Hybrid Deterministic-Stochastic Methods for Data Fitting Addendum: Application to the Hinge Loss

Mark Schmidt and Armand Joulin

January 10, 2012

1 Hinge Loss

In the main paper [1], we present numerical experiments on various logistic regression models (§5). A widely-used alternative to logistic regression models are *margin-based* methods. Rather than seeking to optimize a regularized likelihood, these methods optimize a regularized convex penalty that upper bounds the misclassification error. This penalty is zero if the example is classified correctly by a sufficiently large margin, and grows as the margin decreases. Thus, penalties of this form have a special *sparse* structure in terms of i that we can take advantage of.

Although it applies more generally, we will demonstrate the usage of this sparse structure on the particular case of *smooth support vector machines* [2], where a classifier is estimated by solving the optimization problem

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{M} [\max\{0, 1 - b_i a_i^T x\}]^2 + \frac{\lambda}{2}||x||^2.$$

This problem is once-differentiable with a Lipschitz-continuous gradient, and is strongly-convex. For this problem, we typically expect the maximum to be zero for many i in the solution x_*. Thus, these instances i have no influence on the solution, in the sense that if we remove these data samples then x_* is not changed. This type of problem structure suggests a certain algorithmic modification to reduce the effective number of passes through the data set: we keep track of lower bounds on the quantities $b_i a_i^T x_k$ for all i in the current sample $\mathcal{B}_k$, and on iteration k we *ignore* the data samples i where we can guarantee that $b_i a_i^T x_k \geq 1$. This combines our growing batch strategy with a *shrinking* strategy, and if the problem is sparse in terms of i it allows us to achieve a linear convergence rate *without eventually looking at every data sample on every iteration.*

Obtaining a bound on $b_i a_i^T x_k$ is simple, we first consider the case where we evaluated $\gamma_{ki} \triangleq b_i a_i^T x_k$ on the previous iteration (and we either pre-compute the norms $||a_i||$ of the feature vectors or have an upper bound on them). If we denote the difference between successive iterations by

$$d_k = x_{k+1} - x_k,$$

then we can simply use

$$b_i a_i^T x_{k+1} = b_i a_i^T (x_k + d_k) = \gamma_{ki} + b_i a_i^T d_k \geq \gamma_{ki} - |a_i^T d_k| \geq \gamma_{ki} - ||a_i|| \cdot ||d_k||.$$

Thus, given the γ_{ki} for all i we can use this bound to prune the set of training samples that need to be re-visited on iteration $k+1$ for a cost of $O(n + |\mathcal{B}_{k+1}|)$, which is small in comparison to the normal iteration cost of $O(n|\mathcal{B}_{k+1}|)$.

In general we do not want our bound for example i on iteration $k+1$ to depend on γ_{ki}, since the goal of the shrinking procedure is to avoid computing γ_{ki} on every iteration. Fortunately, we can obtain a similar

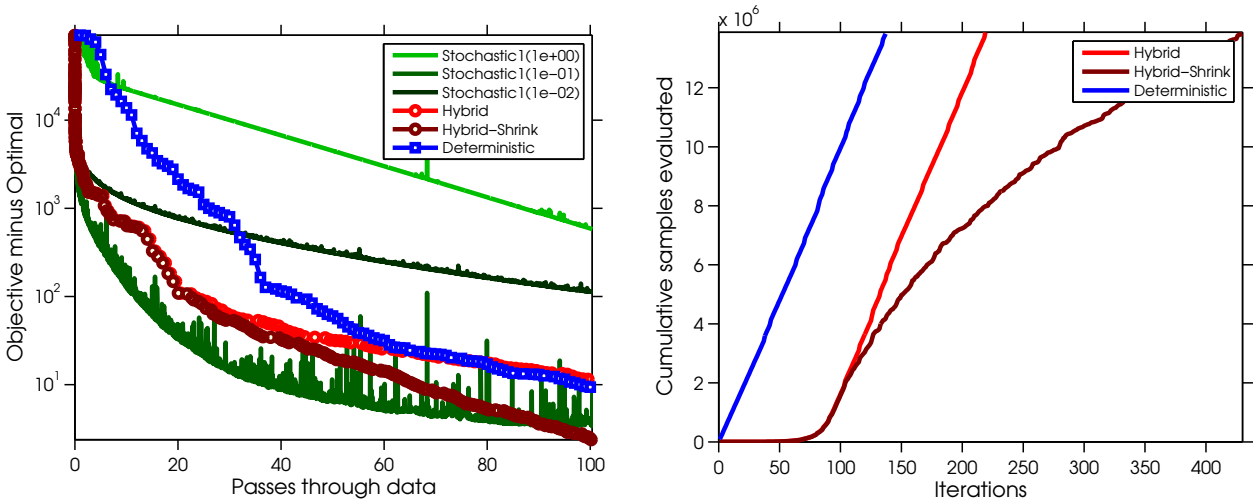

Figure 1: Smooth support vector machine experiments for different optimization strategies for spam classification. The stochastic method is run with 3 different fixed steplengths.

