_i \\ \text{s.t.} \quad & \tilde{\mathbf{c}}_i^T \tilde{\mathbf{w}} \geq 1 - s_i, \quad i = 1, \dots, j, \\ & \tilde{\mathbf{d}}_i^T \tilde{\mathbf{w}} \leq -1 + s_{i+j}, \quad i = 1, \dots, l, \\ & \mathbf{s} \geq \mathbf{0}. \end{aligned} \quad (3)$$

This augmentation is a standard way to handle the intercept term in large-scale linear classification; for example, it is used in LIBLINEAR Fan et al. (2008). The resulting model differs from the classical SVM, which treats the intercept as an unregularized scalar, because the intercept t is included in the regularized vector $\tilde{\mathbf{w}}$. We will call the formulation obtained by this augmentation H-SVM (homogeneous SVM). A particular advantage of H-SVM is that the corresponding dual feasible set is a box, so projection is efficient to compute. One consequence of regularizing the intercept is that the resulting model need not be invariant under translations of the feature vectors. We investigate the practical significance of this issue in Section 5.

2.1 The dual H-SVM formulation

Let $n := j + l$ and define the matrix $\tilde{H} \in \mathbb{R}^{n \times (d+1)}$ by

$$\tilde{H}^T := (\tilde{C}^T \quad -\tilde{D}^T),$$

where $\tilde{C} = (\tilde{\mathbf{c}}_1^T, \dots, \tilde{\mathbf{c}}_j^T) \in \mathbb{R}^{j \times (d+1)}$ and $\tilde{D} = (\tilde{\mathbf{d}}_1^T, \dots, \tilde{\mathbf{d}}_l^T) \in \mathbb{R}^{l \times (d+1)}$. For $\mathbf{r} := (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^j \times \mathbb{R}^l$, the dual of (3) is

$$\begin{aligned} \max \quad & -\frac{1}{2} \|\tilde{H}^T \mathbf{r}\|^2 + \mathbf{e}^T \mathbf{r} \\ \text{s.t.} \quad & \mathbf{r} \in \Omega, \end{aligned} \quad (4)$$

where

$$\Omega := \{\mathbf{r} : \mathbf{0} \leq \mathbf{r} \leq \gamma \mathbf{e}\}.$$

Projection onto Ω is given by component-wise clipping, while minimization with respect to a single coordinate also has a closed-form solution. Thus, (4) is suitable for both projected-gradient algorithms and box-constrained coordinate descent.

The complementarity conditions associated with the upper bounds in (4) imply that every optimal primal-dual pair satisfies $s_i(\gamma - u_i) = 0$ for all $i \in [j]$ and $s_{j+i}(\gamma - v_i) = 0$ for all $i \in [l]$. Thus, if we have an optimal dual solution with $u_i < \gamma$, then $s_i = 0$ and $\mathbf{c}_i$ is correctly classified. The same applies for $\mathbf{v}$. This suggests identifying correctly classified points by determining which optimal dual coordinates are strictly below the upper box constraint.

The difficulty with this approach is that the dual problem may have multiple optimal solutions. In particular, a point $\mathbf{c}_i$ may satisfy $u_i < \gamma$ at one optimal solution but $u_i = \gamma$ at another. An example of such a case is given in Appendix A. Consequently, convergence to an arbitrary optimal solution of the dual does not guarantee that all correctly classified points can be identified through the strict inequalities

$$u_i < \gamma \quad \text{and} \quad v_i < \gamma.$$

This obstruction is not specific to a particular algorithm for solving (4). Different first-order methods may converge to different optimal solutions, and the points certified by the strict coordinate inequalities may depend on the optimal solution selected by the algorithm. Recent results establish convergence of the iterates generated by FISTA (Bot et al., 2025); however, these results do not identify which minimizer FISTA converges to.

2.2 The perturbed H-SVM dual

To ensure that the dual problem has a unique optimal solution, we introduce a strongly concave perturbation of the dual objective. Specifically, for $\mu > 0$, we consider

$$\begin{aligned} \max \quad & g(\mathbf{r}) := -\frac{1}{2}\|\tilde{H}^T \mathbf{r}\|^2 + \mathbf{e}^T \mathbf{r} - \frac{\mu}{2}\|\mathbf{r}\|^2 \\ \text{s.t.} \quad & \mathbf{r} \in \Omega. \end{aligned} \tag{5}$$

Equivalently, we may write

$$g(\mathbf{u}, \mathbf{v}) = -\frac{1}{2}\|C^T \mathbf{u} - D^T \mathbf{v}\|^2 - \frac{1}{2}(\mathbf{e}^T \mathbf{u} - \mathbf{e}^T \mathbf{v})^2 + \mathbf{e}^T \mathbf{u} + \mathbf{e}^T \mathbf{v} - \frac{\mu}{2}\|\mathbf{u}\|^2 - \frac{\mu}{2}\|\mathbf{v}\|^2.$$

As shown below, g is μ -strongly concave, and hence this problem has a unique optimal solution, which we denote

$$\mathbf{r}^* := (\mathbf{u}^*, \mathbf{v}^*).$$

To relate the perturbed dual to its associated primal classifier, we derive its associated primal problem:

$$\begin{aligned} \min_{\mathbf{w}, t, \mathbf{s}, \mathbf{z}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}t^2 + \gamma \mathbf{e}^T \mathbf{s} + \frac{1}{2\mu}\|\mathbf{z}\|^2 \\ \text{s.t.} \quad & C\mathbf{w} + t\mathbf{e} + \mathbf{z}_1 \geq \mathbf{e} - \mathbf{s}_1, \\ & D\mathbf{w} + t\mathbf{e} - \mathbf{z}_2 \leq -\mathbf{e} + \mathbf{s}_2, \\ & \mathbf{s} \geq \mathbf{0}, \end{aligned} \tag{6}$$

where $\mathbf{s}_1 = \mathbf{s}(1 : j)$, $\mathbf{s}_2 = \mathbf{s}(j + 1 : j + l)$ and $\mathbf{z}_1 = \mathbf{z}(1 : j)$, $\mathbf{z}_2 = \mathbf{z}(j + 1 : j + l)$. The KKT conditions relating (5) and (6) are

$$\begin{aligned}
 \mathbf{u} &= \mathbf{z}_1 / \mu \\
 \mathbf{v} &= \mathbf{z}_2 / \mu \\
 \mathbf{w} &= C^T \mathbf{u} - D^T \mathbf{v} \\
 t &= \mathbf{e}^T \mathbf{u} - \mathbf{e}^T \mathbf{v} \\
 \mathbf{z}_1 &= (1 - t)\mathbf{e} - \mathbf{s}_1 - C\mathbf{w} + \boldsymbol{\zeta}_1 \\
 \mathbf{z}_2 &= (1 + t)\mathbf{e} - \mathbf{s}_2 + D\mathbf{w} + \boldsymbol{\zeta}_2 \\
 \mathbf{s}_1^T (\mathbf{u} - \gamma \mathbf{e}) &= 0 \\
 \mathbf{s}_2^T (\mathbf{v} - \gamma \mathbf{e}) &= 0 \\
 \boldsymbol{\zeta}_1^T \mathbf{u} &= 0 \\
 \boldsymbol{\zeta}_2^T \mathbf{v} &= 0 \\
 \mathbf{0} &\leq \mathbf{u} \leq \gamma \mathbf{e} \\
 \mathbf{0} &\leq \mathbf{v} \leq \gamma \mathbf{e} \\
 \mathbf{s} &\geq \mathbf{0} \\
 \boldsymbol{\zeta} &\geq \mathbf{0},
 \end{aligned} \tag{7}$$

where $\boldsymbol{\zeta}_1 = \boldsymbol{\zeta}(1 : j)$ and $\boldsymbol{\zeta}_2 = \boldsymbol{\zeta}(j + 1 : j + l)$. In particular, for the optimal classifier $\tilde{\mathbf{w}}^* = (\mathbf{w}^*, t^*)$ and optimal dual solution $\mathbf{r}^*$, there exists the relation

$$\begin{pmatrix} \mathbf{w}^* \\ t^* \end{pmatrix} = \tilde{H}^T \mathbf{r}^* = \begin{pmatrix} C^T \mathbf{u}^* - D^T \mathbf{v}^* \\ \mathbf{e}^T \mathbf{u}^* - \mathbf{e}^T \mathbf{v}^* \end{pmatrix}. \tag{8}$$

If we suppose $\mathbf{w}, t$ are known in (6), then the entries of $\mathbf{s}$ and $\mathbf{z}$ separate. That is, if for some $i = 1, \dots, j$ we let $\theta = \mathbf{c}_i^T \mathbf{w} + t$ or for some $i = j + 1, \dots, l$, $\theta = -\mathbf{d}_{i-j}^T \mathbf{w} - t$, then the optimal s_i, z_i solve the problem

$$\psi(\theta) := \min \left\{ \gamma s + \frac{1}{2\mu} z^2 : \theta \geq 1 - s - z, s \geq 0 \right\}, \tag{9}$$

which can be written in closed form as the following differentiable approximation to hinge loss:

$$\psi(\theta) = \begin{cases} -\frac{1}{2}\mu\gamma^2 + \gamma(1 - \theta), & \theta \in (-\infty, 1 - \mu\gamma], \\ \frac{1}{2\mu}(1 - \theta)^2, & \theta \in [1 - \mu\gamma, 1], \\ 0, & \theta \in [1, +\infty). \end{cases} \tag{10}$$

In the context of (6), θ can be viewed as the signed margin of a point $\mathbf{h}_i$ with corresponding class label y_i , where $y_i = 1$ and $\mathbf{h}_i = \mathbf{c}_i$ for $i = 1, \dots, j$, and $y_i = -1$ and $\mathbf{h}_i = \mathbf{d}_{i-j}$ for $i = j + 1, \dots, l$:

$$\theta = y_i(\mathbf{w}^T \mathbf{h}_i + t).$$

With a closed-form solution to $\psi(\theta)$, we may now write (6) as an unconstrained convex optimization problem which will be helpful for our analysis.

$$\min_{\mathbf{w}, t} f(\mathbf{w}, t) := \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{2} t^2 + \sum_{i=1}^j \psi(\mathbf{c}_i^T \mathbf{w} + t) + \sum_{i=1}^l \psi(-\mathbf{d}_i^T \mathbf{w} - t). \tag{11}$$

2.3 A practical stopping condition for FISTA

The algorithm we propose applies the strongly convex variant of FISTA (Beck, 2017, Chapter 7) to the perturbed H-SVM dual (5). Our aim is to terminate the algorithm once its iterates provide a certificate of progress on the underlying classification task. We first prove that g is strongly concave and L -smooth, allowing the application of FISTA. Then we describe FISTA and the termination test, and establish a classification guarantee that does not require any geometric assumptions on the data.

Theorem 1 *Suppose all points lie within a fixed distance R from the origin:*

$$\|\mathbf{c}_i\| \leq R, \quad \|\mathbf{d}_i\| \leq R.$$

Then g is μ -strongly concave and L -smooth with

$$L = \mu + \|\tilde{H}\|^2 \leq \mu + n(R^2 + 1).$$

Proof We have

$$\nabla g(\mathbf{r}) = -\tilde{H}\tilde{H}^T \mathbf{r} + \mathbf{e} - \mu \mathbf{r},$$

and therefore

$$\nabla^2 g(\mathbf{r}) = -\tilde{H}\tilde{H}^T - \mu I.$$

Hence g is μ -strongly concave. Moreover,

$$\|\nabla^2 g(\mathbf{r})\|_2 = \|\tilde{H}\tilde{H}^T + \mu I\|_2 = \|\tilde{H}\|^2 + \mu.$$

Since every row of $\tilde{H}$ has norm at most $\sqrt{R^2 + 1}$,

$$\|\tilde{H}\|^2 \leq \|\tilde{H}\|_F^2 \leq n(R^2 + 1).$$

The stated bound follows. ■

We now describe the algorithm and its stopping condition for general parameter values $\mu, \gamma > 0$. The classification guarantee below assumes $\mu\gamma < 1$. Specific parameter choices will be introduced in Section 2.4.

Let $\kappa := L/\mu$, and define the momentum parameter

$$\beta := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}.$$

Starting from $\mathbf{r}_0 = \mathbf{y}_0 = \mathbf{0}$, FISTA computes the iterates according to

$$\mathbf{r}_{k+1} = \text{proj}_\Omega \left(\mathbf{y}_k + \frac{1}{L} \nabla g(\mathbf{y}_k) \right), \tag{12}$$

$$\mathbf{y}_{k+1} = \mathbf{r}_{k+1} + \beta(\mathbf{r}_{k+1} - \mathbf{r}_k). \tag{13}$$

Here $\mathbf{r}_{k+1}$ is the feasible dual iterate, while $\mathbf{y}_k$ is the momentum point at which the gradient is evaluated. The projected-gradient step provides a computable bound on the distance to $\mathbf{r}^*$. Define

$$\epsilon_{k+1}^{\text{FISTA}} := \kappa \|\mathbf{y}_k - \mathbf{r}_{k+1}\|. \tag{14}$$

Theorem 2 Let $\mathbf{y} \in \mathbb{R}^n$, and define $\mathbf{r} := \text{proj}_\Omega(\mathbf{y} + \frac{1}{L}\nabla g(\mathbf{y}))$, $\epsilon := \kappa\|\mathbf{y} - \mathbf{r}\|$. Then

$$\|\mathbf{r} - \mathbf{r}^*\| \leq \epsilon.$$

In particular, the FISTA iterates satisfy

$$\|\mathbf{r}_{k+1} - \mathbf{r}^*\| \leq \epsilon_{k+1}^{\text{FISTA}}.$$

Proof Write

$$Q := \tilde{H}\tilde{H}^T + \mu I, \quad g(\mathbf{r}) = -\frac{1}{2}\mathbf{r}^T Q \mathbf{r} + \mathbf{e}^T \mathbf{r}.$$

The optimality conditions for the projection give

$$\mathbf{z} := \nabla g(\mathbf{y}) + L(\mathbf{y} - \mathbf{r}) \in \mathcal{N}_\Omega(\mathbf{r}),$$

where $\mathcal{N}_\Omega$ denotes the normal cone of Ω . Define the following term, which represents a subgradient of the convex function $-g(\mathbf{r}) + i(\mathbf{r})$:

$$\mathbf{q} := Q\mathbf{r} - \mathbf{e} + \mathbf{z} = (LI - Q)(\mathbf{y} - \mathbf{r}). \quad (15)$$

Since $\mu I \preceq Q \preceq LI$, we have

$$\|\mathbf{q}\| \leq (L - \mu)\|\mathbf{y} - \mathbf{r}\| \leq L\|\mathbf{y} - \mathbf{r}\| = \mu\epsilon. \quad (16)$$

Meanwhile, at the optimal solution,

$$\mathbf{z}^* := \mathbf{e} - Q\mathbf{r}^* \in \mathcal{N}_\Omega(\mathbf{r}^*).$$

Thus, after adding and subtracting $\mathbf{z}^*$ from $\mathbf{q}$, we may write

$$\mathbf{q} = Q(\mathbf{r} - \mathbf{r}^*) + \mathbf{z} - \mathbf{z}^*.$$

By monotonicity of the normal cone,

$$(\mathbf{r} - \mathbf{r}^*)^T \mathbf{q} \geq (\mathbf{r} - \mathbf{r}^*)^T Q(\mathbf{r} - \mathbf{r}^*) \geq \mu\|\mathbf{r} - \mathbf{r}^*\|^2.$$

If $\mathbf{r} \neq \mathbf{r}^*$, applying Cauchy-Schwarz gives

$$\|\mathbf{r} - \mathbf{r}^*\| \leq \frac{\|\mathbf{q}\|}{\mu} \leq \epsilon.$$

If $\mathbf{r} = \mathbf{r}^*$ the bound is immediate. ■

Our stopping condition terminates at the first iteration for which at least one coordinate satisfies

$$\gamma - r_{k+1,i} > \epsilon_{k+1}^{\text{FISTA}}, \quad \text{for some } i \in [n]. \quad (17)$$

Writing $\mathbf{r}_{k+1} = (\mathbf{u}_{k+1}, \mathbf{v}_{k+1})$, we return the classifier computed by

$$\mathbf{w}_{k+1} = C^T \mathbf{u}_{k+1} - D^T \mathbf{v}_{k+1}, \quad t_{k+1} = \mathbf{e}^T \mathbf{u}_{k+1} - \mathbf{e}^T \mathbf{v}_{k+1}.$$

Equivalently, $(\mathbf{w}_{k+1}, t_{k+1}) = \tilde{H}^T \mathbf{r}_{k+1}$.

In the following theorem, we show that (17) means that the point corresponding to the dual coordinate that triggered the stopping condition is correctly classified by both the optimal classifier and the returned classifier. For $\mathbf{r} \in \Omega$, write $(\mathbf{w}(\mathbf{r}), t(\mathbf{r})) = \tilde{H}^T \mathbf{r}$, and define the corresponding signed margins by

$$\theta_i(\mathbf{r}) := \begin{cases} \mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}), & i = 1, \dots, j, \\ -\mathbf{d}_{i-j}^T \mathbf{w}(\mathbf{r}) - t(\mathbf{r}), & i = j+1, \dots, n. \end{cases}$$

In vector form, $\boldsymbol{\theta}(\mathbf{r}) = \tilde{H} \tilde{H}^T \mathbf{r}$.

Theorem 3 *Suppose $\mu\gamma < 1$. Let $\mathbf{r}$ be the projected gradient step for some $\mathbf{y} \in \mathbb{R}^n$, and $\epsilon = \kappa \|\mathbf{y} - \mathbf{r}\|$, as in Theorem 2. If $\gamma - r_i > \epsilon$, then*

$$\theta_i(\mathbf{r}) > 1 - \mu\gamma > 0, \quad \theta_i(\mathbf{r}^*) > 1 - \mu\gamma > 0.$$

Proof Let $\mathbf{q}$ and $\mathbf{z}$ be the subgradient and normal vectors defined in the proof of Theorem 2. Since $r_i < \gamma$ and $\mathbf{z} \in \mathcal{N}_\Omega(\mathbf{r})$, we have $z_i \leq 0$. Otherwise, one could construct a vector $\boldsymbol{\omega} \in \Omega$ with $\omega_j = r_j$ for $j \neq i$ and $\omega_i = \gamma$, where $\mathbf{z}^T(\boldsymbol{\omega} - \mathbf{r}) > 0$. By (15),

$$\theta_i(\mathbf{r}) = 1 - \mu r_i + q_i - z_i.$$

Using (16), the fact that $z_i \leq 0$, and the stopping condition, we obtain

$$\begin{aligned} \theta_i(\mathbf{r}) &\geq 1 - \mu r_i - \mu\epsilon \\ &= 1 - \mu(r_i + \epsilon) \\ &> 1 - \mu\gamma. \end{aligned}$$

For the optimal classifier, Theorem 2 gives

$$r_i^* \leq r_i + \epsilon < \gamma.$$

Hence by the same reasoning as for $\mathbf{z}$, the normal vector $\mathbf{z}^* = \mathbf{e} - Q\mathbf{r}^*$ satisfies $z_i^* \leq 0$. Therefore

$$\theta_i(\mathbf{r}^*) = 1 - \mu r_i^* - z_i^* \geq 1 - \mu r_i^* > 1 - \mu\gamma.$$

■

The stopping condition we introduced in this section uses quantities available during FISTA iterations and does not require knowledge of the geometric structure of the data. The theorem above guarantees correctness of each point satisfying the test. In Section 3, we introduce a geometric data model under which the existence of a single point satisfying the stopping condition guarantees that the returned classifier correctly classifies every well-classified point. We also bound the number of FISTA iterations required for the test to be satisfied.

