# The Kernelized Stochastic Batch Perceptron

Andrew Cotter, Shai Shalev-Shwartz, Nathan Srebro

{cotter,nati}@ttic.edu, shais@cs.huji.ac.il

## Overview

The kernelized Stochastic Batch Perceptron (SBP) is a fast kernel SVM optimization algorithm with learning runtime guarantees which are better than those of any other known approach. It also works well in practice, and a fast implementation (with source code) is available.

## Preliminaries

Let $x_i$ be a list of $n$ training vectors with associated labels $y_i$, and $K(x,x')$ a kernel function. We seek a set of coefficients $\alpha_i$ which will determine the classification of a previously-unseen testing example $x$ as:

$$\operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x)\right)$$

For notational simplicity, define:

$$Q_{ij} = y_i y_j K(x_i, x_j), c = Q\alpha$$

We call $c$ the vector of "responses".

## Generalization

| | Overall | $\epsilon = \Omega(L(u))$ |
|---|---|---|
| SBP | $\left(\frac{L(u)+\epsilon}{\epsilon}\right)^3 \frac{\lVert u\rVert^4}{\epsilon}$ | $\frac{\lVert u\rVert^4}{\epsilon}$ |
| Dual decomposition | $\left(\frac{L(u)+\epsilon}{\epsilon}\right)^2 \frac{\lVert u\rVert^4}{\epsilon^2}$ | $\frac{\lVert u\rVert^4}{\epsilon^2}$ |
| SGD on avg. hinge | $\left(\frac{L(u)+\epsilon}{\epsilon}\right) \frac{\lVert u\rVert^4}{\epsilon^3}$ | $\frac{\lVert u\rVert^4}{\epsilon^3}$ |

Upper bounds on the runtime (and number of kernel evaluations) required to guarantee $L(u)+\varepsilon$ generalization error. $L(u)$ is the expected hinge loss suffered by an arbitrary reference classifier $u$.

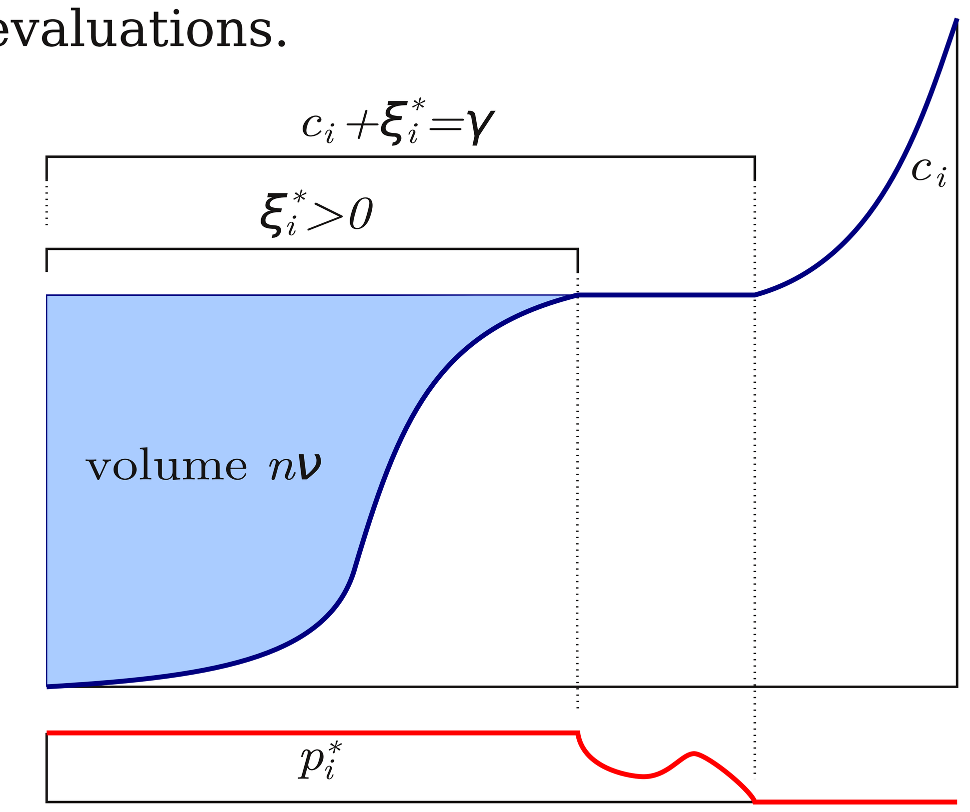
Dual decomposition approaches include most popular kernel SVM optimizers. SGD on the regularized objective (e.g. Pegasos) behaves similarly to SGD on the average hinge loss with a norm constraint (e.g. NORMA).

The relevant regime is often that in which the approximation and estimation errors are of the same order of magnitude (i.e. $\varepsilon=\Theta(L(u))$).