bound based on the latest γ_{ji} that was computed,

$$b_i a_i^T x_{k+1} = b_i a_i^T (x_j + \sum_{m=j}^{k} d_m) \geq \gamma_{ji} - ||a_i|| \sum_{m=j}^{k} ||d_m||. \tag{1}$$

By keeping track of the sum of the norms $\sum_{m=j}^{k} ||d_m||$, the cost of deciding which examples to exclude can again be implemented in $O(n + |\mathcal{B}_{k+1}|)$.

In Figure 1, we repeat our experiment on the spam data set but using the smooth support vector machine objective function. The method labeled *hybrid-shrink* uses the bound (1) to avoid evaluating elements in the batch where the bound guarantees they will not contribute to the objective function. In this plot we see that the stochastic method, with a carefully chosen step size, can outperform both the deterministic and hybrid methods. We expect that this is because the function is only once-differentiable, which may hurt the performance of methods that use a quasi-Newton Hessian approximation (though the stochastic method performs poorly without a carefully chosen step size). However, the *hybrid-shrink* method eventually outperforms all the other methods. Note that the plot on the right reflects that the shrinking has a greater effect the longer we run the method, which makes sense because we expect $||d_k||$ to start getting smaller as we approach the minimizer.

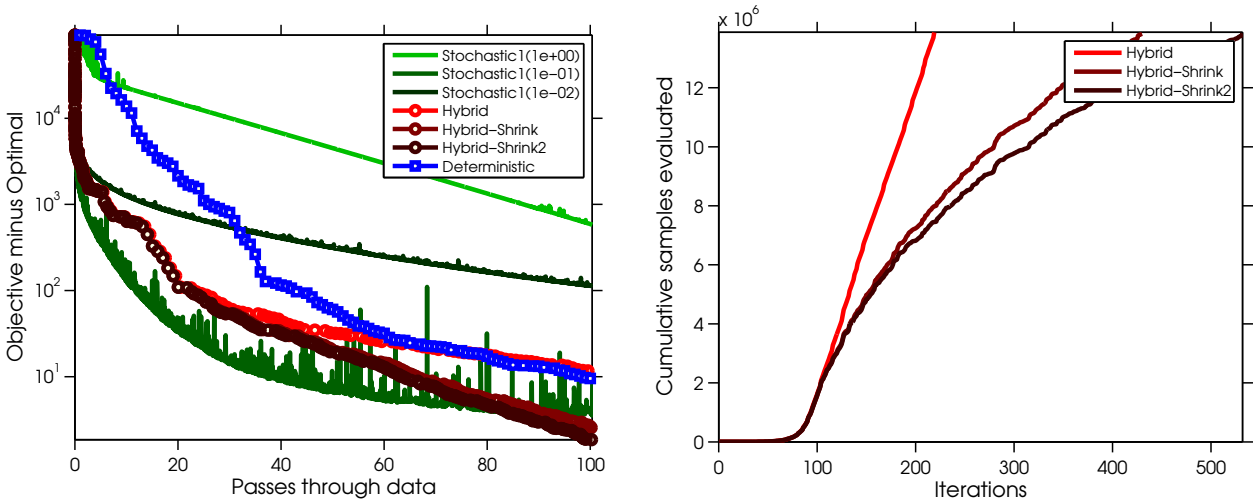

Figure 2: Smooth support vector machine experiments for different optimization strategies for spam classification. The stochastic method is run with 3 different fixed steplengths.

2 Better Bound

Instead of using the bound (1), we could use the bound

$$b_i a_i^T x_{k+1} = b_i a_i^T (x_j + \sum_{m=j}^{k} d_m) \geq \gamma_{ji} - ||a_i|| \cdot || \sum_{m=j}^{k} d_m||.$$

This bound is tighter than (1), but is more expensive to compute because we need to know previous d_k values rather than their norms. In particular, if we need to look back K iterations, then using this bound costs $O(Kn + |\mathcal{B}_{k+1}|)$. We plot the results when the hybrid method uses this bound above, where we see that it gives a further small improvement.

References

[1] M.P. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *Arxiv preprint arXiv:1104.2373*, 2011.

[2] Y.-J. Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. *Computational Optimization and Applications*, 20:5–22, 2001.