Safe screening for support vector machines

Julian Zimmert
Department of Computer Science
Humboldt University of Berlin
julian.zimmert@informatik.hu-berlin.de

Christian Schroeder de Witt
Department of Computer Science
Humboldt University of Berlin
christian.schroeder@hu-berlin.de

Giancarlo Kerg
Department of Computer Science
Humboldt University of Berlin
giancarlo.kerg@hu-berlin.de

Marius Kloft
Department of Computer Science
Humboldt University of Berlin
kloft@hu-berlin.de

Abstract

The L_2-regularized hinge loss kernel SVM is one of the most important and most studied machine learning algorithms. Unfortunately, its computational training time complexity is generally unsuitable for big data. Empirical runtimes can however often be reduced using shrinking heuristics on the training sample set, which exploit the fact that non-support vectors do not affect the decision boundary. These shrinking heuristics are neither well understood nor especially reliable. We present the first safe removal bound for data points which does not rely on spectral properties of the kernel matrix. From there a relaxation provides us with a shrinking heuristic that is more reliable and performs favorably compared to a state-of-the-art shrinking heuristic suggested by Joachims [1], opening up an opportunity to improve the state of the art.

1 Introduction

Kernel-based learning algorithms [2] have found diverse applications due to their distinct merits such as their solid mathematical foundation [3] and modularity, which allows one to obtain non-linear learning algorithms from simpler linear ones in a canonical way. Particularly successful is the kernel support vector machine (kSVM) [4, 5], which has been shown to perform remarkably well across a wide range of problem settings [6]. Unfortunately, its worst-case training time complexity scales as $\mathcal{O}(n^2(d+n))$, where n is the number of training samples and d the dimensionality of the input space [7]. This generally prevents application to big data.

A first step towards reduced runtime complexity is achieved by employing exact *screening rules* or approximate *shrinking heuristics* that allow for the exclusion of training points prior to or early in the training process [8, 1]. The underlying principle here is that in kSVMs the decision boundary is represented as weighted average of so-called *support vectors*—meaning that safely-identified non-support vectors can be omitted from the training process. The shrinking heuristic used by two of the most commonly used state-of-the-art kSVM solvers, LIBSVM [9] and SVMLight[1], dates back to Joachims[1]. However, these shrinking heuristics are not well understood theoretically in the sense that there is a lack in theoretical bounds indicating when a training point may be safely removed or not. Which is why at some later stage one has to verify *a posteriori* whether every single previous individual sample omission was justified. If not so, the algorithm has to be warm-started after going through a costly *descreening* process of run-time complexity $\mathcal{O}(nn_{SV})$, where n_{SV} is the current number of support vectors.

In this paper, we derive *safe* sample removal bounds by exploiting the strong convexity properties of the kSVM primal objective, thus advancing ideas put forward in [8][10][11][12][13][14][15]. Unlike our predecessors, we however succeed in constructing a convex duality gap function in the primal variable. Let $\omega : \mathbb{R}^n \mapsto \mathcal{H}$ be the implicit linear mapping from the space of the dual variables α to the reproducing kernel Hilbert space (RKHS). Then our main contribution is, at the tth iteration, an exact bound on $||\omega(\alpha_t)-\omega(\alpha^*)||_{\mathcal{H}}$, where $(\cdot)^*$ denotes optimality. This implies a bound on the dual variable gradient at the optimum which, combined with standard optimality conditions (K.K.T.)[16], constitutes a demonstrably efficient safe screening rule on its own. Upon further relaxation, our safe screening rule yields a f-parameterized shrinking heuristic. We demonstrate that, while being less aggressive with data point removal, *f-safe shrinking* is more reliable than and performs favourably compared to Joachims' shrinking heuristic.