2.4 Optimal margins and dual coordinates

It will be useful in our analysis to relate the signed margins of the optimal classifier to the coordinates of the optimal solution of the perturbed dual. For each $\mathbf{c}_i$ and $\mathbf{d}_i$, define

$$\theta_{\mathbf{c},i}^* := \mathbf{c}_i^T \mathbf{w}^* + t^*, \quad \theta_{\mathbf{d},i}^* := -\mathbf{d}_i^T \mathbf{w}^* - t^*$$

These quantities are positive when the corresponding points are correctly classified by the optimal hyperplane given by $(\mathbf{w}^*, t^*)$.

The complementarity conditions for (5) and (6) imply that $u_i^* < \gamma$ implies $s_i^* = 0$ and $v_i^* < \gamma$ implies $s_{j+i}^* = 0$. Unlike in the unperturbed H-SVM primal, however, zero slack does not imply that the corresponding signed margin is at least one, because the perturbed primal contains the additional variables $\mathbf{z}$.

Obtaining precise bounds requires fixing the model parameters. For the remainder of the paper, we choose the model parameters $\gamma := 64/n$ and $\mu := n/128$. These choices give $\mu\gamma = 1/2$, so the quadratic portion of ψ lies between the breakpoints $\theta = 1/2$ and $\theta = 1$. Hence, if the signed margin $\theta \geq 1$, the corresponding point makes no contribution to the objective, but as the margin decreases to $\theta \in [1/2, 1]$, the contribution begins to increase quadratically, before becoming linear when $\theta \leq 1/2$. This behavior is depicted in Figure 1. Our choice of γ follows the common convention of normalizing the hinge loss term by the sample size, which ensures that the objective function remains well-scaled as n varies. Such normalization appears throughout the literature, for example in (Shalev-Shwartz and Ben-David, 2014, Equation (15.4)) and (Shalev-Shwartz et al., 2007, Equation (1)).

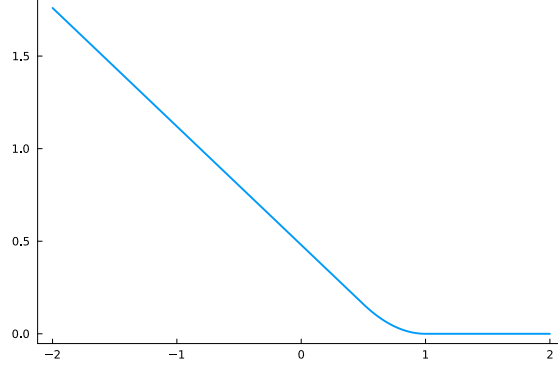


Figure 1: A plot of $\psi(\theta)$, where $n = 100$, $\gamma = 64/n$, $\mu = n/128$.

The following result relates each optimal dual coordinate with the corresponding optimal signed margin. In particular, it shows that $u_i^* < \gamma$ implies $\theta_{\mathbf{c},i}^* > 1/2$ and likewise for v_i^* and $\theta_{\mathbf{d},i}^*$. Thus, every optimal dual coordinate known to be strictly below γ certifies that the corresponding point is correctly classified. Our analysis will show that the dual coordinates associated with all well-classified points are uniformly bounded away from γ .

Lemma 4 *Let $(\mathbf{u}^*, \mathbf{v}^*)$ be the unique optimal solution of the perturbed H-SVM dual, and let $(\mathbf{w}^*, t^*)$ be the corresponding optimal classifier. Then*

$$u_i^* = \text{proj}_{[0,\gamma]} \left(\frac{1 - \theta_{\mathbf{c},i}^*}{\mu} \right), \quad \forall i \in [j],$$

and

$$v_i^* = \text{proj}_{[0,\gamma]} \left(\frac{1 - \theta_{\mathbf{d},i}^*}{\mu} \right), \quad \forall i \in [l].$$

Equivalently,

$$u_i^* = -\psi'(\theta_{\mathbf{c},i}^*), \quad \forall i \in [j], \quad \text{and} \quad v_i^* = -\psi'(\theta_{\mathbf{d},i}^*), \quad \forall i \in [l].$$

The proof is given in Appendix B.

We next use the relationship established in Lemma 4 to translate lower bounds on the optimal signed margins into upper bounds on the corresponding optimal dual coordinates.

Lemma 5 *If $\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K$ for some $K \in [0, 1/2)$, then $u_i^* \leq 2K\gamma < \gamma$. Likewise, if $-\mathbf{d}_i^T \mathbf{w}^* - t^* \geq 1 - K$ for some $K \in [0, 1/2)$, then $v_i^* \leq 2K\gamma < \gamma$.*

Proof Fix $i \in [j]$ and suppose $\theta_{\mathbf{c},i}^* \geq 1 - K$. From Lemma 4

$$u_i^* = \text{proj}_{[0,\gamma]} \left(\frac{1 - \theta_{\mathbf{c},i}^*}{\mu} \right),$$

and since $\theta_{\mathbf{c},i}^* \geq 1 - K$ and $\mu\gamma = 1/2$,

$$\frac{1 - \theta_{\mathbf{c},i}^*}{\mu} \leq \frac{K}{\mu} = 2K\gamma < \gamma.$$

It follows that $u_i^* \leq 2K\gamma < \gamma$. The proof for v_i^* is analogous. ■

The preceding lemma shows that a uniform lower bound on the optimal signed margins yields a uniform fractional separation of the associated optimal dual coordinates from the upper box constraint. In particular, if $\mathbf{c}_i^T \mathbf{w}^* + t^* \geq \frac{1}{2} + \eta$ for some $\eta \in (0, 1/2]$, then taking $K = \frac{1}{2} - \eta$ gives $u_i^* \leq (1 - 2\eta)\gamma$. The same conclusion holds for v_i^* .

3 Data-dependent classification and termination guarantees

We now introduce a geometric data model under which the proposed stopping condition guarantees separation of the two clusters of well-classified points. We then derive a priori iteration bounds for achieving a prescribed positive margin and for satisfying the stopping condition.

3.1 Data model

We will assume that the data form two clusters, each with a center, plus noisy points. Suppose $\sigma_1, \sigma_2 \in [1/2, 1]$ and $0 < \rho < \min(\sigma_1, \sigma_2)$. We assume that there are j_0 “well-classified” points in $B(\sigma_1 \mathbf{e}_1, \rho)$, there are l_0 “well-classified” points in $B(-\sigma_2 \mathbf{e}_1, \rho)$, and the n_{noise} noisy points are contained in the ball that encloses the two balls of *well-classified* points:

$$B \left(\frac{\sigma_1 - \sigma_2}{2} \mathbf{e}_1, \frac{\sigma_1 + \sigma_2}{2} + \rho \right).$$

In our model we have $j_0 + l_0 + n_{\text{noise}} = n$, where $j_0 \geq n/3$, $l_0 \geq n/3$, and $n_{\text{noise}} \leq \nu n$ with

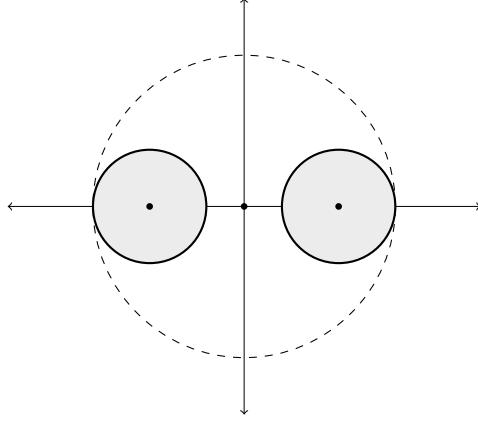


Figure 2: An illustration of the assumption on the data, where we have two balls of well-classified points centered at -1 and $+1$, and a larger ball containing the two, which also includes the noisy points.

$\nu > 0$. These assumptions imply that every data point lies within distance

$$R := \max\{\sigma_1, \sigma_2\} + \rho < 2$$

of the origin. Under the data model, our parameter choices give a uniform bound on the condition number κ . Taking $L = \mu + \|\tilde{H}\|^2$, Theorem 1 and the bound $R < 2$ give

$$\kappa = \frac{L}{\mu} \leq 1 + 128(R^2 + 1) < 641.$$

Thus, within our model, the condition number is bounded by a constant independent of both the number of observations and the ambient dimension. We use this bound in the iteration-complexity analysis below.

3.2 Correctness of the classifier at termination

In this section we establish a lower bound on the optimal margins of the well-classified points. Lemma 5 translates these bounds into upper bounds on the corresponding optimal dual coordinates. We then show that the proposed stopping condition guarantees separation of the well-classified points by the returned classifier.

Margin bounds for well-classified points. Define $\bar{\sigma} := \frac{\sigma_1 + \sigma_2}{2}$, $\delta := \frac{(\sigma_1 - \sigma_2)^2}{4}$, and

$$\xi := \sqrt{\frac{1 + \delta}{(\bar{\sigma} - \rho)^2} + 128\nu \left(\frac{3}{4} + \frac{\bar{\sigma} + \rho}{\bar{\sigma} - \rho} \right)}.$$

Let $\mathcal{C} \subseteq \{1, \dots, j\}$ denote the indices of the well-classified points in the $\mathbf{c}$ -class, and let $\mathcal{D} \subseteq \{1, \dots, l\}$ denote the indices of the well-classified points in the $\mathbf{d}$ -class.

Lemma 6 *Suppose the data satisfy the assumptions of Section 3.1 and let $(\mathbf{w}^*, t^*)$ solve (11). Then $\|(\mathbf{w}^*, t^*)\| \leq \xi$. In particular, it must be that $\|\mathbf{w}^*\| \leq \xi$ and $|t^*| \leq \xi$.*

Proof For readability, let $m := \frac{\sigma_1 - \sigma_2}{2}$, $a := \bar{\sigma} - \rho$, and $b := \bar{\sigma} + \rho$. Recall the objective in (11):

$$f(\mathbf{w}, t) = \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{2} t^2 + \sum_{i=1}^j \psi(\mathbf{c}_i^T \mathbf{w} + t) + \sum_{i=1}^l \psi(-\mathbf{d}_i^T \mathbf{w} - t)$$

and note that $\bar{\mathbf{w}} = \mathbf{e}_1/a$, $\bar{t} = -m/a$ is a feasible solution, where the fact that $\rho < \min(\sigma_1, \sigma_2)$ ensures that $a > 0$. For all i we have

$$\mathbf{c}_i^T \bar{\mathbf{w}} + \bar{t} = \frac{c_{i,1} - m}{a}, \quad -\mathbf{d}_i^T \bar{\mathbf{w}} - \bar{t} = \frac{-d_{i,1} + m}{a}.$$

For $i \in \mathcal{C}$, $c_{i,1} - m \geq \sigma_1 - \rho - m = a$. With the above this gives

$$\mathbf{c}_i^T \bar{\mathbf{w}} + \bar{t} \geq 1, \quad \forall i \in \mathcal{C}.$$

For $i \in \mathcal{D}$, $-d_{i,1} + m \geq \sigma_2 - \rho + m = a$, and as for the $\mathbf{c}$ case, this gives

$$-\mathbf{d}_i^T \bar{\mathbf{w}} - \bar{t} \geq 1, \quad \forall i \in \mathcal{D}.$$

Thus, the well-classified points make no contribution to the objective at $(\bar{\mathbf{w}}, \bar{t})$. On the other hand, every noisy point lies in the ball centered at $m\mathbf{e}_1$ with radius b . Hence, for noisy $\mathbf{c}_i$ and $\mathbf{d}_i$,

$$c_{i,1} - m \geq m - b - m = -b, \quad -d_{i,1} + m \geq -m - b + m = -b.$$

Therefore, the signed margins of the noisy points must be bounded below by $-b/a$. Since $-b/a < 0$,

$$\psi(-b/a) = -\frac{1}{2}\mu\gamma^2 + \gamma \left(1 + \frac{b}{a}\right) = \gamma \left(\frac{3}{4} + \frac{b}{a}\right).$$

Applying the above with the fact that ψ is nonincreasing yields

$$\psi(\theta) \leq \psi(-b/a) = \gamma \left(\frac{3}{4} + \frac{b}{a}\right),$$

where θ represents the signed margin of a noisy point.

The sum terms in (11) must be bounded as follows, since $n_{\text{noise}} \leq \nu n$ and $\gamma = 64/n$:

$$\sum_{i=1}^j \psi(\mathbf{c}_i^T \bar{\mathbf{w}} + \bar{t}) + \sum_{i=1}^l \psi(-\mathbf{d}_i^T \bar{\mathbf{w}} - \bar{t}) \leq 64\nu \left(\frac{3}{4} + \frac{b}{a}\right).$$

Moreover, $\frac{1}{2} \|\bar{\mathbf{w}}\|^2 + \frac{1}{2} \bar{t}^2 = \frac{1+m^2}{2a^2}$, so we have

$$f(\bar{\mathbf{w}}, \bar{t}) \leq \frac{1+m^2}{2a^2} + 64\nu \left(\frac{3}{4} + \frac{b}{a}\right),$$

and thus

$$f(\mathbf{w}^*, t^*) \leq f(\bar{\mathbf{w}}, \bar{t}) \leq \frac{1+m^2}{2a^2} + 64\nu \left(\frac{3}{4} + \frac{b}{a}\right).$$

We may now obtain a bound on $\|(\mathbf{w}^*, t^*)\|$ by noting that $\|(\mathbf{w}^*, t^*)\|^2 \leq 2f(\mathbf{w}^*, t^*)$:

$$\begin{aligned} \|(\mathbf{w}^*, t^*)\| &\leq \sqrt{2f(\mathbf{w}^*, t^*)} \\ &\leq \sqrt{\frac{1+m^2}{a^2} + 128\nu \left(\frac{3}{4} + \frac{b}{a}\right)} \\ &\leq \sqrt{\frac{1+\delta}{(\bar{\sigma}-\rho)^2} + 128\nu \left(\frac{3}{4} + \frac{\bar{\sigma}+\rho}{\bar{\sigma}-\rho}\right)}. \end{aligned}$$

■

Let $p_c := j_0/n$ and $p_d := l_0/n$. By our assumptions in Section 3.1, $p_c, p_d \geq 1/3$. Define

$$K_c := \frac{\xi \sqrt{1 + (\sigma_2 - \rho)^2} + 128\bar{\sigma}\nu}{128p_c(\sigma_1 + \sigma_2 - 2\rho)},$$

and

$$K_d := \frac{\xi \sqrt{1 + (\sigma_1 - \rho)^2} + 128\bar{\sigma}\nu}{128p_d(\sigma_1 + \sigma_2 - 2\rho)}.$$

Lemma 7 *Suppose the data satisfy the assumptions of Section 3.1. If $K_c < 1/2$, then there exists some $i \in \mathcal{C}$ such that*

$$\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K_c.$$

Likewise if $K_d < 1/2$, there exists some $i \in \mathcal{D}$ such that

$$-\mathbf{d}_i^T \mathbf{w}^* - t^* \geq 1 - K_d.$$

The proof is given in Appendix B.

Corollary 8 *Suppose the data satisfy the assumptions of Section 3.1. If $K_c < 1/2$, for any $i \in \mathcal{C}$,*

$$\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K_c - 2\rho\xi.$$

Likewise if $K_d < 1/2$, for any $i \in \mathcal{D}$,

$$-\mathbf{d}_i^T \mathbf{w}^* - t^* \geq 1 - K_d - 2\rho\xi.$$

Proof Again we prove this for $\mathbf{c}$ -points only because the proof for the $\mathbf{d}$ -points is analogous.

Fix an arbitrary $i \in \mathcal{C}$. Let i' be the index whose existence is guaranteed by Lemma 7.

We may write

$$\begin{aligned} \mathbf{c}_i^T \mathbf{w}^* + t^* &= \mathbf{c}_{i'}^T \mathbf{w}^* + t^* + (\mathbf{c}_i - \mathbf{c}_{i'})^T \mathbf{w}^* \\ &\geq \mathbf{c}_{i'}^T \mathbf{w}^* + t^* - \|\mathbf{c}_i - \mathbf{c}_{i'}\| \cdot \|\mathbf{w}^*\| \\ &\geq \mathbf{c}_{i'}^T \mathbf{w}^* + t^* - 2\rho\|\mathbf{w}^*\| \\ &\geq \mathbf{c}_{i'}^T \mathbf{w}^* + t^* - 2\rho\xi \\ &\geq 1 - K_c - 2\rho\xi. \end{aligned}$$

The second inequality came from the fact that $\mathbf{c}_i, \mathbf{c}_{i'} \in B(\sigma_1 \mathbf{e}_1, \rho)$, so $\|\mathbf{c}_i - \mathbf{c}_{i'}\| \leq 2\rho$. The third inequality applies the bound on $\|\mathbf{w}^*\|$ from Lemma 6, and the final inequality applies Lemma 7. $\blacksquare$

Define $K := \max\{K_c, K_d\} + 2\rho\xi$. If $K < 1/2$, Corollary 8 implies that every $i \in \mathcal{C}$ satisfies

$$\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K,$$

and every $i \in \mathcal{D}$ satisfies

$$-\mathbf{d}_i^T \mathbf{w}^* - t^* \geq 1 - K.$$

Moreover, as $\rho, \nu \rightarrow 0$,

$$\xi \rightarrow \frac{2\sqrt{1+\delta}}{\sigma_1 + \sigma_2},$$

and thus

$$K \rightarrow K_0 := \frac{\sqrt{1+\delta}}{64(\sigma_1 + \sigma_2)^2} \max\left(\frac{\sqrt{1+\sigma_2^2}}{p_c}, \frac{\sqrt{1+\sigma_1^2}}{p_d}\right).$$

Since $p_c, p_d \geq 1/3$ and $\sigma_1, \sigma_2 \in [1/2, 1]$, we obtain the inequality

$$K_0 \leq \frac{3\sqrt{34}}{256} \approx 0.0683 < \frac{1}{2}.$$

Therefore, for sufficiently small ρ, ν , we have $K < 1/2$. We will make the assumption that $K < 1/2$ from now on.

Separation at termination. We can now prove under the data model that once one point satisfies the stopping condition, all well-classified points must be separated by the returned classifier.