## SBP Algorithm

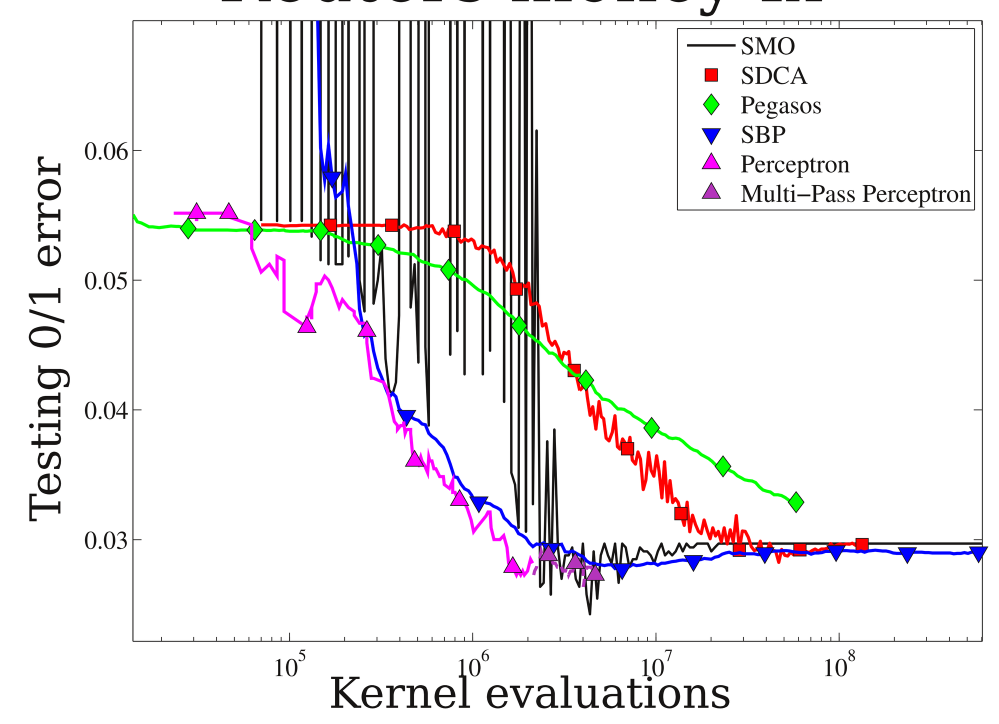
1. Use responses $c$ to find minimax-optimal $\xi^*$ and $p^*$
2. Sample $i$ from $p^*$
3. Let $\alpha_i = P(\alpha_i + \eta_i)$, and update the responses $c$

One can find minimax-optimal $\xi^*$ and $p^*$ using the "water-filling" procedure illustrated below (see our paper for details). This can be accomplished in $O(n)$ time using a divide-and-conquer algorithm. This cost is negligible, since we, like every other kernel SVM optimizer of which we are aware, must calculate one row of the kernel matrix per itation while updating the responses, at a cost of $n$ kernel evaluations.

