Table 2: Experimental results on the four binary classification tasks derived from RCV1. ”Train” denotes the number of training corrections, while ”Test” gives the fraction of misclassified patterns in the test set. Only the results corresponding to the best test set accuracy are shown. In bold are the smallest figures achieved for each of the 8 combinations of dataset (RCV1_x, $x = 70, 101, 4, 59$) and phase (training or test).

	FO		HO_2		SO	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
RCV1_{70}	993	7.20%	941	**6.83%**	**880**	6.95%
RCV1_{101}	673	6.39%	**665**	5.81%	677	**5.48%**
RCV1_{4}	803	6.14%	**783**	**5.94%**	819	6.05%
RCV1_{59}	767	6.45%	762	**6.04%**	**760**	6.84%

Table 3: Experimental results on the OCR tasks. ”Train” denotes the total number of training corrections, summed over the 10 categories, while ”Test” denotes the fraction of misclassified patterns in the test set. Only the results corresponding to the best test set accuracy are shown. For the sparse version of HO_2 we also reported (in parentheses) the number of matrix updates during training. In bold are the smallest figures achieved for each of the 8 combinations of dataset (USPS or MNIST), kernel type (Gaussian or Polynomial), and phase (training or test).

		FO		HO_2		*Sparse* HO_2		SO	
		TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
USPS	GAUSS	1385	6.53%	**945**	**4.76%**	965 (440)	5.13%	1003	5.05%
	POLY	1609	7.37%	1090	5.71%	1081 (551)	**5.52%**	**1054**	5.53%
MNIST	GAUSS	5834	2.10%	**5351**	**1.79%**	5363 (2596)	1.81%	5684	1.82%
	POLY	8148	3.04%	**6404**	2.27%	6476 (3311)	2.28%	6440	**2.03%**

[7] N. Cesa-Bianchi, A. Conconi & C. Gentile (2005). A second-order perceptron algorithm. *SIAM Journal of Computing*, 34(3), 640–668.

[8] N. Cesa-Bianchi, C. Gentile, & L. Zaniboni (2006). Worst-case analysis of selective sampling for linear-threshold algorithms. *JMLR*, 7, 1205–1230.

[9] C. Cortes & V. Vapnik (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

[10] O. Dekel, S. Shalev-Shwartz, & Y. Singer (2006). The Forgetron: a kernel-based Perceptron on a fixed budget. *NIPS 18*, MIT Press, pp. 259–266.

[11] C. Gentile (2001). A new approximate maximal margin classification algorithm. *JMLR*, 2, 213–242.

[12] C. Gentile (2003). The Robustness of the p-norm Algorithms. *Machine Learning*, 53(3), pp. 265–299.

[13] A.J. Grove, N. Littlestone & D. Schuurmans (2001). General convergence results for linear discriminant updates. *Machine Learning Journal*, 43(3), 173–210.

[14] S. Gruvberger, et al. (2001). Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.*, *61*, 5979–5984.

[15] J. Kivinen, M.K. Warmuth, & P. Auer (1997). The perceptron algorithm vs. winnow: linear vs. logarithmic mistake bounds when few input variables are relevant. *Artificial Intelligence*, 97, 325–343.

[16] Y. Le Cun, et al. (1995). Comparison of learning algorithms for handwritten digit recognition. *ICANN 1995*, pp. 53–60.

[17] Y. Li & P. Long (2002). The relaxed online maximum margin algorithm. *Machine Learning*, 46(1-3), 361–387.

[18] N. Littlestone (1988). Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2(4), 285–318.

[19] N. Littlestone & M.K. Warmuth (1994). The weighted majority algorithm. *Information and Computation*, 108(2), 212–261.

[20] P. Long & X. Wu (2004). Mistake bounds for maximum entropy discrimination. *NIPS 2004*.

[21] A.B.J. Novikov (1962). On convergence proofs on perceptrons. *Proc. of the Symposium on the Mathematical Theory of Automata, vol. XII*, pp. 615–622.

[22] Reuters: 2000. `http://about.reuters.com/researchandstandards/corpus/`.

[23] S. Shalev-Shwartz & Y. Singer (2006). Online Learning Meets Optimization in the Dual. *COLT 2006*, pp. 423–437.

[24] B. Schoelkopf & A. Smola (2002). *Learning with kernels*. MIT Press.

[25] Vovk, V. (2001). Competitive on-line statistics. International Statistical Review, 69, 213-248.