Theorem 9 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$ and $n \geq 79$. For some $\mathbf{y} \in \mathbb{R}^n$, define $\mathbf{r} := \text{proj}_\Omega(\mathbf{y} + \frac{1}{L}\nabla g(\mathbf{y}))$ and $\epsilon := \kappa\|\mathbf{y} - \mathbf{r}\|$. Let $\tilde{\mathbf{w}}(\mathbf{r}) = (\mathbf{w}(\mathbf{r}), t(\mathbf{r})) = \tilde{H}^T \mathbf{r}$. If $\gamma - r_i > \epsilon$ for some $i \in [n]$, then*

$$\mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) > 0, \quad \forall i \in \mathcal{C},$$

and

$$-\mathbf{d}_i^T \mathbf{w}(\mathbf{r}) - t(\mathbf{r}) > 0, \quad \forall i \in \mathcal{D}.$$

Proof Recall $\tilde{\mathbf{w}}^* = (\mathbf{w}^*, t^*) = \tilde{H}^T \mathbf{r}^*$ is the optimal classifier. We showed in the proof of Theorem 2 that for $\mathbf{z} = \nabla g(\mathbf{y}) + L(\mathbf{y} - \mathbf{r}) \in \mathcal{N}_\Omega(\mathbf{r})$ and $\mathbf{z}^* = \nabla g(\mathbf{r}^*) \in \mathcal{N}_\Omega(\mathbf{r}^*)$, the subgradient $\mathbf{q} = Q\mathbf{r} - \mathbf{e} + \mathbf{z}$ satisfies

$$\|\mathbf{q}\| \leq \mu\epsilon, \quad \mathbf{q} = Q(\mathbf{r} - \mathbf{r}^*) + \mathbf{z} - \mathbf{z}^*.$$

Since $\mathcal{N}_\Omega$ is monotone,

$$\begin{aligned} \mathbf{q}^T(\mathbf{r} - \mathbf{r}^*) &= (\mathbf{r} - \mathbf{r}^*)^T Q(\mathbf{r} - \mathbf{r}^*) + (\mathbf{r} - \mathbf{r}^*)^T (\mathbf{z} - \mathbf{z}^*) \\ &\geq (\mathbf{r} - \mathbf{r}^*)^T Q(\mathbf{r} - \mathbf{r}^*) \\ &= \|\tilde{H}^T(\mathbf{r} - \mathbf{r}^*)\|^2 + \mu\|\mathbf{r} - \mathbf{r}^*\|^2. \end{aligned}$$

But since $\tilde{\mathbf{w}}(\mathbf{r}) = \tilde{H}^T \mathbf{r}$ and $\tilde{\mathbf{w}}^* = \tilde{H}^T \mathbf{r}^*$, we get

$$\begin{aligned} \|\tilde{\mathbf{w}}(\mathbf{r}) - \tilde{\mathbf{w}}^*\|^2 &\leq \|\mathbf{q}\| \cdot \|\mathbf{r} - \mathbf{r}^*\| - \mu \|\mathbf{r} - \mathbf{r}^*\|^2 \\ &= \frac{\|\mathbf{q}\|^2}{4\mu} - \mu \left(\|\mathbf{r} - \mathbf{r}^*\| - \frac{\|\mathbf{q}\|}{2\mu} \right)^2 \\ &\leq \frac{\|\mathbf{q}\|^2}{4\mu}. \end{aligned}$$

This implies that $\|\tilde{\mathbf{w}}(\mathbf{r}) - \tilde{\mathbf{w}}^*\| \leq \frac{\|\mathbf{q}\|}{2\sqrt{\mu}} \leq \frac{\sqrt{\mu}}{2} \epsilon$. Now, observe that if $\gamma - r_i > \epsilon$, then $\epsilon < \gamma$. Hence,

$$\|\tilde{\mathbf{w}}(\mathbf{r}) - \tilde{\mathbf{w}}^*\| < \frac{\sqrt{\mu}\gamma}{2} = \frac{64\sqrt{n/128}}{2n} = \sqrt{\frac{8}{n}}.$$

But from Corollary 8, for all $i \in \mathcal{C}$, $\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K$, and

$$\begin{aligned} \mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) &= \mathbf{c}_i^T \mathbf{w}^* + t^* + \begin{pmatrix} \mathbf{c}_i \\ 1 \end{pmatrix}^T \begin{pmatrix} \mathbf{w}(\mathbf{r}) - \mathbf{w}^* \\ t(\mathbf{r}) - t^* \end{pmatrix} \\ &\geq 1 - K - \|\tilde{\mathbf{c}}_i\| \cdot \|\tilde{\mathbf{w}}(\mathbf{r}) - \tilde{\mathbf{w}}^*\| \\ &> 1 - K - \sqrt{\frac{8(R^2 + 1)}{n}} \\ &> \frac{1}{2} - \sqrt{\frac{8((6/5)^2 + 1)}{79}} \\ &= \frac{1}{2} - \sqrt{\frac{488}{1975}} > 0, \end{aligned}$$

where we used that $K = \max\{K_c, K_d\} + 2\rho\xi \geq 2\rho\xi$ and $\xi \geq \frac{1}{\bar{\sigma} - \rho}$. With the assumption $K < 1/2$, this implies $\frac{2\rho}{\bar{\sigma} - \rho} < 1/2$. Rearranging gives $4\rho < \bar{\sigma} - \rho$ and

$$\rho < \frac{\bar{\sigma}}{5} \leq \frac{1}{5},$$

which ultimately yields the bound $R = \max\{\sigma_1, \sigma_2\} + \rho < 6/5$.

The same argument, applied to the signed margins $-\mathbf{d}_i^T \mathbf{w}(\mathbf{r}) - t(\mathbf{r})$, proves the result for $i \in \mathcal{D}$. ■

3.3 Data-dependent iteration complexity

In this section we derive two iteration bounds. The first gives the number of FISTA iterations required to guarantee separation of the well-classified points with a prescribed positive margin. The second bounds the number of iterations until the proposed stopping condition is satisfied. Both results follow from the standard convergence theory of FISTA. The bounds (25) and (26), derived in Appendix B, give the corresponding distance and iteration estimates. We first specialize these estimates to our initialization and parameter choices.

Theorem 10 *Let $\epsilon > 0$ and let $\mathbf{r}_k$ be the iterates generated by the strongly convex variant of FISTA applied to (5). Suppose $\mathbf{r}_0 = \mathbf{0}$, $\gamma = 64/n$, and $\mu = n/128$. Then*

$$\|\mathbf{r}_k - \mathbf{r}^*\| \leq \frac{128}{\sqrt{n}} \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{k/2},$$

and $\|\mathbf{r}_k - \mathbf{r}^*\| \leq \epsilon$ whenever

$$k \geq 2\sqrt{\kappa} \log \left(\frac{128}{\sqrt{n}\epsilon} \right).$$

The proof is given in Appendix B.

A priori complexity of separating well-classified points. Suppose $K < 1/2$. By Corollary 8, $\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K$ for all $i \in \mathcal{C}$ and $-\mathbf{d}_i^T \mathbf{w}^* - t^* \geq 1 - K$ for all $i \in \mathcal{D}$. Let $\bar{K} \in (K, 1)$, and consider an arbitrary $\mathbf{r} = (\mathbf{u}, \mathbf{v}) \in \Omega$ with corresponding classifier

$$\tilde{\mathbf{w}}(\mathbf{r}) = \tilde{H}^T \mathbf{r} = \begin{pmatrix} \mathbf{w}(\mathbf{r}) \\ t(\mathbf{r}) \end{pmatrix},$$

where $\mathbf{w}(\mathbf{r}) = C^T \mathbf{u} - D^T \mathbf{v}$, $t(\mathbf{r}) = \mathbf{e}^T \mathbf{u} - \mathbf{e}^T \mathbf{v}$. Define

$$\Delta_{\text{sep}} := \frac{\bar{K} - K}{\sqrt{R^2 + 1} \|\tilde{H}\|}. \quad (18)$$

The following theorem shows that any approximate dual solution satisfying $\|\mathbf{r} - \mathbf{r}^*\| \leq \Delta_{\text{sep}}$ yields a classifier $(\mathbf{w}(\mathbf{r}), t(\mathbf{r}))$ that separates the well-classified points with signed margin at least $1 - \bar{K}$.

Theorem 11 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$, let $\bar{K} \in (K, 1)$, and let $\mathbf{r} \in \Omega$ satisfy $\|\mathbf{r} - \mathbf{r}^*\| \leq \Delta_{\text{sep}}$. Then the classifier $(\mathbf{w}(\mathbf{r}), t(\mathbf{r}))$ separates the well-classified points:*

$$\begin{aligned} \mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) &\geq 1 - \bar{K}, \quad \forall i \in \mathcal{C}, \\ \mathbf{d}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) &\leq -1 + \bar{K}, \quad \forall i \in \mathcal{D}. \end{aligned}$$

Proof Under the assumption that $\|\mathbf{c}_i\| \leq R$ for all i and $\|\mathbf{d}_i\| \leq R$ for all i , we have

$$\begin{aligned} \|\tilde{\mathbf{c}}_i\| &\leq \sqrt{R^2 + 1}, \quad \forall i \in \mathcal{C}, \\ \|\tilde{\mathbf{d}}_i\| &\leq \sqrt{R^2 + 1}, \quad \forall i \in \mathcal{D}. \end{aligned}$$

Since $\|\mathbf{r} - \mathbf{r}^*\| \leq \Delta_{\text{sep}}$,

$$\begin{aligned} \|\tilde{\mathbf{w}}(\mathbf{r}) - \tilde{\mathbf{w}}^*\| &= \|\tilde{H}^T (\mathbf{r} - \mathbf{r}^*)\| \\ &\leq \|\tilde{H}\| \cdot \|\mathbf{r} - \mathbf{r}^*\| \\ &\leq \|\tilde{H}\| \cdot \Delta_{\text{sep}} \\ &\leq \frac{\bar{K} - K}{\sqrt{R^2 + 1}}. \end{aligned}$$

Fix some $i \in \mathcal{C}$. We have

$$\begin{aligned}
 \mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) &= \tilde{\mathbf{c}}_i^T \tilde{\mathbf{w}}(\mathbf{r}) \\
 &= \tilde{\mathbf{c}}_i^T \tilde{\mathbf{w}}^* + \tilde{\mathbf{c}}_i^T (\tilde{\mathbf{w}}(\mathbf{r}) - \tilde{\mathbf{w}}^*) \\
 &\geq (1 - K) - \|\tilde{\mathbf{c}}_i\| \cdot \|\tilde{\mathbf{w}}(\mathbf{r}) - \tilde{\mathbf{w}}^*\| \\
 &\geq (1 - K) - \frac{\left(\sqrt{R^2 + 1}\right) (\bar{K} - K)}{\sqrt{R^2 + 1}} \\
 &= 1 - \bar{K}.
 \end{aligned}$$

By similar steps, $-\mathbf{d}_i^T \mathbf{w}(\mathbf{r}) - t(\mathbf{r}) \geq 1 - \bar{K}$. ■

Corollary 12 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$. Let $\bar{K} \in (K, 1)$, and consider FISTA initialized at $\mathbf{r}_0 = \mathbf{0}$, with $\gamma = 64/n$ and $\mu = n/128$. If*

$$k \geq 2\sqrt{\kappa} \log \left(\frac{128(R^2 + 1)}{\bar{K} - K} \right),$$

then the classifier given by $\tilde{H}^T \mathbf{r}_k$ separates the well-classified points with signed margin at least $1 - \bar{K}$.

Proof As in the proof of Theorem 1, $\|\tilde{H}\| \leq \sqrt{n(R^2 + 1)}$. Thus,

$$\Delta_{\text{sep}} \geq \frac{\bar{K} - K}{\sqrt{n(R^2 + 1)}}.$$

Hence, we may apply Theorem 10 to the quantity on the right hand side above to obtain a bound on the number of iterations required to achieve the smaller accuracy

$$\|\mathbf{r}_k - \mathbf{r}^*\| \leq \frac{\bar{K} - K}{\sqrt{n(R^2 + 1)}}.$$

The conclusion follows from Theorem 11. ■

Complexity of the proposed stopping condition. We now give an upper bound on the number of iterations required before the proposed stopping condition is guaranteed to hold. Suppose $K < 1/2$ and fix some $i \in \mathcal{C}$. To show that the coordinate corresponding to $\mathbf{c}_i$ triggers the stopping condition, it suffices to determine the number of iterations before $\gamma - r_{k,i} > \epsilon_k^{\text{FISTA}}$. Since Corollary 8 and Lemma 5 imply $u_i^* \leq 2K\gamma$, we may use Theorem 2 to write

$$\gamma - r_{k,i} \geq \gamma - u_i^* - |r_{k,i} - u_i^*| \geq (1 - 2K)\gamma - \epsilon_k^{\text{FISTA}},$$

where $\epsilon_k^{\text{FISTA}} = \kappa \|\mathbf{y}_{k-1} - \mathbf{r}_k\|$. Thus, $\epsilon_k^{\text{FISTA}} < (1 - 2K)\gamma/2$ guarantees that $\mathbf{c}_i$ satisfies the stopping condition. An identical argument can be made for $\mathbf{d}_i$ when $i \in \mathcal{D}$.

Under the data model of Section 3.1, the choices $\gamma = 64/n$ and $\mu = n/128$ give $\kappa = L/\mu < 641$. Consequently, for fixed $K < 1/2$, the iteration bound in the following theorem grows logarithmically with n .

Theorem 13 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$. Consider FISTA initialized at $\mathbf{r}_0 = \mathbf{0}$, with $\gamma = 64/n$ and $\mu = n/128$. The stopping condition (17) is guaranteed to be satisfied within the first k iterations whenever*

$$k > 2 + 2\sqrt{\kappa} \log \left(\frac{16\kappa\sqrt{n}}{1-2K} \right).$$

Proof For $k \geq 2$, (13) gives

$$\begin{aligned} \mathbf{y}_{k-1} - \mathbf{r}_k &= \mathbf{r}_{k-1} + \beta(\mathbf{r}_{k-1} - \mathbf{r}_{k-2}) - \mathbf{r}_k \\ &= (1 + \beta)(\mathbf{r}_{k-1} - \mathbf{r}^*) + \beta(\mathbf{r}^* - \mathbf{r}_{k-2}) + (\mathbf{r}^* - \mathbf{r}_k). \end{aligned}$$

Since $\beta \geq 0$,

$$\epsilon_k^{\text{FISTA}} \leq \kappa(1 + \beta)\|\mathbf{r}_{k-1} - \mathbf{r}^*\| + \kappa\beta\|\mathbf{r}_{k-2} - \mathbf{r}^*\| + \kappa\|\mathbf{r}_k - \mathbf{r}^*\|.$$

By Theorem 10 along with the fact that $\beta < 1$ and $1 - 1/\sqrt{\kappa} \in (0, 1)$, we may bound the right hand side above to obtain

$$\begin{aligned} \epsilon_k^{\text{FISTA}} &\leq \frac{128\kappa}{\sqrt{n}} \left((1 + \beta) \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{\frac{k-1}{2}} + \beta \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{\frac{k-2}{2}} + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{\frac{k}{2}} \right) \\ &\leq \frac{4 \cdot 128\kappa}{\sqrt{n}} \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{\frac{k-2}{2}}. \end{aligned}$$

Since $1 - x \leq e^{-x}$ for all x , we may again upper bound $\epsilon_k^{\text{FISTA}}$:

$$\epsilon_k^{\text{FISTA}} \leq \frac{512\kappa}{\sqrt{n}} \exp \left(\frac{2-k}{2\sqrt{\kappa}} \right).$$

Using $(1 - 2K)\gamma/2 = (1 - 2K)32/n$ and the assumption that $K < 1/2$, any

$$k > 2 + 2\sqrt{\kappa} \log \left(\frac{16\sqrt{n}\kappa}{1-2K} \right)$$

gives $\epsilon_k^{\text{FISTA}} < (1 - 2K)\gamma/2$. It follows from the preceding argument that $\gamma - r_{k,i} > \epsilon_k^{\text{FISTA}}$ for such k , so the stopping condition must be satisfied. $\blacksquare$

4 Coordinate descent

The separation guarantees developed above are not specific to FISTA. Under our data model, any first-order method with a computable error bound $\|\mathbf{r}_k - \mathbf{r}^*\| \leq \epsilon_k$ produces a classifier that separates the well-classified points once $\epsilon_k \leq \Delta_{\text{sep}}$. We now consider a dual coordinate descent method of the type used by LIBLINEAR (Hsieh et al., 2008). Unlike FISTA, which updates all dual coordinates simultaneously through a projected-gradient step, coordinate descent updates one dual coordinate at a time while maintaining an up-to-date primal vector $\tilde{\mathbf{w}}$. We derive a computable error bound for coordinate descent applied to (5), and compare the resulting stopping condition with the one used by LIBLINEAR.

Negating the objective of (5), we obtain the equivalent minimization problem

$$\min_{\mathbf{r} \in \Omega} h(\mathbf{r}) := \frac{1}{2} \mathbf{r}^T Q \mathbf{r} - \mathbf{e}^T \mathbf{r}, \quad (19)$$

where $Q := \tilde{H} \tilde{H}^T + \mu I$. We consider coordinate descent applied to this problem with $\mu > 0$. The coordinate updates have the same form as those described in Hsieh et al. (2008) and used by LIBLINEAR's solver type 3, but that solver treats the unperturbed problem with $\mu = 0$. The error bounds and classification guarantees derived below concern the perturbed problem ($\mu > 0$). We return to the unperturbed LIBLINEAR setting in Section 4.1.

We may define the composite objective

$$F_{\text{CD}}(\mathbf{r}) := h(\mathbf{r}) + i_{\Omega}(\mathbf{r}), \quad (20)$$

so that (19) is equivalent to $\min F_{\text{CD}}(\mathbf{r})$. Since $Q \succeq \mu I$, the function h , and consequently F_{CD} , is μ -strongly convex.

Let $G(\mathbf{r}) := \nabla h(\mathbf{r}) = Q\mathbf{r} - \mathbf{e}$. Coordinate descent at some iterate $\mathbf{r}$ updates coordinate i by amount d so that $\mathbf{r}^+ = \mathbf{r} + d\mathbf{e}_i$. Since Q is symmetric,

$$\begin{aligned} h(\mathbf{r} + d\mathbf{e}_i) &= \frac{1}{2} (\mathbf{r} + d\mathbf{e}_i)^T Q (\mathbf{r} + d\mathbf{e}_i) - \mathbf{e}^T (\mathbf{r} + d\mathbf{e}_i) \\ &= h(\mathbf{r}) + d(Q\mathbf{r} - \mathbf{e})_i + \frac{1}{2} Q_{ii} d^2 \\ &= h(\mathbf{r}) + G_i(\mathbf{r})d + \frac{1}{2} Q_{ii} d^2. \end{aligned}$$

Hence, finding the coordinate descent update d is equivalent to solving the following problem with respect to d

$$\min \left\{ G_i(\mathbf{r})d + \frac{1}{2} Q_{ii} d^2 : d \in [-r_i, \gamma - r_i] \right\}.$$

It is not difficult to see that this yields the following closed-form update

$$r_i^+ = \text{proj}_{[0, \gamma]} \left(r_i - \frac{G_i(\mathbf{r})}{Q_{ii}} \right).$$

The algorithm performs the coordinate update without explicitly forming Q . Denote $\tilde{\mathbf{w}}(\mathbf{r}) = \tilde{H}^T \mathbf{r}$ and the i th row of $\tilde{H}$ by $\tilde{\mathbf{h}}_i$. We may write

$$G_i(\mathbf{r}) = \tilde{\mathbf{h}}_i^T \tilde{\mathbf{w}}(\mathbf{r}) + \mu r_i - 1,$$

and $Q_{ii} = \|\tilde{\mathbf{h}}_i\|^2 + \mu$. If in the algorithm coordinate i changes by d , the primal vector $\tilde{\mathbf{w}}$ can be updated as

$$\tilde{\mathbf{w}}(\mathbf{r} + d\mathbf{e}_i) = \tilde{\mathbf{w}}(\mathbf{r}) + d\tilde{\mathbf{h}}_i.$$

Thus, the cost of computing $G_i(\mathbf{r})$ and updating $\mathbf{r}$ and $\tilde{\mathbf{w}}$ is proportional to the number of nonzero entries in the corresponding training data point.

We next derive a computable error bound for the coordinate descent iterates. For any $\mathbf{r} \in \Omega$, define $\mathbf{p}(\mathbf{r}) \in \mathbb{R}^n$ coordinate-wise by

$$p_i(\mathbf{r}) := \begin{cases} \min\{G_i(\mathbf{r}), 0\}, & r_i = 0, \\ G_i(\mathbf{r}), & r_i \in (0, \gamma), \\ \max\{G_i(\mathbf{r}), 0\}, & r_i = \gamma. \end{cases} \quad (21)$$

The vector $\mathbf{p}(\mathbf{r})$ is the minimum-norm subgradient of F_{CD} at $\mathbf{r}$. Hence, $\mathbf{p}(\mathbf{r}) = \mathbf{0}$ iff $\mathbf{r}$ is optimal for (19). The minimum-norm subgradient provides a bound on the distance to the optimal solution, thus providing a computable error bound as desired.