2 Related Work

Ghaoui et al. [15] were the first to develop a safe screening algorithm in order to remove training *features* from optimization problems that are sparse in the primal. Subsequently, Ogawa et al. [10] introduced a safe screening rule that allows for efficient *a priori* removal of kSVM training *samples*. However, their screening can only be applied after solving the unscreened kSVM problem at least twice in advance, which makes their method feasible only in computational path scenarios. Another loosely-related result building up on Ogawa et al. is the work by Ndiaye et al. [8].

Hsieh et al. [17] suggest a *divide-and-conquer* algorithm which allows for fast parallelization of large kSVM problems through initial training set m-segmentation using *kernel kmeans clustering* at $\mathcal{O}(nmd)$. The optimal solutions to the resulting m independently-solved kSVMs are then used to warm-start the global kSVM problem. [17] also present a screening rule that can be used to remove training samples prior to the conquer step, however, their associated removal bounds $\propto \lambda_{\text{min}}^{-1}$ are in practice meaningless as they require prior knowledge of λ_{min}, the smallest eigenvalue of the kernel matrix. Note that λ_{min} is not cheaply available and secondly can be very small or even zero when the kernel matrix is merely *semi*-definite.

Another attempt at improving kSVM performance is the work by *Steinwart et al.*[18], who do not offer any novel screening rules, but present an improved working set selection scheme for the dual variables.

3 Problem Setup and Notation

Let $\mathcal{X}$ and $\mathcal{Y}$ be input and output spaces, respectively, and let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X}\times\mathcal{Y}$ be a set of training samples. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel corresponding to a mapping $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS). The primal and dual optimization tasks of the kernel SVM are then to minimize the primal and maximize the dual objective functions $P : \mathcal{H} \to \mathbb{R}$ and $D : [0, C]^n \to \mathbb{R}$, respectively, defined as [19]

$$P(w) := \frac{1}{2}||w||_{\mathcal{H}}^2 + C\sum_{i=1}^{n} \left[(1 - y_i \left\langle \phi(x_i), w\right\rangle\right]_+ , \qquad D(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2}(\alpha \circ y)^T K(\alpha \circ y).$$

Here $[z]_+$ denotes $\max(0, z)$ and $C \in\,]0, +\infty[$ is the regularization parameter. Using the Lagrangian formalism it can be shown [19] that the primal and dual optima w^* and α^* are related by the linear function $\omega : [0, C]^n \to \mathcal{H}$ defined as $\omega(\alpha) := \sum_{i=1}^{n} y_i\alpha_i\phi(x_i)$, through the identity $w^* = \omega(\alpha^*)$. We define the duality gap functions $G_P : \mathcal{H} \mapsto \mathbb{R}$ and $G_D : [0, C]^n \mapsto \mathbb{R}$ as

$$G_P(w) := \min_{\substack{\alpha\in[0,C]^n:\\ \omega(\alpha)=w}} G_D(\alpha) = P(w) - \max_{\substack{\alpha\in[0,C]^n:\\ \omega(\alpha)=w}} D(\alpha)$$

and $G_D(\alpha) := P(\omega(\alpha)) - D(\alpha)$, respectively. Note that Slater's condition holds in the optimum [19] and hence $G_P(w^*) = 0 = G_D(\alpha^*)$.

4 Bounding the Primal Distance

In this section we present our main result: a bound on the gradients of the dual objective function, which we use in the subsequent section to derive a safe dual variable removal rule and a novel shrinking heuristic.

Proposition 4.1. *The duality gap $G_P : \omega\left([0, C]^n\right) \to \mathbb{R}$ is strongly convex with parameter 2 [16] satisfying:*

$$G_P(w_1) \geq G_P(w_2) + \langle\nabla G_P(w_2), w_1 - w_2\rangle + ||w_1 - w_2||_{\mathcal{H}}^2,\ \ \forall w_1, w_2 \in \omega\left([0, C]^n\right)$$

Proof. The definition of the function ω implies that $D(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2}||\omega(\alpha)||^2_{\mathcal{H}}$. Thus

$$G_P(w) = C\sum_{i=1}^n [(1 - y_i \langle \phi(x_i), w\rangle)]_+ + ||w||^2_{\mathcal{H}} - \underbrace{\max_{\substack{\alpha\in[0,C]^n:\\ \omega(\alpha)=w}} \sum_{i=1}^n \alpha_i}_{=:h(w)}$$

