INTRODUCTION TO MACHINE LEARNING

Introduction to Machine Learning

Alex Smola and S.V.N. Vishwanathan

Yahoo! Labs
Santa Clara
–and–
Departments of Statistics and Computer Science
Purdue University
–and–
College of Engineering and Computer Science
Australian National University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

Typeface Monotype Times 10/13pt *System* LaTeX 2ε [ALEXANDER J. SMOLA AND S.V.N. VISHWANATHAN]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data available

ISBN 0 521 82583 0 hardback

AUTHOR: VISHY
REVISION: 256
TIMESTAMP: DECEMBER 6, 2011
URL: SVN://VISHY@REPOS.STAT.PURDUE.EDU/THEBOOK/TRUNK/BOOK/THEBOOK.TEX

Contents

1

Online Learning and Boosting

So far the learning algorithms we considered assumed that all the training data is available before building a model for predicting labels on unseen data points. In many modern applications data is available only in a streaming fashion, and one needs to predict labels on the fly. To describe a concrete example, consider the task of spam filtering. As emails arrive the learning algorithm needs to classify them as spam or ham. Tasks such as these are tackled via online learning. Online learning proceeds in rounds. At each round a training example is revealed to the learning algorithm, which uses its current model to predict the label. The true label is then revealed to the learner which incurs a loss and updates its model based on the feedback provided. This protocol is summarized in Algorithm 1.1. The goal of online learning is to minimize the total loss incurred. By an appropriate choice of labels and loss functions, this setting encompasses a large number of tasks such as classification, regression, and density estimation. In our spam detection example, if an email is misclassified the user can provide feedback which is used to update the spam filter, and the goal is to minimize the number of misclassified emails.

1.1 Halving Algorithm

The halving algorithm is conceptually simple, yet it illustrates many of the concepts in online learning. Suppose we have access to a set of n experts, that is, functions f_i which map from the input space $\mathcal{X}$ to the output space $\mathcal{Y} = \{\pm 1\}$. Furthermore, assume that one of the experts is consistent, that is, there exists a $j \in \{1, \ldots, n\}$ such that $f_j(x_t) = y_t$ for $t = 1, \ldots, T$. The halving algorithm maintains a set $\mathcal{C}_t$ of consistent experts at time t. Initially $\mathcal{C}_0 = \{1, \ldots, n\}$, and it is updated recursively as

$$\mathcal{C}_{t+1} = \{i \in \mathcal{C}_t \text{ s.t. } f_i(x_{t+1}) = y_{t+1}\}. \tag{1.1}$$

The prediction on a new data point is computed via a majority vote amongst the consistent experts: $\hat{y}_t = \text{majority}(\mathcal{C}_t)$.

Lemma 1.1 *The Halving algorithm makes at most* $\log_2(n)$ *mistakes.*

Algorithm 1.1 Protocol of Online Learning

for $t = 1, \ldots, T$ do **do**
  Get training instance x_t
  Predict label $\hat{y}_t$
  Get true label y_t
  Incur loss $l(\hat{y}_t, x_t, y_t)$
  Update model
end for

Proof Let M denote the total number of mistakes. The halving algorithm makes a mistake at iteration t if at least half the consistent experts $\mathcal{C}_t$ predict the wrong label. This in turn implies that

$$|\mathcal{C}_{t+1}| \leq \frac{|\mathcal{C}_t|}{2} \leq \frac{|\mathcal{C}_0|}{2^M} = \frac{n}{2^M}.$$

On the other hand, since one of the experts is consistent it follows that $1 \leq |\mathcal{C}_{t+1}|$. Therefore, $2^M \leq n$. Solving for M completes the proof. ■

1.2 Weighted Majority

We now turn to the scenario where none of the experts is consistent. Therefore, the aim here is not to minimize the number mistakes but to minimize regret.