Theorem 14 *Suppose $\mu > 0$. For every $\mathbf{r} \in \Omega$,*

$$\|\mathbf{r} - \mathbf{r}^*\| \leq \frac{1}{\mu} \|\mathbf{p}(\mathbf{r})\|.$$

Proof Since F_{CD} is μ -strongly convex, its subdifferential is μ -strongly monotone:

$$\langle \mathbf{p}(\mathbf{r}) - \mathbf{p}(\mathbf{r}^*), \mathbf{r} - \mathbf{r}^* \rangle \geq \mu \|\mathbf{r} - \mathbf{r}^*\|^2.$$

Using Cauchy-Schwarz and the fact that $\mathbf{p}(\mathbf{r}^*) = \mathbf{0}$, one obtains the inequality

$$\mu \|\mathbf{r} - \mathbf{r}^*\| \leq \|\mathbf{p}(\mathbf{r})\|.$$

■

We can now define the computable error bound for outer iteration k as

$$\epsilon_k^{\text{CD}} := \frac{\|\mathbf{p}(\mathbf{r}_k)\|}{\mu}.$$

Theorem 14 gives $\|\mathbf{r}_k - \mathbf{r}^*\| \leq \epsilon_k^{\text{CD}}$. Consequently, under the data model, whenever $\epsilon_k^{\text{CD}} \leq \Delta_{\text{sep}}$, Theorem 11 guarantees that the classifier associated with $\mathbf{r}_k$ separates the well-classified points with signed margin $1 - \bar{K}$. As in Theorem 3, we show that if $\gamma - r_{k,i} > \epsilon_k^{\text{CD}}$, the corresponding point is correctly classified by both the optimal classifier and the returned classifier.

Theorem 15 *Let $\mathbf{r} \in \Omega$ and suppose $\mu\gamma \in (0, 1)$. If $\gamma - r_i > \epsilon^{\text{CD}} := \|\mathbf{p}(\mathbf{r})\|/\mu$ for some $i \in [n]$, then*

$$\theta_i(\mathbf{r}) > 1 - \mu\gamma > 0, \quad \theta_i(\mathbf{r}^*) > 1 - \mu\gamma > 0.$$

In particular, the corresponding point is correctly classified by both classifiers $\tilde{H}^T \mathbf{r}$ and $\tilde{H}^T \mathbf{r}^*$.

Proof Since $\mathbf{p}(\mathbf{r})$ is a subgradient of F_{CD} at $\mathbf{r}$, there exists $\mathbf{z} \in \mathcal{N}_{\Omega}(\mathbf{r})$ such that

$$\mathbf{p}(\mathbf{r}) = \tilde{H} \tilde{H}^T \mathbf{r} + \mu \mathbf{r} - \mathbf{e} + \mathbf{z}.$$

Recall $\boldsymbol{\theta}(\mathbf{r}) = \tilde{H} \tilde{H}^T \mathbf{r}$, so $\boldsymbol{\theta}(\mathbf{r}) = \mathbf{e} - \mu \mathbf{r} + \mathbf{p}(\mathbf{r}) - \mathbf{z}$. The stopping condition implies $r_i < \gamma$, which by the same reasoning as in the proof of Theorem 3 implies $z_i \leq 0$. Therefore,

$$\begin{aligned} \theta_i(\mathbf{r}) &= 1 - \mu r_i + p_i(\mathbf{r}) - z_i \\ &\geq 1 - \mu r_i - \|\mathbf{p}(\mathbf{r})\| \\ &= 1 - \mu r_i - \mu \epsilon^{\text{CD}} \\ &> 1 - \mu\gamma > 0. \end{aligned}$$

Moreover, the error bound $\|\mathbf{r} - \mathbf{r}^*\| \leq \epsilon^{\text{CD}}$ gives

$$r_i^* \leq r_i + |r_i - r_i^*| \leq r_i + \epsilon^{\text{CD}} < \gamma,$$

while optimality means there exists $\mathbf{z}^* \in \mathcal{N}_\Omega(\mathbf{r}^*)$ such that

$$\tilde{H}\tilde{H}^T\mathbf{r}^* + \mu\mathbf{r}^* - \mathbf{e} + \mathbf{z}^* = \mathbf{0}.$$

From $r_i^* < \gamma$ and $\mathbf{z}^* \in \mathcal{N}_\Omega(\mathbf{r}^*)$, $z_i^* \leq 0$, and

$$\begin{aligned} \theta_i(\mathbf{r}^*) &= 1 - \mu r_i^* - z_i^* \\ &\geq 1 - \mu r_i^* \\ &> 1 - \mu\gamma > 0. \end{aligned}$$

■

We may also prove that under the data model, one point satisfying the stopping condition implies all well-classified points must be separated by the returned classifier.

Theorem 16 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$, $\gamma = 64/n$, $\mu = n/128$, and $n \geq 79$. Let $\mathbf{r} \in \Omega$ and $\tilde{\mathbf{w}}(\mathbf{r}) = (\mathbf{w}(\mathbf{r}), t(\mathbf{r})) = \tilde{H}^T\mathbf{r}$. If $\gamma - r_i > \epsilon^{\text{CD}}$ for some $i \in [n]$, then*

$$\mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) > 0, \quad \forall i \in \mathcal{C},$$

and

$$-\mathbf{d}_i^T \mathbf{w}(\mathbf{r}) - t(\mathbf{r}) > 0, \quad \forall i \in \mathcal{D}.$$

Proof Recall $\tilde{\mathbf{w}}^* = (\mathbf{w}^*, t^*) = \tilde{H}^T\mathbf{r}^*$ is the optimal classifier. By the proof of the previous theorem, there exist $\mathbf{z} \in \mathcal{N}_\Omega(\mathbf{r})$ and $\mathbf{z}^* \in \mathcal{N}_\Omega(\mathbf{r}^*)$ such that

$$\mathbf{p}(\mathbf{r}) = \tilde{H}\tilde{H}^T\mathbf{r} + \mu\mathbf{r} - \mathbf{e} + \mathbf{z}, \quad 0 = \tilde{H}\tilde{H}^T\mathbf{r}^* + \mu\mathbf{r}^* - \mathbf{e} + \mathbf{z}^*.$$

Subtracting the second equation from the first, we obtain

$$\mathbf{p}(\mathbf{r}) = \tilde{H}\tilde{H}^T(\mathbf{r} - \mathbf{r}^*) + \mu(\mathbf{r} - \mathbf{r}^*) + \mathbf{z} - \mathbf{z}^*.$$

Using the fact that $\mathcal{N}_\Omega$ is monotone, taking the inner product of the above with $\mathbf{r} - \mathbf{r}^*$ gives

$$\|\mathbf{p}(\mathbf{r})\| \cdot \|\mathbf{r} - \mathbf{r}^*\| \geq \mathbf{p}(\mathbf{r})^T(\mathbf{r} - \mathbf{r}^*) \geq \|\tilde{H}^T(\mathbf{r} - \mathbf{r}^*)\|^2 + \mu\|\mathbf{r} - \mathbf{r}^*\|^2.$$

Thus,

$$\|\tilde{H}^T(\mathbf{r} - \mathbf{r}^*)\|^2 \leq \frac{\|\mathbf{p}(\mathbf{r})\|^2}{4\mu} - \mu \left(\|\mathbf{r} - \mathbf{r}^*\| - \frac{\|\mathbf{p}(\mathbf{r})\|}{2\mu} \right)^2 \leq \frac{\|\mathbf{p}(\mathbf{r})\|^2}{4\mu},$$

and

$$\|\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^*\| = \|\tilde{H}^T(\mathbf{r} - \mathbf{r}^*)\| \leq \frac{\|\mathbf{p}(\mathbf{r})\|}{2\sqrt{\mu}}.$$

Since the stopping condition implies $\epsilon^{\text{CD}} < \gamma$ and the definition of ϵ^{CD} gives $\|\mathbf{p}(\mathbf{r})\| = \mu\epsilon^{\text{CD}}$, we obtain

$$\|\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^*\| < \frac{\mu\gamma}{2\sqrt{\mu}} = \frac{\sqrt{\mu}\gamma}{2} = \sqrt{\frac{8}{n}}.$$

The remainder of the proof follows directly from the proof of Theorem 9. $\blacksquare$

There is an important distinction between evaluating our stopping condition for coordinate descent and for FISTA. During an outer iteration of coordinate descent, the subgradient entries are evaluated at different intermediate iterates, so they do not, in general, give the components of $\mathbf{p}(\mathbf{r}_k)$ at the end of the iteration. Computing ϵ_k^{CD} therefore requires an additional computation of the full subgradient $\mathbf{p}(\mathbf{r}_k)$. Since the coordinate descent algorithm maintains $\tilde{\mathbf{w}} = \tilde{H}^T \mathbf{r}$, each component of the subgradient can be computed from (21) in $\mathcal{O}(\text{nnz}(\tilde{\mathbf{h}}_i))$ operations. Thus, computing $\mathbf{p}(\mathbf{r}_k)$ itself requires $\mathcal{O}(\text{nnz}(\tilde{H}))$ operations, which is of the same order as an outer iteration of coordinate descent.

In practice, the full subgradient need not be computed after every outer iteration. Instead, inexpensive optimality checks of the same form as those used by LIBLINEAR (see Section 4.1) can be used to determine when to compute the full subgradient and evaluate our stopping condition.

4.1 A limitation of the LIBLINEAR stopping condition

In this section, we set $\mu = 0$ and consider the unperturbed problem solved by LIBLINEAR's solver type 3. The definitions of G and $\mathbf{p}$ above are used with $\mu = 0$. The LIBLINEAR solver does not use $\|\mathbf{p}(\mathbf{r})\|$ as a stopping condition. Before each outer iteration, LIBLINEAR shuffles the coordinate order. Suppose during outer iteration k the coordinates are visited in the order $i_1, \dots, i_n$, and let $\mathbf{r}_k^{s-1}$ denote the iterate immediately before coordinate i_s is updated in outer loop k . LIBLINEAR computes $p_{i_s}(\mathbf{r}_k^{s-1})$ and then updates the coordinate i_s , producing the next intermediate iterate $\mathbf{r}_k^s$. The following quantities are recorded during outer iteration k :

$$M^k := \max_{s \in [n]} p_{i_s}(\mathbf{r}_k^{s-1}), \quad m^k := \min_{s \in [n]} p_{i_s}(\mathbf{r}_k^{s-1}).$$

LIBLINEAR stops when $M^k - m^k$, $|M^k|$, and $|m^k|$ are all below some user-defined tolerance. Because $\mathbf{r}$ is modified after each coordinate update, the subgradient entries used to compute M^k and m^k are evaluated at different intermediate values of $\mathbf{r}$. As a result, M^k and m^k cannot in general be interpreted as the maximum and minimum components of $\mathbf{p}(\mathbf{r})$ at any single iterate $\mathbf{r}$.

The fact that LIBLINEAR does not compute a full subgradient but instead computes the entries at different intermediate iterates allows for the algorithm's stopping condition to be satisfied despite the current classifier being poor.

Set

$$q := \sqrt{\frac{3d}{2} - 1},$$

and consider the dataset defined by the four points:

$$\mathbf{c}_1 = \mathbf{e}, \quad \mathbf{d}_1 = \frac{1}{q}\mathbf{e}, \quad \mathbf{d}_2 = -\mathbf{e}, \quad \mathbf{d}_3 = -\frac{1}{q}\mathbf{e}.$$

The sample mean of every feature is zero, while the sample variance of each feature is

$$\frac{1}{3} \left(1 + \frac{1}{q^2} + 1 + \frac{1}{q^2} \right) = \frac{2}{3} \left(1 + \frac{1}{q^2} \right) = \frac{d}{q^2},$$

since $q^2 + 1 = 3d/2$. Dividing each feature by its standard deviation gives the scaled points

$$\bar{\mathbf{c}}_1 = \frac{q}{\sqrt{d}}\mathbf{e}, \quad \bar{\mathbf{d}}_1 = \frac{1}{\sqrt{d}}\mathbf{e}, \quad \bar{\mathbf{d}}_2 = -\frac{q}{\sqrt{d}}\mathbf{e}, \quad \bar{\mathbf{d}}_3 = -\frac{1}{\sqrt{d}}\mathbf{e}.$$

These are the points LIBLINEAR sees after scaling by the standard deviation, as is done in our numerical experiments. Note that all points lie on the line $x_1 = \dots = x_d$, so this example is a dimension d generalization of separating the numbers $1, 1/\sqrt{2}, -1, -1/\sqrt{2}$ on the real line. Interestingly, the difference between $\bar{\mathbf{c}}_1$ and $\bar{\mathbf{d}}_1$ grows as the dimension increases:

$$\|\bar{\mathbf{c}}_1 - \bar{\mathbf{d}}_1\| = q - 1 = \mathcal{O}(\sqrt{d}).$$

Using $\gamma = 64/n = 16$ and LIBLINEAR's solver type 3, suppose that the coordinates corresponding to $\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_2, \mathbf{c}_1$ are visited in order during the first outer loop, and then in the second outer loop visited in the order corresponding to $\mathbf{c}_1, \mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_2$. Starting at LIBLINEAR's default starting point $\mathbf{r} = \mathbf{0}$, the subgradient entries computed in order by LIBLINEAR in the second sweep are

$$0, \quad -\frac{2(q+1)}{q^2+1}, \quad \frac{2(q-1)}{q^2+1}, \quad 0.$$

At the end of the second sweep, $\tilde{\mathbf{w}} = (\mathbf{0}, -1)$. This classifier places each point in the $\mathbf{d}$ -class, misclassifying $\mathbf{c}_1$. Observe that

$$M^k = \frac{2(q-1)}{q^2+1}, \quad m^k = -\frac{2(q+1)}{q^2+1},$$

and

$$M^k - m^k = \frac{4q}{q^2+1}.$$

Thus, for any positive tolerance, there exists a sufficiently large d where LIBLINEAR stops after the second outer iteration with a classifier that places all points in the same class. For the default tolerance used in LIBLINEAR of 0.1, the algorithm stops after two outer iterations with the bad classifier $\tilde{\mathbf{w}} = (\mathbf{0}, -1)$ for all $d \geq 1066$. This is just one example of this phenomenon. We explore premature termination of LIBLINEAR on this dataset numerically in Section 5.3.

5 Implementation and numerical experiments

Our numerical experiments address two questions. First, we compare two FISTA implementations with standard linear SVM solvers on benchmark datasets. Then, we examine two aspects of robustness that arise from our analysis: the effect of regularizing the intercept in the H-SVM formulation, and examples of where the LIBLINEAR stopping condition can fail.

Theorem 3 implies that any coordinate satisfying

$$\gamma - r_{k+1,i} > \epsilon_{k+1}^{\text{FISTA}},$$

Algorithm 1 Robustly terminated FISTA for the perturbed H-SVM dual

Input: Smoothness constant L , strong-concavity parameter μ , misclassification penalty parameter γ , matrix of $\mathbf{c}$ -points $C \in \mathbb{R}^{j \times d}$, matrix of $\mathbf{d}$ -points $D \in \mathbb{R}^{l \times d}$, stopping parameter $p \in [0, 1]$, maximum number of iterations **maxit**, gradient function $\nabla g : \mathbb{R}^n \rightarrow \mathbb{R}^n$, where $n := j + l$.

Initialize $\mathbf{y}_0 = \mathbf{r}_0 = \mathbf{0} \in \mathbb{R}^n$, $\kappa = L/\mu$, $T = \emptyset$.

for $k = 0$ **to** **maxit** **do**

$$\mathbf{r}_{k+1} = \text{proj}_{\Omega} \left(\mathbf{y}_k + \frac{1}{L} \nabla g(\mathbf{y}_k) \right)$$

$$\mathbf{y}_{k+1} = \mathbf{r}_{k+1} + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right) (\mathbf{r}_{k+1} - \mathbf{r}_k)$$

$$\epsilon_{k+1}^{\text{FISTA}} \leftarrow \kappa \|\mathbf{y}_k - \mathbf{r}_{k+1}\|$$

$$T \leftarrow T \cup \{i \in [n] : \gamma - r_{k+1,i} > \epsilon_{k+1}^{\text{FISTA}}\}$$

if $|T| \geq \max\{1, pn\}$ **then**

break

end if

end for

$$\mathbf{u}_{k+1} \leftarrow \mathbf{r}_{k+1}(1 : j)$$

$$\mathbf{v}_{k+1} \leftarrow \mathbf{r}_{k+1}(j+1 : n)$$

$$\mathbf{w}_{k+1} \leftarrow C^T \mathbf{u}_{k+1} - D^T \mathbf{v}_{k+1}$$

$$t_{k+1} \leftarrow \mathbf{e}^T \mathbf{u}_{k+1} - \mathbf{e}^T \mathbf{v}_{k+1}$$

return $\mathbf{w}_{k+1}, t_{k+1}$

corresponds to a training point that is correctly classified by the optimal perturbed H-SVM classifier. Throughout the FISTA iterations, the algorithm can keep a list T_k of all such points:

$$T_k := \bigcup_{s=1}^k \{i \in [n] : \gamma - r_{s,i} > \epsilon_s^{\text{FISTA}}\}$$

In practice, one may use the number of certified points as a stopping condition. For some choice of $p \in [0, 1]$, stop once

$$|T_k| \geq \max\{1, pn\}.$$

Thus, $p = 0$ corresponds to stopping when the first training point is certified, while, for example, $p = 1/2$ requires at least half of the training points to have been certified. Recall, under our data model, with $K < 1/2$ and $n \geq 79$, Theorem 9 guarantees that stopping with $p = 0$ yields a classifier that separates all the well-classified points.

For FISTA applied to H-SVM, stopping with $\mathbf{r}_{k+1} = (\mathbf{u}_{k+1}, \mathbf{v}_{k+1})$ gives the classifier

$$\mathbf{w}_{k+1} = C^T \mathbf{u}_{k+1} - D^T \mathbf{v}_{k+1}, \quad t_{k+1} = \mathbf{e}^T \mathbf{u}_{k+1} - \mathbf{e}^T \mathbf{v}_{k+1}.$$

On the other hand, for the classical SVM, $\mathbf{w}_{k+1}$ is computed the same way, but t_{k+1} is the result of solving the unconstrained one-dimensional convex problem (36) given in Appendix C.3. In our implementation, we solve this problem by bisection.

5.1 Benchmark results

We evaluate FISTA applied to H-SVM (see Algorithm 1) and FISTA applied to the classical SVM against three widely used algorithms for solving SVM problems, namely Pegasos (Shalev-Shwartz et al., 2007), LIBLINEAR (Fan et al., 2008), and LIBSVM (Chang and Lin, 2011). Ten benchmark datasets¹ are chosen to span a range of sample sizes and feature dimensions, and to include both sparse and dense regimes. For each dataset we report test accuracy and training time in seconds, as well as characteristics of the dataset. The results are split into two tables, where n_{tr} and n_{te} denote the number of training and test examples respectively, d denotes the number of features, and Avg. nnz/ex. is an abbreviation for the average number of nonzero entries per example in the training set, computed as

$$\text{Avg. nnz/ex.} = \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \|\mathbf{h}_i\|_0,$$

where $\mathbf{h}_i$ denotes a single training example. Table 1 contains five standard benchmarks, including several sparse problems, while Table 2 contains five large-scale dense benchmarks. Within each row of the tables, boldface indicates the highest test accuracy and underlining indicates the shortest training time. Ties at the displayed precision receive the same formatting.

For each dataset, we used the official train/test split if possible. If no official split was available, we used a stratified random split, with 70% of the data for training and 30% for testing. All preprocessing parameters are computed using the training set and are then applied unchanged to the test set.

Let $X \in \mathbb{R}^{n_{\text{tr}} \times d}$ denote the training data matrix, and let $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ denote the vectors of training-set feature means and standard deviations. LIBLINEAR, LIBSVM, and Pegasos received as input $X \text{Diag}(\boldsymbol{\sigma})^{-1}$. We did not explicitly center the data, since explicit centering destroys sparsity and, in experiments on explicitly centered data, the three solvers performed worse than with scaling alone.