For invertible kernel matrices, the last term is trivially linear in w, for semi-definite kernels we can prove its convexity as follows:

Note that $h(w)$ is for any w the solution of a linear program. Hence, by strong duality,

$$h(w) = \min_{\lambda\in\mathcal{H}} \max_{\alpha\in[0,C]^n} \sum_{i=1}^n \alpha_i - \langle \omega(\alpha) - w, \lambda\rangle = \min_{\lambda\in\mathcal{H}} f(\lambda) + \langle w, \lambda\rangle = -g(w)$$

where $g(w) = \max_{\lambda\in\mathcal{H}} -f(\lambda) - \langle w, \lambda\rangle$ and $f(\lambda) = \max_{\alpha\in[0,C]^n} \sum_{i=1}^n \alpha_i - \langle \omega(\alpha), \lambda\rangle$. Danskin's theorem shows that both f and g are convex functions. Thus one can write $G_p(w) = A(w) + ||w||^2_{\mathcal{H}}$, where A is convex, and the proposition follows. □

Corollary 4.2. *Let w^* be the primal optimum. Then for all $\alpha \in [0, C]^n$, we have*

$$||\omega(\alpha) - w^*||_{\mathcal{H}} \leq \sqrt{G_D(\alpha)}$$

Proof. Reconsidering the strong convexity inequality from the above property, we have

$$\begin{aligned} G_P(\omega(\alpha)) &\geq G_P(w^*) + \langle \nabla G_P(w^*), \omega(\alpha) - w^*\rangle + ||\omega(\alpha) - w^*||^2_{\mathcal{H}} \\ &\geq G_P(w^*) + ||\omega(\alpha) - w^*||^2_{\mathcal{H}} \end{aligned}$$

where the last inequality follows from the optimality of w^* implying $\langle \nabla G_P(w^*), \omega(\alpha) - w^*\rangle \geq 0$. The strong duality of SVM implies $G_P(w^*) = 0$, thus $G_D(\alpha) \geq G_P(\omega(\alpha)) \geq ||\omega(\alpha) - w^*||^2_{\mathcal{H}}$. □

Corollary 4.3. *Let α^* be the dual optimum. Denote k_{ij} the entries of the associated kernel matrix, then for all $i = 1, \ldots, n$ and all $\alpha \in [0, C]^n$ we have:*

$$|\nabla D(\alpha^*)_i - \nabla D(\alpha)_i| \leq \sqrt{k_{ii} \cdot G_D(\alpha)}.$$

Proof. It is straightforward to see that for all $\alpha \in [0, C]^n$ we have $\nabla D(\alpha)_i = 1 - y_i \langle \omega(\alpha), \phi(x_i)\rangle$. Thus, in particular

$$\begin{aligned} \nabla D(\alpha^*)_i &= 1 - y_i \langle w^*, \phi(x_i)\rangle \\ &= 1 - y_i \langle \omega(\alpha), \phi(x_i)\rangle - y_i \langle w^* - \omega(\alpha), \phi(x_i)\rangle \\ &= \nabla D(\alpha)_i - y_i \langle w^* - \omega(\alpha), \phi(x_i)\rangle \end{aligned}$$

thus $|\nabla D(\alpha^*)_i - \nabla D(\alpha)_i| = |\langle w^* - \omega(\alpha), \phi(x_i)\rangle|$. We can bound the last term by Cauchy-Schwarz:

$$|\langle w^* - \omega(\alpha), \phi(x_i)\rangle| \leq ||w^* - \omega(\alpha)||_{\mathcal{H}} \sqrt{k_{ii}} \leq \sqrt{k_{ii} G_D(\alpha)}$$

where the last inequality follows from the previous corollary. □

5 Safe screening and f-Safe shrinking heuristic

Remark 5.1. *By the K.K.T. conditions [16] it holds in the optimal point:*

If $\nabla D(\alpha^*)_i > 0$, then $\alpha^*_i = C$; if $\nabla D(\alpha^*)_i < 0$, then $\alpha^*_i = 0$.