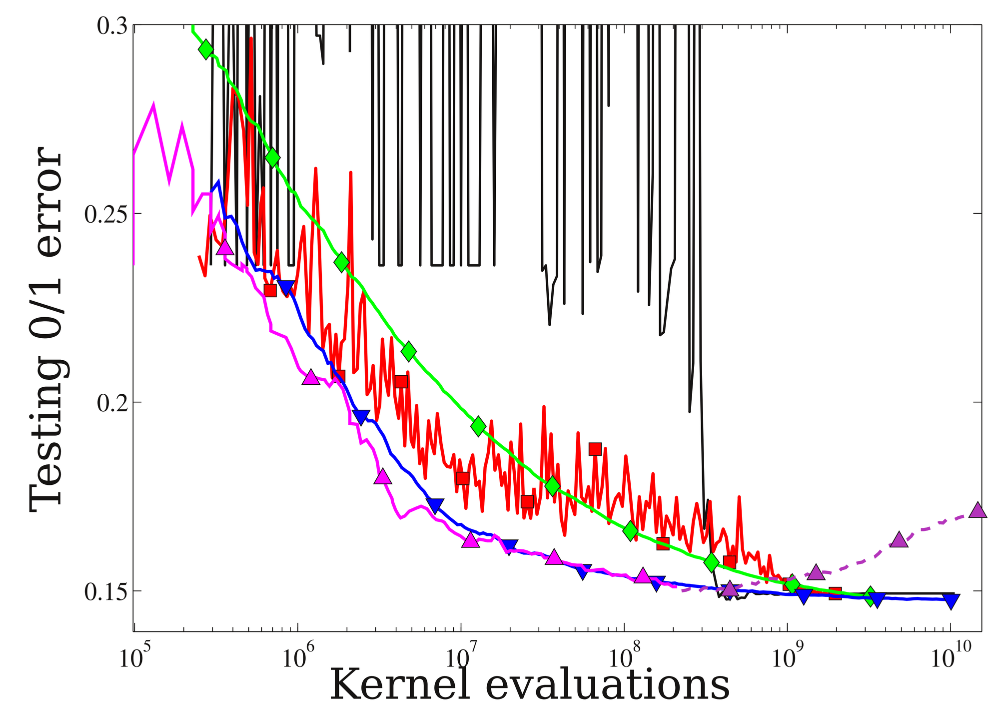



## Experiments

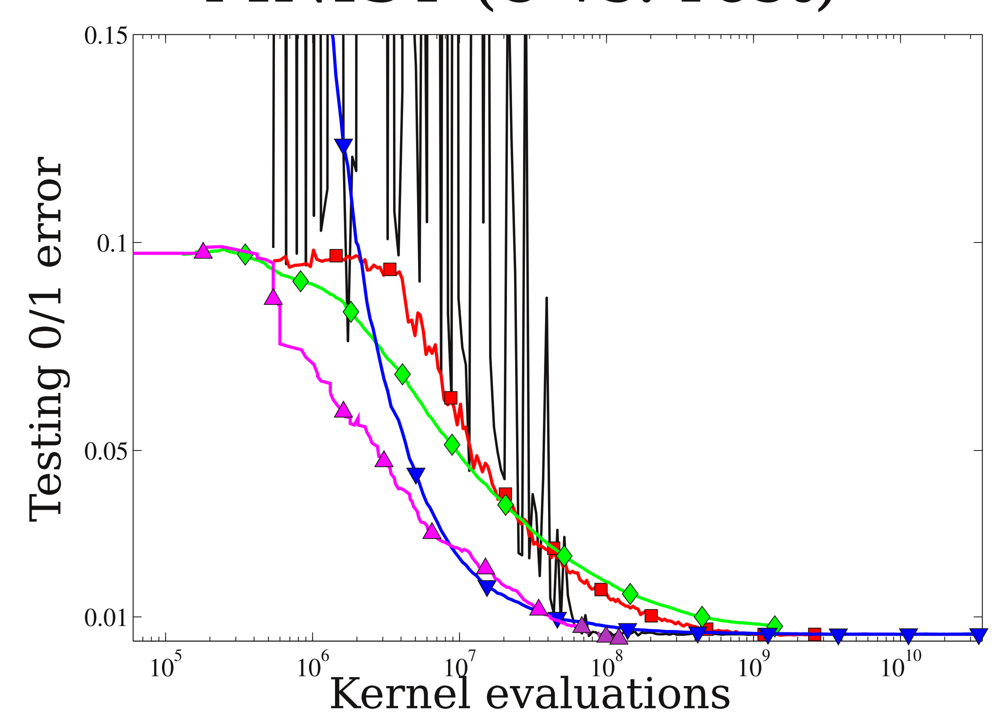
### Reuters money-fx




### Adult




### MNIST (8 vs. rest)




## Three Objective Functions

### Primal objective:

$$\underset{\alpha\in\mathbb{R}^n,\, b\in\mathbb{R}}{\text{minimize}} : \frac{1}{2}\alpha^T Q\alpha + C\sum_{i=1}^{n}\max(0, 1-c_i)$$

Often optimized using stochastic gradient descent (NORMA, Pegasos, ....)

### Dual objective:

$$\underset{\alpha\in\mathbb{R}^n}{\text{maximize}} : \mathbf{1}^T\alpha - \frac{1}{2}\alpha^T Q\alpha$$

$$\text{subject to} : \forall i\,(0 \le \alpha_i \le C)$$

Coordinate ascent-like algorithms (SMO, SVM-Light, LIBSVM, ....)

### Slack-constrained objective:

$$\underset{\alpha\in\mathbb{R}^n}{\text{maximize}} : \max_{\xi\in\mathbb{R}^n}\min_{p\in\Delta^n} p^T(c+\xi)$$

$$\text{subject to} : \alpha^T Q\alpha \le 1$$

$$: \xi \succeq 0, \mathbf{1}^T\xi \le n\nu$$

SIMBA (Hazan et al.) in linear setting. In Kernel setting, we use SGD.

All three of these objectives are equivalent in that by varying either $C$ or $\nu$ one explores the same Pareto optimal frontier (i.e. for every $C$ there exists a $\nu$ giving the same solution, and vice-versa).

For suboptimal solutions, exact equivalence breaks down: $\varepsilon^2$-suboptimal solutions to the dual objective may be only $\varepsilon$-suboptimal in the primal. However, $\varepsilon$-suboptimal solutions to the slack-constrained objective are *better* than $\varepsilon$-suboptimal in terms of average hinge loss (which is what the primal objective minimizes, plus regularization). Hence, SGD on the slack-constrained objective converges more rapidly than SGD on the primal *in terms of what the primal itself seeks to minimize*.

This leads to a better bound on generalization performance for SGD on the slack-constrained objective than that achieved by any other known method.

ttic.uchicago.edu/~cotter/projects/SBP