Our FISTA implementations additionally apply centering implicitly. We implicitly work with the standardized matrix Z or $\tilde{Z}$

$$Z := (X - \mathbf{e}\boldsymbol{\mu}^T) \text{Diag}(\boldsymbol{\sigma})^{-1}, \quad \tilde{Z} = \begin{pmatrix} Z & \mathbf{e} \end{pmatrix},$$

without ever forming Z or $\tilde{Z}$. For $\mathbf{v} = (\bar{\mathbf{v}}, v_0) \in \mathbb{R}^{d+1}$ and $\mathbf{u} \in \mathbb{R}^{n_{\text{tr}}}$,

$$\begin{aligned} \tilde{Z}\mathbf{v} &= X \text{Diag}(\boldsymbol{\sigma})^{-1} \bar{\mathbf{v}} + \mathbf{e} (v_0 - \boldsymbol{\mu}^T \text{Diag}(\boldsymbol{\sigma})^{-1} \bar{\mathbf{v}}), \\ \tilde{Z}^T \mathbf{u} &= \begin{pmatrix} \text{Diag}(\boldsymbol{\sigma})^{-1} X^T \mathbf{u} - \text{Diag}(\boldsymbol{\sigma})^{-1} \boldsymbol{\mu} (\mathbf{e}^T \mathbf{u}) \\ \mathbf{e}^T \mathbf{u} \end{pmatrix}. \end{aligned}$$

These products are efficient to compute, since $\text{Diag}(\boldsymbol{\sigma})^{-1} \bar{\mathbf{v}}$ amounts to the vector with entries $\bar{v}_j / \sigma_j$. Thus the cost of centering is absorbed into low-rank correction terms in the matrix-vector products, while preserving the storage advantages of the original sparse representation.

1. Datasets were accessed from <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>, except those denoted with an asterisk, which were accessed from <https://archive.ics.uci.edu/>.

After training, a classifier obtained in standardized coordinates is converted to the original feature coordinates. If $(\bar{\mathbf{w}}, \bar{t})$ denotes the classifier in standardized coordinates, then the equivalent classifier in the original coordinates is given by

$$\mathbf{w} = \text{Diag}(\boldsymbol{\sigma})^{-1} \bar{\mathbf{w}}, \quad t = \bar{t} - \boldsymbol{\mu}^T \text{Diag}(\boldsymbol{\sigma})^{-1} \bar{\mathbf{w}}.$$

Across all benchmarks, for Algorithm 1 we use $\gamma = 64/n_{\text{tr}}$ and $\mu = n_{\text{tr}}/128$. We choose `maxit` = 10^4 , but the maximum number of iterations is never reached on any of the experiments. FISTA is stopped when the first training point is certified, corresponding to $p = 0$. The strongly convex variant of FISTA (Beck, 2017, Section 10.7.7) is used throughout and we always initialize FISTA with starting point zero. We use $L = \mu + \|\tilde{Z}\|_2^2$ for H-SVM and $L = \mu + \|Z\|_2^2$ for classical SVM. We compute L using power iteration applied implicitly to $\tilde{Z}^T \tilde{Z}$ or $Z^T Z$, terminating when the residual falls below 10^{-10} .

The FISTA and Pegasos implementations were written in Julia, while LIBSVM and LIBLINEAR used the their Julia bindings². For the two FISTA methods, LIBLINEAR, and LIBSVM, we use the common coefficient

$$\gamma = C_{\text{LIBLINEAR}} = C_{\text{LIBSVM}} = \frac{64}{n_{\text{tr}}},$$

so that the misclassification penalty coefficient is aligned across the methods. For LIBLINEAR, we use solver type 3 with the bias parameter set to 1. Under this choice, LIBLINEAR augments the data exactly as we do in (5), making solver type 3 the closest standard LIBLINEAR comparison. The two problems are not identical because our perturbed H-SVM additionally contains the term $-\mu \|\mathbf{r}\|^2/2$. The official LIBLINEAR implementation uses a hard-coded allowance of 300 outer iterations. We do not change this, except in Section 5.3 for the MNIST inversion experiments. The default stopping tolerance in LIBLINEAR is 0.1, which we leave as is. For LIBSVM, we use a linear kernel and keep the default tolerance of 10^{-3} for its KKT-based stopping condition described in (Chang and Lin, 2011, Section 4.1.2). The remaining solver parameters are left at their default values. LIBSVM solves the classical soft-margin SVM formulation and is thus comparable to our SVM formulation as opposed to H-SVM.

We use our own sparse implementation of Pegasos. Under the parameterization used by Pegasos, the misclassification penalty $64/n_{\text{tr}}$ is equivalent to setting $\lambda = 1/64$ in the algorithm. The stepsize we use is $\eta_k = 1/(\lambda k)$ where k denotes the k th iteration, and projection of $\mathbf{w}$ onto the ball $\|\mathbf{w}\|_2 \leq 1/\sqrt{\lambda}$ is performed after each iteration. Our implementation of Pegasos additionally maintains an unregularized bias term b , updated alongside $\mathbf{w}$ on margin violations, as suggested in (Shalev-Shwartz et al., 2007, Section 6). The standard Pegasos algorithm stops after a prescribed number of iterations (Shalev-Shwartz et al., 2007, Section 2). In our implementation, we prescribe the number of epochs, using 200 for the datasets in Table 1 and 30 for those in Table 2.

All experiments in Table 1 and Table 2 completed, except LIBLINEAR on the `epsilon` dataset. Each attempt to run LIBLINEAR on `epsilon` caused the computer to run out of memory. “OOM” is used in Table 2 to denote that this experiment could not be completed due to insufficient memory. All experiments were run on a laptop with an 11th Gen Intel Core i7-1165G7 CPU (4 cores / 8 threads, 2.80 GHz) and 16 GB RAM.

2. See <https://github.com/JuliaML/LIBSVM.jl> and <https://github.com/JuliaML/LIBLINEAR.jl>

Table 1: Test accuracy and runtime (seconds) for sparse and standard benchmark problems. A $\dagger$ is used to denote when the maximum number of iterations were reached.

Dataset	n_{tr}	n_{te}	d	Avg. nnz/ex.	FISTA SVM		FISTA H-SVM		LIBLINEAR		LIBSVM		Pegasos	
					acc	time	acc	time	acc	time	acc	time	acc	time
a9a	32561	16281	123	13.87	0.8501	0.980	0.8501	0.455	0.8487	<u>0.414</u>	0.8490	34.609	0.8477	1.166
w8a	49749	14951	300	11.65	0.9864	1.677	0.9846	0.841	0.9855	<u>0.727</u>	0.9855	8.018	0.9837	2.345
cod-rna	59535	271617	8	8.00	0.9534	1.128	0.9539	0.443	0.9511	<u>0.419</u>	0.9532	32.968	0.9027	1.230
rcv1.binary	20242	677399	47236	74.05	0.9323	3.038	0.9328	3.958	0.9286	<u>1.338</u>	0.9300	434.213	0.9302	66.615
bank_marketing*	20183	8649	63	19.14	0.9054	1.049	0.9032	<u>0.516</u>	0.9016	1.181 $\dagger$	0.9025	22.349	0.9070	1.141

Table 2: Test accuracy and runtime (seconds) for large dense problems. “OOM” denotes when a solver ran out of memory.

Dataset	n_{tr}	n_{te}	d	Avg. nnz/ex.	FISTA SVM		FISTA H-SVM		LIBLINEAR		Pegasos	
					acc	time	acc	time	acc	time	acc	time
skin_nonskin	171540	73517	3	2.95	0.9367	3.993	0.9363	0.542	0.9457	0.454	0.9197	<u>0.441</u>
SUSY	4500000	500000	18	17.79	0.7816	117.814	0.7815	24.429	0.7795	<u>14.242</u>	0.7800	72.402
HEPMAS*	7000000	3500000	27	27.00	0.8160	175.976	0.8160	43.049	0.8161	<u>20.624</u>	0.8161	181.370
HIGGS	10500000	500000	28	25.79	0.6359	238.232	0.6365	53.336	0.6333	<u>38.787</u>	0.5943	389.022
epsilon	400000	100000	2000	2000.00	0.8972	970.471	0.8972	<u>875.985</u>	OOM		0.8970	879.224

5.2 Effect of intercept regularization in the H-SVM formulation

We conclude our numerical experiments by investigating the sensitivity of H-SVM to the representation of the feature vectors. Recall, H-SVM differs from the classical SVM by including the intercept term t in the regularized vector $\tilde{\mathbf{w}} = (\mathbf{w}, t)$. As a consequence, H-SVM is not invariant under translations of the data. We illustrate this effect on MNIST images (LeCun et al., 1998) with inverted black and white pixels.

Consider an MNIST image represented by a vector $\mathbf{x} \in [0, 1]^d$, and define its corresponding inverted image by

$$\bar{\mathbf{x}} := \mathbf{e} - \mathbf{x}.$$

Suppose $(\mathbf{w}, t)$ encodes an optimal classifier for the classical SVM (1). Then, one can construct an equivalent classifier $(-\mathbf{w}, t + \mathbf{e}^T \mathbf{w})$ for the inverted data:

$$\begin{aligned} (-\mathbf{w})^T \bar{\mathbf{x}} + (t + \mathbf{e}^T \mathbf{w}) &= (-\mathbf{w})^T (\mathbf{e} - \mathbf{x}) + t + \mathbf{e}^T \mathbf{w} \\ &= \mathbf{w}^T \mathbf{x} + t. \end{aligned}$$

Clearly, applying this transformation again would convert a classifier for the inverted data back to the original data. Since $\|-\mathbf{w}\| = \|\mathbf{w}\|$, the classical SVM formulation is invariant under the data inversion. However, in the case of H-SVM, $\|(\mathbf{w}, t)\|$ and $\|(-\mathbf{w}, t + \mathbf{e}^T \mathbf{w})\|$ need not be equal, so solving H-SVM on the inverted data need not give an equivalent classifier compared with solving on the original data.

Centering removes this issue. Let $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ denote the feature-wise mean and standard deviation of the original training data. Given an image $\mathbf{x}$, the centered and scaled feature vector is

$$\mathbf{z} := \text{Diag}(\boldsymbol{\sigma})^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

Table 3: Test accuracy and prediction agreement for classifiers trained on original and inverted MNIST representations.

Digits	Method	Black acc.	White acc.	Agreement
3 vs 5	FISTA SVM	0.9685	0.9532	0.9711
	FISTA H-SVM	0.9685	0.9522	0.9711
	LIBLINEAR	0.9690	0.7166	0.7266
1 vs 9	FISTA SVM	0.9949	0.9888	0.9911
	FISTA H-SVM	0.9949	0.9860	0.9865
	LIBLINEAR	0.9949	0.4706	0.4674
2 vs 7	FISTA SVM	0.9820	0.9796	0.9927
	FISTA H-SVM	0.9811	0.9796	0.9917
	LIBLINEAR	0.9816	0.9680	0.9777

For the inverted training data, the feature-wise mean and standard deviation are given by $\bar{\boldsymbol{\mu}} = \mathbf{e} - \boldsymbol{\mu}$ and $\bar{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$. Therefore,

$$\begin{aligned}
\text{Diag}(\bar{\boldsymbol{\sigma}})^{-1}(\bar{\mathbf{x}} - \bar{\boldsymbol{\mu}}) &= \text{Diag}(\boldsymbol{\sigma})^{-1}((\mathbf{e} - \mathbf{x}) - (\mathbf{e} - \boldsymbol{\mu})) \\
&= -\text{Diag}(\boldsymbol{\sigma})^{-1}(\mathbf{x} - \boldsymbol{\mu}) \\
&= -\mathbf{z}.
\end{aligned}$$

Hence, after centering and scaling, inversion simply reduces to negating the features. The transformation $(\mathbf{w}, t) \mapsto (-\mathbf{w}, t)$ then gives $(-\mathbf{w})^T(-\mathbf{z}) + t = \mathbf{w}^T\mathbf{z} + t$ and

$$\|(-\mathbf{w}, t)\| = \|(\mathbf{w}, t)\|.$$

Thus, solving H-SVM on the original and inverted data is equivalent if the data are centered. Note that centering is what gives the equivalence here, not scaling.

We test the effect of inversion on three binary MNIST classification problems: 3 versus 5, 1 versus 9, and 2 versus 7. For each pair of digits, we train separate classifiers on the original images and on their inversions. The same train/test split is used for both runs. We compare FISTA applied to the classical SVM with FISTA applied to H-SVM and LIBLINEAR. We deliberately do not center the data in these experiments, in order to observe the effect regularizing the intercept has on solving the inverted problem.

We report in Table 3 both test accuracy and prediction agreement for the two classifiers, where agreement is the proportion of test examples assigned the same label by the two classifiers. The settings for each algorithm remain unchanged from Section 5, except that the implicit centering used by FISTA is disabled.

Although the classical SVM problem is invariant under inversion, the implementation uses representation-dependent step sizes and early stopping, so the returned approximate classifiers need not have identical predictions. Classical FISTA has the highest or joint-highest agreement for all three digit pairs. FISTA applied to H-SVM also exhibits relatively small changes in accuracy and predictions. LIBLINEAR exhibits substantially larger changes for the 3 versus 5 and 1 versus 9 problems. However, all LIBLINEAR runs on the inverted data reached the 300-iteration limit without satisfying the stopping condition.

Table 4: Effect of increasing the LIBLINEAR iteration limit on the inverted MNIST data.

Digits	Method	Test acc.	Time (s)	Termination
3 vs 5	LIBLINEAR (5000)	0.9600	101.762	iteration limit
	LIBLINEAR (10000)	0.9632	179.267	iteration limit
	FISTA H-SVM	0.9522	53.828	stopping condition
1 vs 9	LIBLINEAR (5000)	0.9310	124.009	iteration limit
	LIBLINEAR (10000)	0.9953	235.922	iteration limit
	FISTA H-SVM	0.9860	59.017	stopping condition
2 vs 7	LIBLINEAR (5000)	0.9791	79.091	iteration limit
	LIBLINEAR (10000)	0.9626	125.935	iteration limit
	FISTA H-SVM	0.9796	57.939	stopping condition

These results describe the representation sensitivity of the algorithms under their stopping conditions. Without centering, H-SVM need not be invariant under inversion. Nevertheless, FISTA remains comparatively stable in these experiments, while LIBLINEAR shows larger changes and reaches its iteration limit on the inverted data. We investigate its termination behavior further in Section 5.3.

5.3 Robustness of stopping conditions

In this section, we illustrate two ways in which LIBLINEAR’s stopping condition can align poorly with the underlying classification task. First, the stopping condition can be overly conservative such that a useful classifier may emerge long before the stopping condition is satisfied. Second, the stopping condition can terminate prematurely despite the classification task not being solved adequately. The parameters used in LIBLINEAR for this section remain unchanged from our benchmark experiments, except for when we explicitly increase the maximum allowed iterations. Note that by default LIBLINEAR uses the tolerance 0.1 for its stopping conditions, which we keep unchanged.

Overly conservative termination on MNIST. The experiments in Section 5.2 revealed that LIBLINEAR reached its hard-coded 300 iteration limit on the inverted MNIST problems. In particular, the stopping condition with default tolerance 0.1 was never satisfied.

To investigate whether additional iterations improve classification accuracy, and caused LIBLINEAR to satisfy its stopping condition, we recompiled LIBLINEAR with iteration limits of 5000 and 10,000. Table 4 reports test accuracy, training time, and termination status compared with FISTA applied to H-SVM with its usual stopping condition.

Additional iterations substantially improved LIBLINEAR’s test accuracy for 3 versus 5 and 1 versus 9. At 10,000 iterations, its accuracy exceeded that of FISTA on both problems. For 2 versus 7, accuracy improved at 5000 iterations but decreased at 10,000, illustrating that additional optimization does not necessarily improve test accuracy. LIBLINEAR reached the iteration limit in every run without satisfying its stopping condition, even when its test accuracy matched or exceeded that of FISTA. Moreover, the 10,000 iteration runs took approximately 2 to 4 times as long as FISTA.

Table 5: Premature termination rate of LIBLINEAR on the four-point dataset introduced in Section 4.1.

d	LIBLINEAR failure rate (%)	FISTA accuracy (%)
500	0	100
1100	34.5	100
2000	34.2	100
5000	47.8	100

Premature termination before obtaining a good classifier. Recall the dataset described in Section 4.1. We gave a specific example of two subsequent coordinate orderings that makes LIBLINEAR stop in two iterations with a separator that misclassifies the only point in the $\mathbf{c}$ -class. This example is general in the sense that for any choice of tolerance, a sufficiently large d will make this sequence of coordinate orderings cause the algorithm to stop prematurely with the classifier that places $\mathbf{c}_1$ in the $\mathbf{d}$ -class incorrectly.

We compare LIBLINEAR and FISTA applied to H-SVM on this dataset for various choices of dimension d . Since LIBLINEAR shuffles the coordinate ordering in each outer iteration, we test the algorithm on seeds 1 to 1000.

Table 5 reports the percentage of LIBLINEAR runs that stop early and fail to classify $\mathbf{c}_1$ correctly. The table also reports the accuracy of the classifier returned by FISTA after it satisfies its stopping condition. For $d \geq 1100$, premature termination occurs in approximately 34% to 48% of the LIBLINEAR runs, whereas FISTA returns a separating classifier in every case. Thus, the failure exhibited in Section 4.1 is not restricted to one particular coordinate ordering, but in fact occurs with substantial frequency under LIBLINEAR’s randomized coordinate orderings.

Stopping conditions used by LIBSVM and Pegasos. We have not analyzed the robustness of the KKT-based tolerance used by LIBSVM, but it seems plausible that this test could also exhibit nonrobust behavior since the tolerance is not directly connected to the data. And it seems almost certain that the Pegasos termination test, that is, stopping after a fixed number of iterations, lacks robustness in the sense defined herein.

6 Conclusion

In this paper, we studied using first-order methods to solve support vector machines from the perspective of the underlying classification task rather than focusing solely on solving the associated optimization problem to a predefined optimization tolerance. We introduced a computable stopping condition which, under our geometric data model, guarantees the separation of the well-classified points once triggered. We further derived a data-dependent accuracy threshold such that any sufficiently accurate dual solution yields a hyperplane that separates the well-classified points. This threshold depends explicitly on geometric properties of the data, thereby relating the computational complexity of obtaining a classifier to the geometry of the instance itself.

We specialized this theory to FISTA and coordinate descent by deriving computable bounds on the distance from the current iterate to the optimal dual solution, which allow our proposed stopping condition to be applied in practice to these algorithms. For FISTA in particular, we additionally obtained data-dependent iteration complexity bounds for both separating the well-classified points and satisfying the proposed stopping condition.

Our results also illustrate limitations of the stopping condition used by LIBLINEAR. In particular, we constructed examples in which the condition is either overly conservative or terminates prematurely with a poor classifier. Comparatively, the stopping condition we use for FISTA performs consistently across the numerical experiments. Indeed, our numerical experiments show that a simple FISTA implementation using the proposed stopping condition is competitive with widely used SVM solvers across a range of benchmarks.

Finally, we highlight two directions for future research. Specifically in the realm of binary classification, we selected the parameters $\gamma = 64/n$ and $\mu = n/128$ because these parameter worked well in both our theoretical developments and computational experiments. However, there could be data sets with either an exceptionally large or exceptionally small number of misclassified points in which other values might be appropriate. Future research could extend our theory so that γ is selected in a data dependent manner. Looking beyond binary classification, there are countless problems in data science that are solved by optimization, both convex and nonconvex, and, to the best of our knowledge, few if any have data dependent termination tests. Therefore, much work remains to be done in this arena.