Using the bound of Corollary 4.3, we are able to give a safe screening rule as follows:

$$\boxed{\text{if } \nabla D(\alpha)_i > \sqrt{k_{ii} G_D(\alpha)}, \text{ then } \alpha^*_i = C; \text{ if } \nabla D(\alpha)_i < -\sqrt{k_{ii} G_D(\alpha)}, \text{ then } \alpha^*_i = 0.}$$

While the safe removal might not be sufficient to reduce the training sample set size, we can derive an efficient shrinking heuristic by introducing a factor $0 < f < 1$. The f-safe shrinking heuristic reads

$$\begin{aligned} \nabla D(\alpha_t)_i > f\sqrt{k_{ii} G_D(\alpha_t)} &\Rightarrow \text{ remove training point}(\alpha_i = C) \\ \nabla D(\alpha_t)_i < -f\sqrt{k_{ii} G_D(\alpha_t)} &\Rightarrow \text{ remove training point}(\alpha_i = 0) \end{aligned} \tag{1}$$

Algorithm 1: (SMO TYPE SVM DUAL SOLVER WITH F-SAFE SHRINKING).

---

input: *kernel matrix* $K = (k(x_i, x_j))_{i,j=1}^{n}$
labels $y_1, \ldots, y_n \in \{-1, 1\}$
optimization precision ε
initialize:
  $\nabla D_i := 1,\ \alpha_i := 0\ \forall i = 1, \ldots, n$
  Gap := nC
  $\mathcal{A} := \{1 \ldots n\}$
while $\neg$ optimality conditions satisfied within ϵ **do**
  while $\neg$ optimality conditions satisfied within ϵ **do**
    Working set optimization:
      update α for a working set $s \subset \mathcal{A}$
    Update: compute new ∇D_i for $i \in \mathcal{A}$ and Gap
    shrinking: reduce $\mathcal{A}$ by points satisfying (1)
  end while
  Reset working set: $\mathcal{A} := \{1 \ldots n\}$
  Update: compute new ∇D_i for $i \in \mathcal{A}$ and Gap
end while
output: ϵ-accurate α

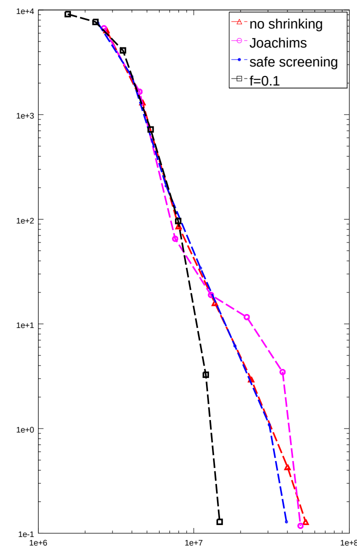

Figure 1: Duality gap (y-axis) vs. number of kernel evaluations (x-axis) [MNIST digits 3/5]

6 Preliminary Experiments

In the following, we study the performance of f-safe shrinking based on MNIST binary digit classification [20]. We use working sets of size one [18] with a stopping rule[9]. All experiments are implemented in MATLAB and use a RBF kernel ($\sigma = 6$) and $C = 1$.

Experiment A [Table 1] shows that our f-safe shrinking heuristic can be easily tuned to substantially outperform Joachims' shrinking. In future work, we will investigate model selection algorithms for f. **Experiment B** [Figure 1] shows that a hardly tuned f-safe shrinking heuristic can have a similar convergence rate as Joachims' shrinking, while effectively avoiding episodes of slow convergence due to remediation of false removals. Note that our safe screening rule only marginally outperforms unscreened kSVMs for practically large ϵ-precision.