Acknowledgments and Disclosure of Funding

Supported in part by a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix A. An example of multiple optimizers of the SVM dual

Consider the one-dimensional case of (4), where $c_1 = c_2 = c_3 = 1, c_4 = -1$ and $d_1 = d_2 = -1, d_3 = 1$, and $\gamma \geq 1$. The well-classified points are clearly c_1, c_2, c_3 and d_1, d_2 . The KKT conditions for this problem are as follows:

$$\begin{aligned}
\mathbf{w} &= C^T \mathbf{u} - D^T \mathbf{v} \\
t &= \mathbf{e}^T \mathbf{u} - \mathbf{e}^T \mathbf{v} \\
\mathbf{z}_1 &= \gamma \mathbf{e} - \mathbf{u} \\
\mathbf{z}_2 &= \gamma \mathbf{e} - \mathbf{v} \\
\mathbf{c}_i^T \mathbf{w} + t &\geq 1 - s_i, \quad \forall i \in [j] \\
\mathbf{d}_i^T \mathbf{w} + t &\leq -1 + s_{i+j}, \quad \forall i \in [l] \\
u_i(1 - s_i - \mathbf{c}_i^T \mathbf{w} - t) &= 0, \quad \forall i \in [j] \\
v_i(1 - s_{i+j} + \mathbf{d}_i^T \mathbf{w} + t) &= 0, \quad \forall i \in [l] \\
\mathbf{z}^T \mathbf{s} &= 0 \\
\mathbf{0} &\leq \mathbf{u} \leq \gamma \mathbf{e} \\
\mathbf{0} &\leq \mathbf{v} \leq \gamma \mathbf{e} \\
\mathbf{s}, \mathbf{z} &\geq \mathbf{0}
\end{aligned} \tag{22}$$

One KKT solution of this problem is given by $\mathbf{w} = 1, t = 0, \mathbf{z}_1 = \gamma \mathbf{e} - \mathbf{u}, \mathbf{z}_2 = \gamma \mathbf{e} - \mathbf{v}$,

$$\mathbf{u} = \begin{pmatrix} \gamma/3 + 1/6 \\ \gamma/3 + 1/6 \\ \gamma/3 + 1/6 \\ \gamma \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} \gamma/2 + 1/4 \\ \gamma/2 + 1/4 \\ \gamma \end{pmatrix}, \quad \mathbf{s}(1:j) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{s}(j+1:j+l) = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix},$$

where $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$. Indeed, this optimal solution is desirable, since for the properly classified points, $u_i < \gamma$ and $v_i < \gamma$. However, we can construct another KKT solution from the above by adding $2\gamma/3 - 1/6$ to u_1 and subtracting $\gamma/3 - 1/12$ from both u_2 and u_3 :

$$\mathbf{u} = \begin{pmatrix} \gamma \\ 1/4 \\ 1/4 \\ \gamma \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} \gamma/2 + 1/4 \\ \gamma/2 + 1/4 \\ \gamma \end{pmatrix}.$$

We still set $\mathbf{z}_1 = \gamma \mathbf{e} - \mathbf{u}$ and $\mathbf{z}_2 = \gamma \mathbf{e} - \mathbf{v}$, and the values of all other variables remain the same. This new optimal solution is undesirable because the well-classified point c_1 has $u_1 = \gamma$, preventing c_1 from being identified through the condition $u_1 < \gamma$. This example can be extended to include arbitrarily many points, in which case arbitrarily many well-classified points can have corresponding optimal dual variables equal to γ .

For both KKT solutions above, we have $t = 0$. Since the stationarity condition for the H-SVM intercept is

$$t = \mathbf{e}^T \mathbf{u} - \mathbf{e}^T \mathbf{v},$$

it follows that

$$\mathbf{e}^T \mathbf{u} = \mathbf{e}^T \mathbf{v}.$$

This is precisely the stationarity condition associated with the unregularized intercept in the classical soft-margin SVM (1). The remaining KKT conditions are identical for the two formulations. Consequently, both solutions above are also KKT solutions of the classical soft-margin SVM. In particular, its dual also has multiple optimal solutions, one of which assigns the upper-bound value γ to the dual variable associated with the well-classified point c_1 .

Appendix B. Additional proofs for H-SVM

Proof of Lemma 4 Recall the definition of ψ in (9). We may write the Lagrangian as

$$\mathcal{L}(s, z, a) = \gamma s + \frac{1}{2\mu} z^2 + a(1 - s - z - \theta),$$

where $s, a \geq 0$. The dual problem is

$$\max_{a \geq 0} \min_{z \in \mathbb{R}, s \geq 0} (\gamma - a)s + \frac{1}{2\mu} z^2 - az + a(1 - \theta),$$

The inner minimization is separable in s and z . In particular, $\min\{(\gamma - a)s : s \geq 0\}$ is unbounded below unless $a \leq \gamma$. For the z problem

$$\min_z \frac{1}{2\mu} z^2 - az,$$

the unique minimizer is $z^* = \mu a$, which gives objective $-\frac{\mu}{2} a^2$. It follows from strong duality that

$$\psi(\theta) = \max \left\{ a(1 - \theta) - \frac{\mu}{2} a^2 : 0 \leq a \leq \gamma \right\}, \quad (23)$$

which has closed-form solution

$$a^*(\theta) = \text{proj}_{[0, \gamma]} \left(\frac{1 - \theta}{\mu} \right).$$

Moreover, for $\theta = \theta_{c,i}^*$ we have that z^* equals the i th coordinate of $\mathbf{z}^*$, and subsequently $a^* = u_i^*$, where we used that $z^* = \mu a^*(\theta_{c,i}^*)$ and the KKT conditions (7). The proof for $\theta_{d,i}^*$ and v_i^* is analogous.

Differentiating $\psi(\theta)$ using the expression (23) gives

$$\psi'(\theta) = -a^* = -\text{proj}_{[0, \gamma]} \left(\frac{1 - \theta}{\mu} \right),$$

since the maximizer of (23) is unique. This completes the proof. ■

Proof of Lemma 7 From the KKT conditions (7),

$$\begin{pmatrix} \mathbf{w}^* \\ t^* \end{pmatrix} = \sum_{i=1}^j u_i^* \begin{pmatrix} \mathbf{c}_i \\ 1 \end{pmatrix} - \sum_{i=1}^l v_i^* \begin{pmatrix} \mathbf{d}_i \\ 1 \end{pmatrix}. \quad (24)$$

Consider the vector $\mathbf{q} := (\mathbf{e}_1, \sigma_2 - \rho)$. For $i \in \mathcal{C}$,

$$\mathbf{q}^T \begin{pmatrix} \mathbf{c}_i \\ 1 \end{pmatrix} = c_{i,1} + \sigma_2 - \rho \geq \sigma_1 + \sigma_2 - 2\rho,$$

while for $i \in \mathcal{D}$,

$$-\mathbf{q}^T \begin{pmatrix} \mathbf{d}_i \\ 1 \end{pmatrix} = -d_{i,1} - \sigma_2 + \rho \geq 0.$$

Meanwhile, for the noisy $\mathbf{c}_i$,

$$\begin{aligned} \mathbf{q}^T \begin{pmatrix} \mathbf{c}_i \\ 1 \end{pmatrix} &= c_{i,1} + \sigma_2 - \rho \\ &\geq \frac{\sigma_1 - \sigma_2}{2} - \frac{\sigma_1 + \sigma_2}{2} - \rho + \sigma_2 - \rho \\ &= -2\rho \\ &> -(\sigma_1 + \sigma_2), \end{aligned}$$

and for the noisy $\mathbf{d}_i$,

$$\begin{aligned} -\mathbf{q}^T \begin{pmatrix} \mathbf{d}_i \\ 1 \end{pmatrix} &= -d_{i,1} - \sigma_2 + \rho \\ &\geq \frac{-\sigma_1 + \sigma_2}{2} - \frac{\sigma_1 + \sigma_2}{2} - \rho - \sigma_2 + \rho \\ &= -(\sigma_1 + \sigma_2). \end{aligned}$$

It follows from (24) that

$$\mathbf{q}^T \begin{pmatrix} \mathbf{w}^* \\ t^* \end{pmatrix} \geq (\sigma_1 + \sigma_2 - 2\rho) \sum_{i \in \mathcal{C}} u_i^* - 2\bar{\sigma}\gamma n_{\text{noise}},$$

where we used that $\sigma_1 + \sigma_2 = 2\bar{\sigma}$, and $u_i^*, v_i^* \leq \gamma$. Rearranging, using Cauchy-Schwarz, $\gamma = 64/n$, and $n_{\text{noise}} \leq \nu n$, we obtain

$$\begin{aligned} (\sigma_1 + \sigma_2 - 2\rho) \sum_{i \in \mathcal{C}} u_i^* &\leq \|\mathbf{q}\| \cdot \|(\mathbf{w}^*, t^*)\| + 128\bar{\sigma}\nu \\ &\leq \xi \sqrt{1 + (\sigma_2 - \rho)^2} + 128\bar{\sigma}\nu. \end{aligned}$$

It follows that there exists an $i \in \mathcal{C}$ such that

$$u_i^* \leq \frac{\xi \sqrt{1 + (\sigma_2 - \rho)^2} + 128\bar{\sigma}\nu}{j_0(\sigma_1 + \sigma_2 - 2\rho)}.$$

If we multiply both sides by $\mu = n/128$, we obtain the inequality

$$\mu u_i^* \leq K_c,$$

and by the assumption that $K_c < 1/2 = \mu\gamma$, it must be that $u_i^* < \gamma$. In view of the projection formula in Lemma 4 and the bound $u_i^* < \gamma$, the upper projection endpoint must be inactive, so

$$\frac{1 - \theta_{c,i}^*}{\mu} \leq u_i^*.$$

After rearranging, we get $\theta_{c,i}^* \geq 1 - \mu u_i^* \geq 1 - K_c$, as desired.

The proof for the **d**-class is analogous, instead using $\mathbf{q} := (\mathbf{e}_1, -(\sigma_1 - \rho))$. ■

FISTA complexity theory. Suppose $\mathbf{r}_k$ are the iterates generated by the strongly convex variant of FISTA applied to

$$F(\mathbf{r}) := -g(\mathbf{r}) + i_\Omega(\mathbf{r}),$$

where g is defined in (5), and recall $\kappa = L/\mu$. Then

$$F(\mathbf{r}_k) - F(\mathbf{r}^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \left(F(\mathbf{r}_0) - F(\mathbf{r}^*) + \frac{\mu}{2}\|\mathbf{r}_0 - \mathbf{r}^*\|^2\right).$$

This follows from (Beck, 2017, Theorem 10.42), since F is μ -strongly convex and $-g$ is L -smooth as shown in Theorem 1.

Lemma 17 *Let $\epsilon > 0$ and suppose $\mathbf{r}_k$ are the iterates generated by the strongly convex variant of FISTA applied to $F(\mathbf{r})$. Then*

$$\|\mathbf{r}_k - \mathbf{r}^*\| \leq \sqrt{\frac{2C_0}{\mu}} \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{k/2}, \quad (25)$$

where $C_0 := F(\mathbf{r}_0) - F(\mathbf{r}^*) + \frac{\mu}{2}\|\mathbf{r}_0 - \mathbf{r}^*\|^2$ and $\kappa = L/\mu$. In particular, if $C_0 > 0$, then $\|\mathbf{r}_k - \mathbf{r}^*\| \leq \epsilon$ whenever

$$k \geq \frac{\log\left(\frac{2C_0}{\mu\epsilon^2}\right)}{\log\left(\frac{1}{1-1/\sqrt{\kappa}}\right)}. \quad (26)$$

If $C_0 = 0$, then $\mathbf{r}_k = \mathbf{r}^*$ for every k .

Proof By strong convexity of F , we have for any $\mathbf{x}, \mathbf{y}$ and $\mathbf{q} \in \partial F(\mathbf{y})$,

$$F(\mathbf{x}) \geq F(\mathbf{y}) + \mathbf{q}^T(\mathbf{x} - \mathbf{y}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Taking $\mathbf{x} = \mathbf{r}_k$, $\mathbf{y} = \mathbf{r}^*$, and using $\mathbf{0} \in \partial F(\mathbf{r}^*)$, we obtain the inequality

$$F(\mathbf{r}_k) - F(\mathbf{r}^*) \geq \frac{\mu}{2}\|\mathbf{r}_k - \mathbf{r}^*\|^2.$$

Combining the two inequalities, we obtain

$$\frac{\mu}{2}\|\mathbf{r}_k - \mathbf{r}^*\|^2 \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k C_0.$$

Rearranging gives

$$\|\mathbf{r}_k - \mathbf{r}^*\| \leq \sqrt{\frac{2C_0}{\mu}} \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{k/2}.$$

If $C_0 = 0$, then the above inequality gives $\mathbf{r}_k = \mathbf{r}^*$ for every k . We may therefore restrict attention to $C_0 > 0$. To show the iteration bound, we want to know for which k

$$\sqrt{\frac{2C_0}{\mu}} \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{k/2} \leq \epsilon$$

is satisfied. Squaring both sides and rearranging gives

$$\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \leq \frac{\epsilon^2 \mu}{2C_0}.$$

Taking the log of both sides gives

$$k \log \left(1 - \frac{1}{\sqrt{\kappa}}\right) \leq \log \left(\frac{\epsilon^2 \mu}{2C_0}\right),$$

where we note that $\kappa > 1$, since Theorem 1 gives $L = \mu + \|\tilde{H}\|^2$ and $\|\tilde{H}\| > 0$ because every augmented data vector has last coordinate ± 1 . Since $1 - 1/\sqrt{\kappa} \in (0, 1)$, dividing by $\log(1 - 1/\sqrt{\kappa})$ swaps the inequality:

$$k \geq \frac{\log \left(\frac{\epsilon^2 \mu}{2C_0}\right)}{\log \left(1 - \frac{1}{\sqrt{\kappa}}\right)},$$

and the desired inequality comes from multiplying the numerator and denominator by -1 . ■

Proof of Theorem 10 From $\mathbf{r}_0 = \mathbf{0}$,

$$\begin{aligned} C_0 &= F(\mathbf{0}) - F(\mathbf{r}^*) + \frac{\mu}{2} \|\mathbf{r}^*\|^2 \\ &= -\frac{1}{2} \|\tilde{H}^T \mathbf{r}^*\|^2 + \mathbf{e}^T \mathbf{r}^* \\ &\leq \mathbf{e}^T \mathbf{r}^*. \end{aligned}$$

Since $\mathbf{r}^* \leq \gamma \mathbf{e}$, we obtain $C_0 \leq \mathbf{e}^T \mathbf{r}^* \leq \gamma n$, and by the assumption that $\gamma = 64/n$, $C_0 \leq 64$. Hence,

$$\frac{2C_0}{\mu} \leq \frac{128^2}{n},$$

where we also used that $\mu = n/128$. Substituting this bound into (25) gives

$$\|\mathbf{r}_k - \mathbf{r}^*\| \leq \frac{128}{\sqrt{n}} \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{k/2}, \quad (27)$$

For the iteration bound, if $C_0 = 0$, then Lemma 17 gives $\mathbf{r}_k = \mathbf{r}^*$ for every k , so the conclusion is immediate. Suppose therefore that $C_0 > 0$. Using $1 - x \leq e^{-x}$ for all x , from (27) we obtain

$$\|\mathbf{r}_k - \mathbf{r}^*\| \leq \frac{128}{\sqrt{n}} \exp\left(\frac{-k}{2\sqrt{\kappa}}\right).$$

Hence, any $k \geq 2\sqrt{\kappa} \log\left(\frac{128}{\sqrt{n}\epsilon}\right)$ gives $\|\mathbf{r}_k - \mathbf{r}^*\| \leq \epsilon$. ■

Appendix C. Theoretical results for classical SVM

Our analysis of the classical SVM in this section will make the same assumptions on the data as in Section 3.1, and will continue to use $\gamma = 64/n$ and $\mu = n/128$. The data assumptions are explicitly invoked in the statements of the results that require them.

The classical soft-margin SVM problem (1) can be written as a sum of a convex smooth term and several convex nonsmooth terms, but the proximal operator of the nonsmooth terms does not admit a simple closed form. As in the H-SVM case, the solution is to apply FISTA to the dual formulation:

$$\begin{aligned} \max_{\mathbf{u}, \mathbf{v}} \quad & -\frac{1}{2} \|C^T \mathbf{u} - D^T \mathbf{v}\|^2 + \mathbf{e}^T \mathbf{u} + \mathbf{e}^T \mathbf{v} \\ \text{s.t.} \quad & \mathbf{e}^T \mathbf{u} = \mathbf{e}^T \mathbf{v}, \\ & \mathbf{0} \leq \mathbf{u} \leq \gamma \mathbf{e}, \\ & \mathbf{0} \leq \mathbf{v} \leq \gamma \mathbf{e}, \end{aligned} \tag{28}$$

where as before, $C = (\mathbf{c}_1^T, \dots, \mathbf{c}_j^T) \in \mathbb{R}^{j \times d}$ and $D = (\mathbf{d}_1^T, \dots, \mathbf{d}_l^T) \in \mathbb{R}^{l \times d}$.

C.1 Projecting onto Ω

Let $\Omega := \{(\mathbf{u}, \mathbf{v}) : \mathbf{e}^T \mathbf{u} = \mathbf{e}^T \mathbf{v}, \mathbf{0} \leq \mathbf{u} \leq \gamma \mathbf{e}, \mathbf{0} \leq \mathbf{v} \leq \gamma \mathbf{e}\}$. The projection $\text{proj}_\Omega(\mathbf{x}, \mathbf{y})$ is the solution to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{u}, \mathbf{v}} \quad & \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 + \frac{1}{2} \|\mathbf{v} - \mathbf{y}\|^2 \\ \text{s.t.} \quad & \mathbf{e}^T \mathbf{u} = \mathbf{e}^T \mathbf{v}, \\ & \mathbf{0} \leq \mathbf{u} \leq \gamma \mathbf{e}, \\ & \mathbf{0} \leq \mathbf{v} \leq \gamma \mathbf{e}. \end{aligned} \tag{29}$$

After dropping constants, (29) can be rewritten in the form of a continuous quadratic knapsack problem:

$$\begin{aligned} \min_{\mathbf{r}} \quad & \frac{1}{2} \mathbf{r}^T I \mathbf{r} - \mathbf{q}^T \mathbf{r} \\ \text{s.t.} \quad & \mathbf{e}^T \mathbf{r} = 0, \\ & \boldsymbol{\ell} \leq \mathbf{r} \leq \boldsymbol{\nu}, \end{aligned} \tag{30}$$

where $\mathbf{r} := (\mathbf{u}, -\mathbf{v})$, $\mathbf{q} := (\mathbf{x}, -\mathbf{y})$, $\boldsymbol{\ell} := (\mathbf{0}, -\gamma \mathbf{e})$, and $\boldsymbol{\nu} := (\gamma \mathbf{e}, \mathbf{0})$. An algorithm for exactly solving (30) in $\mathcal{O}(j+l)$ operations was introduced in Brucker (1984), yielding a linear time method for projecting onto Ω .

C.2 The perturbed dual and its corresponding primal

Let

$$H^T := \begin{pmatrix} C^T & -D^T \end{pmatrix}.$$

As in the H-SVM setting, the unperturbed dual need not have a unique optimal solution (cf. with Section A). We therefore consider the strongly concave perturbation, where $\mu > 0$:

$$\begin{aligned} \max \quad & g(\mathbf{r}) := -\frac{1}{2}\|H^T \mathbf{r}\|^2 + \mathbf{e}^T \mathbf{r} - \frac{\mu}{2}\|\mathbf{r}\|^2 \\ \text{s.t.} \quad & \mathbf{r} \in \Omega, \end{aligned} \tag{31}$$

Our focus in this section will be on solving the formulation (31). The dual of (31) is

$$\begin{aligned} \min_{s,t,\mathbf{w},\mathbf{z}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \mathbf{e}^T \mathbf{s} + \frac{1}{2\mu}\|\mathbf{z}\|^2 \\ \text{s.t.} \quad & C\mathbf{w} + t\mathbf{e} + \mathbf{z}_1 \geq \mathbf{e} - \mathbf{s}_1, \\ & D\mathbf{w} + t\mathbf{e} - \mathbf{z}_2 \leq -\mathbf{e} + \mathbf{s}_2, \\ & \mathbf{s} \geq \mathbf{0}, \end{aligned} \tag{32}$$

where $\mathbf{s}_1 = \mathbf{s}(1:j)$, $\mathbf{s}_2 = \mathbf{s}(j+1:j+l)$ and $\mathbf{z}_1 = \mathbf{z}(1:j)$, $\mathbf{z}_2 = \mathbf{z}(j+1:j+l)$. The KKT conditions relating (31) and (32) are

$$\begin{aligned} \mathbf{u} &= \mathbf{z}_1/\mu \\ \mathbf{v} &= \mathbf{z}_2/\mu \\ \mathbf{w} &= C^T \mathbf{u} - D^T \mathbf{v} \\ \mathbf{e}^T \mathbf{u} &= \mathbf{e}^T \mathbf{v} \\ \mathbf{z}_1 &= (1-t)\mathbf{e} - \mathbf{s}_1 - C\mathbf{w} + \boldsymbol{\zeta}_1 \\ \mathbf{z}_2 &= (1+t)\mathbf{e} - \mathbf{s}_2 + D\mathbf{w} + \boldsymbol{\zeta}_2 \\ \mathbf{s}_1^T (\mathbf{u} - \gamma \mathbf{e}) &= 0 \\ \mathbf{s}_2^T (\mathbf{v} - \gamma \mathbf{e}) &= 0 \\ \boldsymbol{\zeta}_1^T \mathbf{u} &= 0 \\ \boldsymbol{\zeta}_2^T \mathbf{v} &= 0 \\ \mathbf{0} &\leq \mathbf{u} \leq \gamma \mathbf{e} \\ \mathbf{0} &\leq \mathbf{v} \leq \gamma \mathbf{e} \\ \mathbf{s} &\geq \mathbf{0} \\ \boldsymbol{\zeta} &\geq \mathbf{0}, \end{aligned} \tag{33}$$

where $\boldsymbol{\zeta}_1 = \boldsymbol{\zeta}(1:j)$ and $\boldsymbol{\zeta}_2 = \boldsymbol{\zeta}(j+1:j+l)$. Note these conditions are identical to those in (7) except that the condition $t = \mathbf{e}^T \mathbf{u} - \mathbf{e}^T \mathbf{v}$ has become $0 = \mathbf{e}^T \mathbf{u} - \mathbf{e}^T \mathbf{v}$.

As in the analysis of H-SVM, we may write (32) as an unconstrained convex optimization problem:

$$\min_{\mathbf{w},t} f(\mathbf{w},t) := \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^j \psi(\mathbf{c}_i^T \mathbf{w} + t) + \sum_{i=1}^l \psi(-\mathbf{d}_i^T \mathbf{w} - t). \tag{34}$$

C.3 Early stopping for classical SVM

By similar reasoning as in the proof of Theorem 1, g defined in (31) is μ -strongly concave and L -smooth with $L = \mu + \|H\|^2$. For classical SVM, the FISTA iterations remain the same, with $\kappa = L/\mu$. Thus, we may use the same error bound

$$\epsilon_{k+1}^{\text{FISTA}} := \kappa \|\mathbf{y}_k - \mathbf{r}_{k+1}\|.$$

Theorem 18 *Let $\mathbf{y} \in \mathbb{R}^n$, and define $\mathbf{r} := \text{proj}_\Omega(\mathbf{y} + \frac{1}{L}\nabla g(\mathbf{y}))$, $\epsilon := \kappa \|\mathbf{y} - \mathbf{r}\|$. Then*

$$\|\mathbf{r} - \mathbf{r}^*\| \leq \epsilon.$$

In particular, the FISTA iterates satisfy

$$\|\mathbf{r}_{k+1} - \mathbf{r}^*\| \leq \epsilon_{k+1}^{\text{FISTA}}.$$

Proof The proof is identical to the proof of Theorem 2 after replacing $\tilde{H}$ with H and noting that the definitions of Ω and L are different. $\blacksquare$

For classical SVM, we also stop at the first iteration for which at least one coordinate satisfies

$$\gamma - r_{k+1,i} > \epsilon_{k+1}^{\text{FISTA}}, \quad \text{for some } i \in [n]. \quad (35)$$

Computing the classifier to return after stopping is not as simple for classical SVM. While we can obtain $\mathbf{w}(\mathbf{r}) = H^T \mathbf{r}$ from some $\mathbf{r} \in \Omega$, computing the approximate intercept $t(\mathbf{r})$ is less straightforward than in the H-SVM case. We will take $t(\mathbf{r})$ to be an optimal solution to the one-dimensional unconstrained optimization problem

$$\min_t f(\mathbf{w}, t). \quad (36)$$

Recall that $\psi(\theta)$ is convex, hence $\psi'(\theta)$ is nondecreasing. Let $\mathbf{w}^*$ be optimal for (32). Define

$$\eta := \max \left\{ \max_{i \in [j]} |\mathbf{c}_i^T(\mathbf{w} - \mathbf{w}^*)|, \max_{i \in [l]} |\mathbf{d}_i^T(\mathbf{w} - \mathbf{w}^*)| \right\}.$$

Lemma 19 *Let $\mathbf{w} \in \mathbb{R}^d$ and denote $T(\mathbf{w}) := \arg\min_{t \in \mathbb{R}} f(\mathbf{w}, t)$. Then $T(\mathbf{w})$ is a nonempty closed interval. Moreover, for every $t \in T(\mathbf{w})$, there exists $t^* \in T(\mathbf{w}^*)$ such that*

$$|t - t^*| \leq \eta.$$

Proof For the proof, we will use the abuse of notation

$$f'(\mathbf{w}, t) = \frac{\partial}{\partial t} f(\mathbf{w}, t) = \sum_{i=1}^j \psi'(\mathbf{c}_i^T \mathbf{w} + t) - \sum_{i=1}^l \psi'(-\mathbf{d}_i^T \mathbf{w} - t).$$

$T(\mathbf{w})$ is nonempty by the coercivity of $t \mapsto f(\mathbf{w}, t)$. Indeed, taking $|t| \rightarrow \infty$ sends $f(\mathbf{w}, t) \rightarrow \infty$ for fixed $\mathbf{w}$. Since $\psi'(\theta)$ is continuous and nondecreasing, $-\psi'(-\theta)$ is also continuous and nondecreasing. Thus, $f'(\mathbf{w}, t)$ is a continuous and nondecreasing function of t . Writing

$$T(\mathbf{w}) = \{t \in \mathbb{R} : f'(\mathbf{w}, t) = 0\},$$

makes it clear that $T(\mathbf{w})$ is a closed interval. Denote $T(\mathbf{w}^*) := [t^-, t^+]$.

For any $s \in \mathbb{R}$,

$$\mathbf{c}_i^T \mathbf{w}^* + s = \mathbf{c}_i^T \mathbf{w} + s + \mathbf{c}_i^T (\mathbf{w}^* - \mathbf{w}) \leq \mathbf{c}_i^T \mathbf{w} + (s + \eta),$$

and

$$-\mathbf{d}_i^T \mathbf{w}^* - s = -\mathbf{d}_i^T \mathbf{w} - s - \mathbf{d}_i^T (\mathbf{w}^* - \mathbf{w}) \geq -\mathbf{d}_i^T \mathbf{w} - (s + \eta).$$

Thus, by ψ' nondecreasing, we have

$$f'(\mathbf{w}, s + \eta) \geq f'(\mathbf{w}^*, s). \quad (37)$$

Similarly,

$$\mathbf{c}_i^T \mathbf{w}^* + s \geq \mathbf{c}_i^T \mathbf{w} + (s - \eta),$$

and

$$-\mathbf{d}_i^T \mathbf{w}^* - s \leq -\mathbf{d}_i^T \mathbf{w} - (s - \eta).$$

So,

$$f'(\mathbf{w}, s - \eta) \leq f'(\mathbf{w}^*, s). \quad (38)$$

If $t < t^- - \eta$, $f'(\mathbf{w}^*, t + \eta) < 0$ and taking $s = t + \eta$ in (38),

$$f'(\mathbf{w}, t) \leq f'(\mathbf{w}^*, t + \eta) < 0,$$

from which it follows that $t \notin T(\mathbf{w})$. On the other hand, if $t > t^+ + \eta$, $f'(\mathbf{w}^*, t - \eta) > 0$ and taking $s = t - \eta$ in (37),

$$f'(\mathbf{w}, t) \geq f'(\mathbf{w}^*, t - \eta) > 0,$$

so again $t \notin T(\mathbf{w})$. It must then be that

$$T(\mathbf{w}) \subseteq [t^- - \eta, t^+ + \eta],$$

and the result follows. ■

The classical SVM case also benefits from the following relationship between the primal and dual optimal solutions.

Theorem 20 *Let $(\mathbf{u}^*, \mathbf{v}^*)$ be the unique optimal solution of (31), and let $(\mathbf{w}^*, t^*)$ be any optimal primal classifier. Then*

$$u_i^* = \text{proj}_{[0, \gamma]} \left(\frac{1 - \mathbf{c}_i^T \mathbf{w}^* - t^*}{\mu} \right), \quad \forall i \in [j],$$

and

$$v_i^* = \text{proj}_{[0, \gamma]} \left(\frac{1 + \mathbf{d}_i^T \mathbf{w}^* + t^*}{\mu} \right), \quad \forall i \in [l].$$

Equivalently,

$$u_i^* = -\psi'(\mathbf{c}_i^T \mathbf{w}^* + t^*), \quad \forall i \in [j], \quad \text{and} \quad v_i^* = -\psi'(-\mathbf{d}_i^T \mathbf{w}^* - t^*), \quad \forall i \in [l].$$

Proof The proof of this result mirrors that of Lemma 4, since the relevant KKT conditions in (33) are identical to (7). $\blacksquare$

Define

$$\theta_i(\mathbf{w}(\mathbf{r}), t(\mathbf{r})) := \begin{cases} \mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}), & i = 1, \dots, j, \\ -\mathbf{d}_{i-j}^T \mathbf{w}(\mathbf{r}) - t(\mathbf{r}), & i = j+1, \dots, n, \end{cases}$$

where $\mathbf{w}(\mathbf{r}) = H^T \mathbf{r}$ and $t(\mathbf{r})$ is some point in $T(\mathbf{w}(\mathbf{r}))$. We show an analogous result to Theorem 3 for the classical SVM, which proves that the condition $\gamma - r_i > \epsilon$ implies the current classifier correctly classifies the corresponding point, as does the optimal classifier.

Theorem 21 *Suppose $\mu\gamma < 1$. Let $\mathbf{r} = (\mathbf{u}, \mathbf{v})$ be the projected gradient step for some $\mathbf{y} \in \mathbb{R}^n$, and $\epsilon = \kappa \|\mathbf{y} - \mathbf{r}\|$, as in Theorem 18. If $\gamma - r_i > \epsilon$, then*

$$\theta_i(\mathbf{w}(\mathbf{r}), t(\mathbf{r})) > 1 - \mu\gamma > 0, \quad \theta_i(\mathbf{w}(\mathbf{r}^*), t(\mathbf{r}^*)) > 1 - \mu\gamma > 0.$$

Proof Let $\mathbf{w} := H^T \mathbf{r}$ and $t := t(\mathbf{r}) \in T(\mathbf{w}(\mathbf{r}))$ be chosen arbitrarily. Define $\bar{\mathbf{r}} := (\bar{\mathbf{u}}, \bar{\mathbf{v}})$ by

$$\begin{aligned} \bar{u}_i &:= \text{proj}_{[0, \gamma]} \left(\frac{1 - \mathbf{c}_i^T \mathbf{w} - t}{\mu} \right), \quad \forall i \in [j], \\ \bar{v}_i &:= \text{proj}_{[0, \gamma]} \left(\frac{1 + \mathbf{d}_i^T \mathbf{w} + t}{\mu} \right), \quad \forall i \in [l]. \end{aligned}$$

We showed in the proof of Lemma 4 that

$$\psi'(\theta) = -\text{proj}_{[0, \gamma]} \left(\frac{1 - \theta}{\mu} \right).$$

Therefore, using that $t \in \text{argmin}_s f(\mathbf{w}, s)$, we have

$$\begin{aligned} 0 &= \frac{\partial f(\mathbf{w}, t)}{\partial t} \\ &= \sum_{i=1}^j \psi'(\mathbf{c}_i^T \mathbf{w} + t) - \sum_{i=1}^l \psi'(-\mathbf{d}_i^T \mathbf{w} - t) \\ &= -\mathbf{e}^T \bar{\mathbf{u}} + \mathbf{e}^T \bar{\mathbf{v}}. \end{aligned}$$

So $\bar{\mathbf{r}} \in \Omega$. Define the vector $\mathbf{a}$ such that

$$a_i = \begin{cases} 1, & i \in \{1, \dots, j\}, \\ -1, & i \in \{j+1, \dots, n\}. \end{cases}$$

We may write

$$\bar{\mathbf{r}} = \text{proj}_{[0, \gamma]^n} \left(\frac{\mathbf{e} - H\mathbf{w} - \mathbf{a}t}{\mu} \right),$$

which has optimality conditions

$$\bar{\mathbf{q}} := \mathbf{e} - H\mathbf{w} - \mathbf{a}t - \mu\bar{\mathbf{r}} \in \mathcal{N}_{[0, \gamma]^n}(\bar{\mathbf{r}}).$$

Define

$$\bar{\mathbf{z}} := \bar{\mathbf{q}} + \mathbf{a}t = \mathbf{e} - H\mathbf{w} - \mu\bar{\mathbf{r}}.$$

For every $\mathbf{s} \in \Omega$, $\mathbf{a}^T(\mathbf{s} - \bar{\mathbf{r}}) = 0$, therefore,

$$\bar{\mathbf{z}}^T(\mathbf{s} - \bar{\mathbf{r}}) = \bar{\mathbf{q}}^T(\mathbf{s} - \bar{\mathbf{r}}) + t\mathbf{a}^T(\mathbf{s} - \bar{\mathbf{r}}) = \bar{\mathbf{q}}^T(\mathbf{s} - \bar{\mathbf{r}}) \leq 0,$$

because $\bar{\mathbf{q}} \in \mathcal{N}_{[0,\gamma]^n}(\bar{\mathbf{r}})$ and $\mathbf{s} \in \Omega \subseteq [0, \gamma]^n$. Hence, $\bar{\mathbf{z}} \in \mathcal{N}_\Omega(\bar{\mathbf{r}})$, and

$$H\mathbf{w} + \mu\bar{\mathbf{r}} - \mathbf{e} + \bar{\mathbf{z}} = \mathbf{0}. \quad (39)$$

Let $Q := HH^T + \mu I$. Recall from the proof of Theorem 2 that for $\mathbf{z} := \nabla g(\mathbf{y}) + L(\mathbf{y} - \mathbf{r}) \in \mathcal{N}_\Omega(\mathbf{r})$ and $\mathbf{q} := Q\mathbf{r} - \mathbf{e} + \mathbf{z}$,

$$\|\mathbf{q}\| \leq \mu\epsilon.$$

Moreover, $\mathbf{w} = H^T\mathbf{r}$, so $\mathbf{q} = H\mathbf{w} + \mu\mathbf{r} - \mathbf{e} + \mathbf{z}$. Subtracting (39) from this quantity gives

$$\mathbf{q} = \mu(\mathbf{r} - \bar{\mathbf{r}}) + \mathbf{z} - \bar{\mathbf{z}}.$$

By monotonicity of the normal cone,

$$\begin{aligned} \|\mathbf{r} - \bar{\mathbf{r}}\| \cdot \|\mathbf{q}\| &\geq (\mathbf{r} - \bar{\mathbf{r}})^T \mathbf{q} \\ &= \mu\|\mathbf{r} - \bar{\mathbf{r}}\|^2 + (\mathbf{r} - \bar{\mathbf{r}})^T (\mathbf{z} - \bar{\mathbf{z}}) \\ &\geq \mu\|\mathbf{r} - \bar{\mathbf{r}}\|^2. \end{aligned}$$

Thus, $\|\mathbf{r} - \bar{\mathbf{r}}\| \leq \|\mathbf{q}\|/\mu \leq \epsilon$. Hence,

$$\bar{r}_i \leq r_i + \epsilon < \gamma,$$

and we know

$$\bar{r}_i = \text{proj}_{[0,\gamma]} \left(\frac{1 - \theta_i(\mathbf{w}, t)}{\mu} \right).$$

It follows that $1 - \theta_i(\mathbf{w}, t) < \mu\gamma$, i.e. $\theta_i(\mathbf{w}, t) > 1 - \mu\gamma$.

For the optimal classifier, Theorem 18 gives $r_i^* \leq r_i + \epsilon < \gamma$. Let $(\mathbf{w}^*, t^*)$ be any optimal primal classifier. By Theorem 20,

$$r_i^* = \text{proj}_{[0,\gamma]} \left(\frac{1 - \theta_i(\mathbf{w}^*, t^*)}{\mu} \right),$$

but $r_i^* < \gamma$, so

$$\frac{1 - \theta_i(\mathbf{w}^*, t^*)}{\mu} < \gamma \iff \theta_i(\mathbf{w}^*, t^*) > 1 - \mu\gamma.$$

■

C.4 Correctness of the classifier at termination for classical SVM

In this section we show that the stopping condition guarantees separation of the well-classified points by the returned classifier, under the assumptions of our data model.

Lemma 22 *Recall the parameter assumptions imply $\mu\gamma = 1/2$. If $\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K$ for some $K \in [0, 1/2)$, then $u_i^* \leq 2K\gamma < \gamma$. Likewise, if $-\mathbf{d}_i^T \mathbf{w}^* - t^* \geq 1 - K$ for some $K \in [0, 1/2)$, then $v_i^* \leq 2K\gamma < \gamma$.*

Proof The proof of this result is identical to the proof of Lemma 5, where we invoke Theorem 20 instead of Lemma 4. $\blacksquare$

Define $\bar{\sigma} := \frac{\sigma_1 + \sigma_2}{2}$ and

$$\xi := \sqrt{\frac{1}{(\bar{\sigma} - \rho)^2} + 128\nu \left(\frac{3}{4} + \frac{\bar{\sigma} + \rho}{\bar{\sigma} - \rho} \right)}$$

Lemma 23 *Suppose the data satisfy the assumptions of Section 3.1. Let $(\mathbf{w}^*, t^*)$ solve (32). Then $\|\mathbf{w}^*\| \leq \xi$.*

Proof The argument in the proof of Lemma 6 applies verbatim using the same choice of $(\bar{\mathbf{w}}, \bar{t})$, except that the classical objective contains no $t^2/2$ term. The result of this is that the δ term in ξ no longer appears. $\blacksquare$

Define $p_c := j_0/n$, $p_d := l_0/n$, and introduce

$$K := \max \left\{ \frac{\xi + 128\bar{\sigma}\nu}{128p_c(\sigma_1 + \sigma_2 - 2\rho)}, \frac{\xi + 128\bar{\sigma}\nu}{128p_d(\sigma_1 + \sigma_2 - 2\rho)} \right\} + 2\rho\xi.$$

Note that K obtained for classical SVM is no larger than the one obtained for H-SVM. Thus, for sufficiently small ρ, ν , we still have $K < 1/2$.