Dual gap	*Safe screening*	*f=0.32*	*f=0.1*	*f=0.032*	*f=0.01*	*Joachims'*
at removal of 50%	8.3e-01	7.9e+00	6.1e+01	2.6e+02	**1.1e+03**	1.6e+02
at removal of 75%	3.5e-01	3.5e+00	3.3e+01	1.5e+02	**4.3e+02**	7.2e+01
at removal of 87.5%	0	1.0e+00	1.0e+01	6.5e+01	**2.3e+02**	2.7e+01
at removal of 93.8%	0	0	9.7e-01	1.5e+01	**6.0e+01**	9.3e+00
Reshrinkings	**0**	**0**	**0**	14	37	270
Kernel evaluations	6.2e+06	3.3e+06	**1.8e+06**	7.7e+06	4.4e+06	9.4e+06

Table 1: Typical performance metric [MNIST digits 0/7]

7 Conclusion and Outlook

Our key technical contribution is a proven safe screening rule that we extend to a high-performance shrinking heuristic. Preliminary experiments indicate that our f-safe shrinking heuristic consistently outperforms state of the art screening algorithms, including the one used by LIBSVM[1].

As immediate next steps, we will implement our approach in a runtime-optimized environment and extend the solver by an improved working set selection scheme [18], *kernel kmeans clustering* warm-starting [17] and with a hierarchical f-shrinking heuristic. Subsequently, we will present a family of parallelized solvers inspired by state-of-the-art DCSVM[17] and GTSVM[21]. We will also apply our results to SVR and other suitable algorithms. In doing this, we hope to contribute toward making kernel support vector methods big data-friendly.

Acknowledgments

This work was partly funded by the German Research Foundation (DFG) award KL 2698/2-1.

References

[1] T. Joachims, "Advances in kernel methods," ch. Making Large-scale Support Vector Machine Learning Practical, pp. 169–184, Cambridge, MA, USA: MIT Press, 1999.

[2] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[3] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[4] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (D. Haussler, ed.), pp. 144–152, 1992.

[5] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[6] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[7] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," tech. rep., ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.

[8] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon, "GAP Safe screening rules for sparse multi-task and multi-class models," *ArXiv e-prints*, June 2015.

[9] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, May 2011.

[10] K. Ogawa, Y. Suzuki, and I. Takeuchi, "Safe screening of non-support vectors in pathwise svm computation," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (S. Dasgupta and D. Mcallester, eds.), vol. 28, pp. 1382–1390, JMLR Workshop and Conference Proceedings, May 2013.

[11] Z. Xiang and P. Ramadge, "Fast lasso screening tests based on correlations," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 2137–2140, March 2012.

[12] Z. J. Xiang, H. Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *Advances in Neural Information Processing Systems 24* (J. Shawe-taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds.), pp. 900–908, 2011.

[13] L. Dai and K. Pelckmans, "An ellipsoid based, two-stage screening test for bpdn," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pp. 654–658, Aug 2012.

[14] J. Wang, P. Wonka, and J. Ye, "Lasso screening rules via dual polytope projection," *CoRR*, vol. abs/1211.3966, 2012.

[15] L. E. Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," *CoRR*, vol. abs/1009.3515, 2010.

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[17] C. Hsieh, S. Si, and I. S. Dhillon, "A divide-and-conquer solver for kernel support vector machines," *CoRR*, vol. abs/1311.0914, 2013.

[18] I. Steinwart, D. Hush, and C. Scovel, "Training svms without offset," *J. Mach. Learn. Res.*, vol. 12, pp. 141–202, Feb. 2011.

[19] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Neural Networks*, vol. 12, pp. 181–201, May 2001.

[20] Y. Lecun and C. Cortes, "The mnist database of handwritten digits,"

[21] A. Cotter, N. Srebro, and J. Keshet, "A gpu-tailored approach for training kernelized svms," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, (New York, NY, USA), pp. 805–813, ACM, 2011.