Theorem 24 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$. Let $(\mathbf{w}^*, t^*)$ be an arbitrary solution to (32). For any $i \in \mathcal{C}$,*

$$\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K,$$

and for any $i \in \mathcal{D}$,

$$-\mathbf{d}_i^T \mathbf{w}^* - t^* \geq 1 - K.$$

Proof From the KKT conditions (33),

$$\begin{pmatrix} \mathbf{w}^* \\ 0 \end{pmatrix} = \sum_{i=1}^j u_i^* \begin{pmatrix} \mathbf{c}_i \\ 1 \end{pmatrix} - \sum_{i=1}^l v_i^* \begin{pmatrix} \mathbf{d}_i \\ 1 \end{pmatrix} \quad (40)$$

As in the proof of Lemma 7, we can consider the vector $\mathbf{q} := (\mathbf{e}_1, \sigma_2 - \rho)$ and obtain for $i \in \mathcal{C}$,

$$\mathbf{q}^T \begin{pmatrix} \mathbf{c}_i \\ 1 \end{pmatrix} \geq \sigma_1 + \sigma_2 - 2\rho,$$

while for $i \in \mathcal{D}$,

$$-\mathbf{q}^T \begin{pmatrix} \mathbf{d}_i \\ 1 \end{pmatrix} \geq 0.$$

Meanwhile, for the noisy $\mathbf{c}_i$,

$$\mathbf{q}^T \begin{pmatrix} \mathbf{c}_i \\ 1 \end{pmatrix} > -(\sigma_1 + \sigma_2),$$

and for the noisy $\mathbf{d}_i$,

$$-\mathbf{q}^T \begin{pmatrix} \mathbf{d}_i \\ 1 \end{pmatrix} \geq -(\sigma_1 + \sigma_2).$$

It follows from (40) that

$$\mathbf{e}_1^T \mathbf{w}^* = \mathbf{q}^T \begin{pmatrix} \mathbf{w}^* \\ 0 \end{pmatrix} \geq (\sigma_1 + \sigma_2 - 2\rho) \sum_{i \in \mathcal{C}} u_i^* - 2\bar{\sigma}\gamma n_{\text{noise}},$$

where we used that $\sigma_1 + \sigma_2 = 2\bar{\sigma}$, and $u_i^*, v_i^* \leq \gamma$. Rearranging, using Cauchy-Schwarz, $\gamma = 64/n$, and $n_{\text{noise}} \leq \nu n$, we obtain

$$\begin{aligned} (\sigma_1 + \sigma_2 - 2\rho) \sum_{i \in \mathcal{C}} u_i^* &\leq \|\mathbf{w}^*\| + 128\bar{\sigma}\nu \\ &\leq \xi + 128\bar{\sigma}\nu. \end{aligned}$$

It follows that there exists an $i \in \mathcal{C}$ such that

$$u_i^* \leq \frac{\xi + 128\bar{\sigma}\nu}{j_0(\sigma_1 + \sigma_2 - 2\rho)}.$$

If we multiply both sides by $\mu = n/128$, we obtain the inequality

$$\mu u_i^* \leq \frac{\xi + 128\bar{\sigma}\nu}{128p_c(\sigma_1 + \sigma_2 - 2\rho)},$$

and by the assumption that $K < 1/2 = \mu\gamma$, the right hand side in the above is less than $\mu\gamma$, so it must be that $u_i^* < \gamma$. In view of the projection formula in Theorem 20 and the bound $u_i^* < \gamma$, the upper projection endpoint must be inactive, so

$$\frac{1 - \mathbf{c}_i^T \mathbf{w}^* - t^*}{\mu} \leq u_i^*.$$

After rearranging, we get

$$\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - \mu u_i^* \geq 1 - \frac{\xi + 128\bar{\sigma}\nu}{128p_c(\sigma_1 + \sigma_2 - 2\rho)}.$$

The remainder of the proof follows exactly the same as in the H-SVM case, where we apply the proof of Corollary 8 and the comments thereafter.

The proof for the $\mathbf{d}$ -class is analogous, instead using $\mathbf{q} := (\mathbf{e}_1, -(\sigma_1 - \rho))$. ■

We may now show that, under the assumptions of the data model, once one point satisfies the stopping condition, all well-classified points must be separated by the returned classifier.

Theorem 25 Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$ and $n \geq 185$. For some $\mathbf{y} \in \mathbb{R}^n$, define $\mathbf{r} := \text{proj}_\Omega(\mathbf{y} + \frac{1}{L}\nabla g(\mathbf{y}))$ and $\epsilon := \kappa\|\mathbf{y} - \mathbf{r}\|$. Let $\mathbf{w}(\mathbf{r}) = H^T \mathbf{r}$ and $t(\mathbf{r}) \in T(\mathbf{w}(\mathbf{r}))$ be chosen arbitrarily. If $\gamma - r_i > \epsilon$ for some $i \in [n]$, then

$$\mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) > 0, \quad \forall i \in \mathcal{C},$$

and

$$-\mathbf{d}_i^T \mathbf{w}(\mathbf{r}) - t(\mathbf{r}) > 0, \quad \forall i \in \mathcal{D}.$$

Proof Let $\mathbf{w}^* := H^T \mathbf{r}^*$, $\mathbf{w} := \mathbf{w}(\mathbf{r}) = H^T \mathbf{r}$, and $t := t(\mathbf{r}) \in T(\mathbf{w}(\mathbf{r}))$. Denote $Q := HH^T + \mu I$. As in the proof of Theorem 21, $\mathbf{z} := \nabla g(\mathbf{y}) + L(\mathbf{y} - \mathbf{r}) \in \mathcal{N}_\Omega(\mathbf{r})$, while $\mathbf{q} := Q\mathbf{r} - \mathbf{e} + \mathbf{z}$ satisfies $\|\mathbf{q}\| \leq \mu\epsilon$. By optimality of $\mathbf{r}^*$, there exists some $\mathbf{z}^* \in \mathcal{N}_\Omega(\mathbf{r}^*)$ such that

$$\mathbf{0} = Q\mathbf{r}^* - \mathbf{e} + \mathbf{z}^*.$$

Subtracting the above from $\mathbf{q}$ gives the equation

$$\mathbf{q} = Q(\mathbf{r} - \mathbf{r}^*) + \mathbf{z} - \mathbf{z}^*.$$

Monotonicity of $\mathcal{N}_\Omega$ implies

$$\begin{aligned} \|\mathbf{q}\| \cdot \|\mathbf{r} - \mathbf{r}^*\| &\geq (\mathbf{r} - \mathbf{r}^*)^T \mathbf{q} \\ &= \|H^T(\mathbf{r} - \mathbf{r}^*)\|^2 + \mu\|\mathbf{r} - \mathbf{r}^*\|^2 + (\mathbf{r} - \mathbf{r}^*)^T(\mathbf{z} - \mathbf{z}^*) \\ &\geq \|\mathbf{w} - \mathbf{w}^*\|^2 + \mu\|\mathbf{r} - \mathbf{r}^*\|^2. \end{aligned}$$

We then obtain the inequality

$$\begin{aligned} \|\mathbf{w} - \mathbf{w}^*\|^2 &\leq \|\mathbf{q}\| \cdot \|\mathbf{r} - \mathbf{r}^*\| - \mu\|\mathbf{r} - \mathbf{r}^*\|^2 \\ &= \frac{\|\mathbf{q}\|^2}{4\mu} - \mu \left(\|\mathbf{r} - \mathbf{r}^*\| - \frac{\|\mathbf{q}\|}{2\mu} \right)^2. \end{aligned}$$

Dropping the negative term on the right and taking the square root of both sides gives

$$\|\mathbf{w} - \mathbf{w}^*\| \leq \frac{1}{2\sqrt{\mu}} \|\mathbf{q}\| \leq \frac{\sqrt{\mu}\epsilon}{2} < \frac{\sqrt{\mu}\gamma}{2} = \sqrt{\frac{8}{n}},$$

where the strict inequality comes from the assumption that $\gamma - r_i > \epsilon$, so $\epsilon < \gamma$. These steps are similar to the ones in the proof of Theorem 9. The difference comes from the fact that we need to bound $|t - t^*|$ as well. From Lemma 19, we can find a $t^* \in T(\mathbf{w}^*)$ satisfying

$$|t - t^*| \leq \eta \leq R\|\mathbf{w} - \mathbf{w}^*\|.$$

By Theorem 24, for all $i \in \mathcal{C}$, $\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K$. Therefore,

$$\begin{aligned} \mathbf{c}_i^T \mathbf{w} + t &= \mathbf{c}_i^T \mathbf{w}^* + t^* + \mathbf{c}_i^T(\mathbf{w} - \mathbf{w}^*) + (t - t^*) \\ &\geq 1 - K - \|\mathbf{c}_i\| \cdot \|\mathbf{w} - \mathbf{w}^*\| - |t - t^*| \\ &\geq 1 - K - 2R\|\mathbf{w} - \mathbf{w}^*\| \\ &\geq 1 - K - R\sqrt{\frac{32}{n}}. \end{aligned}$$

By the definitions of K and ξ ,

$$K \geq 2\rho\xi \geq \frac{2\rho}{\bar{\sigma} - \rho}.$$

Since $K < 1/2$, we have $\rho < \bar{\sigma}/5 \leq 1/5$. Thus,

$$R = \max\{\sigma_1, \sigma_2\} + \rho < 1 + \frac{1}{5} = \frac{6}{5}.$$

Hence, for all $n \geq 185$,

$$\mathbf{c}_i^T \mathbf{w} + t \geq 1 - K - R\sqrt{\frac{32}{n}} > \frac{1}{2} - \frac{6}{5}\sqrt{\frac{32}{185}} > 0.$$

An identical argument shows $-\mathbf{d}_i^T \mathbf{w} - t > 0$ for all $i \in \mathcal{D}$. ■

C.5 Data-dependent iteration complexity for classical SVM

We finish by deriving two iteration bounds. The first gives the number of FISTA iterations required to guarantee separation of the well-classified points. The second gives the number of iterations until our stopping condition is satisfied.

Let $\bar{K} \in (K, 1)$ and define

$$\Delta_{\text{sep}} := \frac{\bar{K} - K}{2R\|H\|}.$$

The following theorem shows that any approximate solution satisfying $\|\mathbf{r} - \mathbf{r}^*\| \leq \Delta_{\text{sep}}$ yields a classifier that separates the well-classified points with signed margin at least $1 - \bar{K}$.

Theorem 26 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$. Let $\bar{K} \in (K, 1)$, and let $\mathbf{r} \in \Omega$ satisfy $\|\mathbf{r} - \mathbf{r}^*\| \leq \Delta_{\text{sep}}$. Then the classifier $(\mathbf{w}(\mathbf{r}), t(\mathbf{r}))$ obtained by computing $\mathbf{w}(\mathbf{r}) = H^T \mathbf{r}$ and $t(\mathbf{r}) \in T(\mathbf{w}(\mathbf{r}))$ separates the well-classified points:*

$$\begin{aligned} \mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) &\geq 1 - \bar{K}, \quad \forall i \in \mathcal{C}, \\ \mathbf{d}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) &\leq -1 + \bar{K}, \quad \forall i \in \mathcal{D}. \end{aligned}$$

Proof Since $\mathbf{w}^* = H^T \mathbf{r}^*$,

$$\|\mathbf{w}(\mathbf{r}) - \mathbf{w}^*\| \leq \|H\| \|\mathbf{r} - \mathbf{r}^*\|.$$

Thus,

$$\eta \leq R\|\mathbf{w} - \mathbf{w}^*\| \leq R\|H\|\Delta_{\text{sep}}.$$

Fix some $i \in \mathcal{C}$. By Lemma 19 we may choose $t^* \in T(\mathbf{w}^*)$ so that $|t(\mathbf{r}) - t^*| \leq \eta$. By Theorem 24, $\mathbf{c}_i^T \mathbf{w}^* + t^* \geq 1 - K$, and

$$\begin{aligned} \mathbf{c}_i^T \mathbf{w}(\mathbf{r}) + t(\mathbf{r}) &= \mathbf{c}_i^T \mathbf{w}^* + t^* + \mathbf{c}_i^T (\mathbf{w}(\mathbf{r}) - \mathbf{w}^*) + (t(\mathbf{r}) - t^*) \\ &\geq \mathbf{c}_i^T \mathbf{w}^* + t^* - \|\mathbf{c}_i\| \cdot \|\mathbf{w}(\mathbf{r}) - \mathbf{w}^*\| - |t(\mathbf{r}) - t^*| \\ &\geq 1 - K - 2R\|H\|\Delta_{\text{sep}} \\ &= 1 - \bar{K}, \end{aligned}$$

where the second inequality used that $|t(\mathbf{r}) - t^*| \leq \eta$ along with the previously obtained upper bound on η .

The proof for the points in the $\mathbf{d}$ -class is analogous. $\blacksquare$

The FISTA convergence theory in Theorem 10 also holds for classical SVM, with $\tilde{H}$ replaced by H and $\kappa = L/\mu = (\mu + \|H\|^2)/\mu$. Indeed, the same proof applies, since $\mathbf{0} \in \Omega$, and

$$C_0 = -\frac{1}{2}\|H^T \mathbf{r}^*\|^2 + \mathbf{e}^T \mathbf{r}^* \leq \mathbf{e}^T \mathbf{r}^* \leq \gamma n = 64.$$

Corollary 27 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$. Let $\bar{K} \in (K, 1)$, and consider FISTA initialized at $\mathbf{r}_0 = \mathbf{0}$, with $\gamma = 64/n$ and $\mu = n/128$. If*

$$k \geq 2\sqrt{\kappa} \log \left(\frac{256R^2}{\bar{K} - K} \right),$$

then the classifier $(\mathbf{w}(\mathbf{r}_k), t(\mathbf{r}_k))$ obtained by computing $\mathbf{w}(\mathbf{r}_k) = H^T \mathbf{r}_k$ and $t(\mathbf{r}_k) \in T(\mathbf{w}(\mathbf{r}_k))$, separates the well-classified points with signed margin at least $1 - \bar{K}$.

Proof Since $\|H\| \leq \|H\|_F \leq \sqrt{n}R$, we have

$$\Delta_{\text{sep}} \geq \frac{\bar{K} - K}{2\sqrt{n}R^2}.$$

Hence, we may apply Theorem 10 to the quantity on the right hand side above to obtain a bound on the number of iterations required to achieve the smaller accuracy

$$\|\mathbf{r}_k - \mathbf{r}^*\| \leq \frac{\bar{K} - K}{2\sqrt{n}R^2}.$$

The conclusion follows from Theorem 26. $\blacksquare$

We now derive an upper bound on the number of iterations required before our stopping condition is satisfied. Suppose the assumptions of Section 3.1 hold with $K < 1/2$ and fix some $i \in \mathcal{C}$. By Theorem 24 and Lemma 22, $u_i^* \leq 2K\gamma$. Applying Theorem 18 gives

$$\gamma - r_{k,i} \geq \gamma - u_i^* - |r_{k,i} - u_i^*| \geq (1 - 2K)\gamma - \epsilon_k^{\text{FISTA}}.$$

Therefore, $\epsilon_k^{\text{FISTA}} < (1 - 2K)\gamma/2$ guarantees that $\mathbf{c}_i$ satisfies the stopping condition. An identical argument can be made for $\mathbf{d}_i$ when $i \in \mathcal{D}$.

Theorem 28 *Suppose the data satisfy the assumptions of Section 3.1, with $K < 1/2$. Consider FISTA initialized at $\mathbf{r}_0 = \mathbf{0}$, with $\gamma = 64/n$ and $\mu = n/128$. The stopping condition $\gamma - r_{k,i} > \epsilon_k^{\text{FISTA}}$ for some $i \in [n]$ is guaranteed to be satisfied within the first k iterations whenever*

$$k > 2 + 2\sqrt{\kappa} \log \left(\frac{16\kappa\sqrt{n}}{1 - 2\bar{K}} \right).$$

Proof The proof is identical to the proof of Theorem 13. $\blacksquare$

References

- David Applegate, Oliver Hinder, Haihao Lu, and Miles Lubin. Faster first-order primal-dual methods for linear programming using restarts and sharpness. *Mathematical Programming*, 201(1):133–184, Sep 2023. ISSN 1436-4646. doi: 10.1007/s10107-022-01901-9. URL <https://doi.org/10.1007/s10107-022-01901-9>.
- Tatyana V. Bandos, Gustavo Camps-Valls, and Emilio Soria-Olivas. Statistical criteria for early-stopping of support vector machines. *Neurocomputing*, 70(13):2588–2592, 2007. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2006.12.019>. URL <https://www.sciencedirect.com/science/article/pii/S092523120700029X>. Selected papers from the 3rd International Conference on Development and Learning (ICDL 2004) Time series prediction competition: the CATS benchmark.
- Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997. URL <https://epubs.siam.org/doi/abs/10.1137/1.9781611974997>.
- Radu Ioan Bot, Jalal Fadili, and Dang-Khoa Nguyen. The Iterates of Nesterov’s Accelerated Algorithm Converge in The Critical Regimes, 2025. URL <https://arxiv.org/abs/2510.22715>.
- Peter Brucker. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3):163–166, 1984. ISSN 0167-6377. doi: [https://doi.org/10.1016/0167-6377\(84\)90010-5](https://doi.org/10.1016/0167-6377(84)90010-5). URL <https://www.sciencedirect.com/science/article/pii/0167637784900105>.
- Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. ISSN 1573-0565. doi: 10.1007/BF00994018. URL <https://doi.org/10.1007/BF00994018>.
- Felipe Cucker and Javier Peña. A primal-dual algorithm for solving polyhedral conic systems with a finite-precision machine. *SIAM Journal on Optimization*, 12(2):522–554, 2002. doi: 10.1137/S1052623401386794. URL <https://doi.org/10.1137/S1052623401386794>.
- Daniel Dadush, Sophie Huiberts, Bento Natura, and László A. Végh. A scaling-invariant algorithm for linear programming whose running time depends only on the constraint matrix, 2023. URL <https://arxiv.org/abs/1912.06252>.
- 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(63):1889–1918, 2005. URL <http://jmlr.org/papers/v6/fan05a.html>.
- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Lib-linear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008. ISSN 1532-4435.

- Robert M. Freund and Jorge R. Vera. Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm. *SIAM Journal on Optimization*, 10(1): 155–176, 1999. doi: 10.1137/S105262349732829X. URL <https://doi.org/10.1137/S105262349732829X>.
- Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th International Conference on Machine Learning*, ICML ’08, page 408–415, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390208. URL <https://doi.org/10.1145/1390156.1390208>.
- T. Joachims. Making large-Scale SVM Learning Practical. LS8-Report 24, Universität Dortmund, LS VIII-Report, 1998.
- Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. MNIST handwritten digit database. <https://web.archive.org/web/20220121031128/https://yann.lecun.org/exdb/mnist/>, 1998. Original: <https://yann.lecun.org/exdb/mnist/> (archived snapshot). Accessed: 2025-07-08.
- Haihao Lu and Jinwen Yang. On the geometry and refined rate of primal-dual hybrid gradient for linear programming, 2023a. arXiv:2307.03664 [math.OC].
- Haihao Lu and Jinwen Yang. On the infimal sub-differential size of primal-dual hybrid gradient method and beyond, 2023b. arXiv:2206.12061 [math.OC].
- Kohei Ogawa, Yoshiki Suzuki, and Ichiro Takeuchi. Safe screening of non-support vectors in pathwise svm computation. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1382–1390, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL <https://proceedings.mlr.press/v28/ogawa13b.html>.
- E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 276–285, 1997. doi: 10.1109/NNSP.1997.622408.
- John C. Platt. A Fast Algorithm for Training Support Vector Machines. Technical report, Microsoft Research, 1998.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, UK, 2014.
- Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal Estimated sub-GrAdient SOLver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, ICML ’07, page 807–814, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273598. URL <https://doi.org/10.1145/1273496.1273598>.

Zikai Xiong and Robert M. Freund. The Role of Level-Set Geometry on the Performance of PDHG for Conic Linear Optimization, 2024. URL <https://arxiv.org/abs/2406.01942>.

Julian Zimmert, Christian Schroeder de Witt, Giancarlo Kerg, and Marius Kloft. Safe screening for support vector machines. In *8th NIPS Workshop on Optimization for Machine Learning (OPT 2015)*, NeurIPS Workshop, 2015. URL <https://opt-ml.org/oldopt/opt15/index.html>.