1.3 Stochastic Mirror Descent

In this section we will consider optimization algorithms for solving the following problem:

$$\min_{w \in \Omega} J(w) \text{ where } J(w) = \sum_{t=1}^{T} f_t(w). \tag{1.2}$$

Suppose we have access to a function ψ which is continuously differentiable and strongly convex with modulus of strong convexity $\sigma > 0$ (see Section **??** for definition of strong convexity), then we can define the Bregman divergence (**??**) corresponding to ψ as

$$\Delta_\psi(w, w') = \psi(w) - \psi(w') - \left\langle w - w', \nabla\psi(w') \right\rangle.$$

Algorithm 1.2 Stochastic Mirror Descent

Input: Initial point w_1, maximum iterations T
for $t = 1, \ldots, T$ **do**
  Compute $\hat{w}_{t+1} = \nabla\psi^* \left(\nabla\psi(w_t) - \eta_t g_t\right)$ with $g_t := \partial_w f_t(w_t)$
  Set $w_{t+1} = P_{\psi,\Omega}\left(\hat{w}_{t+1}\right)$
end for
Return: w_{T+1}

We can also generalize the orthogonal projection (**??**) by replacing the square Euclidean norm with the above Bregman divergence:

$$P_{\psi,\Omega}(w') = \operatorname*{argmin}_{w\in\Omega} \Delta_\psi(w, w'). \tag{1.3}$$

Denote $w^* = P_{\psi,\Omega}(w')$. Just like the Euclidean distance is non-expansive, the Bregman projection can also be shown to be non-expansive in the following sense:

$$\Delta_\psi(w, w') \geq \Delta_\psi(w, w^*) + \Delta_\psi(w^*, w') \tag{1.4}$$

for all $w \in \Omega$. The diameter of Ω as measured by Δ_ψ is given by

$$\operatorname{diam}_\psi(\Omega) = \max_{w,w'\in\Omega} \Delta_\psi(w, w'). \tag{1.5}$$

For the rest of this chapter we will make the following standard assumptions:

- Each f_t is convex and revealed at time instance t.
- Ω is a closed convex subset of $\mathbb{R}^n$ with non-empty interior.
- The diameter $\operatorname{diam}_\psi(\Omega)$ of Ω is bounded by $F < \infty$.
- The set of optimal solutions of (1.2) denoted by Ω^* is non-empty.
- The subgradient $\partial_w f_t(w)$ can be computed for every t and $w \in \Omega$.
- The Bregman projection (1.3) can be computed for every $w' \in \mathbb{R}^n$.
- The gradient $\nabla\psi$, and its inverse $(\nabla\psi)^{-1} = \nabla\psi^*$ can be computed.

The method we employ to solve (1.2) is given in Algorithm 1.2. Before analyzing the performance of the algorithm we would like to discuss three special cases. First, Euclidean distance squared which recovers projected stochastic gradient descent, second Entropy which recovers Exponentiated gradient descent, and third the p-norms for $p > 2$ which recovers the p-norm Perceptron. BUGBUG TODO.

Our key result is Lemma 1.3 given below. It can be found in various guises in different places most notably Lemma 2.1 and 2.2 in [Ned02], Theorem 4.1 and Eq. (4.21) and (4.15) in [BT03], in the proof of Theorem 1 of [Zin03], as well as Lemma 3 of [SSS07]. We prove a slightly general variant; we allow for projections with an arbitrary Bregman divergence and also take into account a generalized version of strong convexity of f_t. Both these modifications will allow us to deal with general settings within a unified framework.

Definition 1.2 *We say that a convex function f is strongly convex with respect to another convex function ψ with modulus λ if*

$$f(w) - f(w') - \langle w - w', \mu \rangle \geq \lambda \Delta_\psi(w, w') \text{ for all } \mu \in \partial f(w'). \tag{1.6}$$

The usual notion of strong convexity is recovered by setting $\psi(\cdot) = \frac{1}{2} \|\cdot\|^2$.

Lemma 1.3 *Let f_t be strongly convex with respect to ψ with modulus $\lambda \geq 0$ for all t. For any $w \in \Omega$ the sequences generated by Algorithm 1.2 satisfy*

$$\Delta_\psi(w, w_{t+1}) \leq \Delta_\psi(w, w_t) - \eta_t \langle g_t, w_t - w \rangle + \frac{\eta_t^2}{2\sigma} \|g_t\|^2 \tag{1.7}$$

$$\leq (1 - \eta_t \lambda) \Delta_\psi(w, w_t) - \eta_t (f_t(w_t) - f_t(w)) + \frac{\eta_t^2}{2\sigma} \|g_t\|^2 . \tag{1.8}$$

Proof We prove the result in three steps. First we upper bound $\Delta_\psi(w, w_{t+1})$ by $\Delta_\psi(w, \hat{w}_{t+1})$. This is a consequence of (1.4) and the non-negativity of the Bregman divergence which allows us to write

$$\Delta_\psi(w, w_{t+1}) \leq \Delta_\psi(w, \hat{w}_{t+1}). \tag{1.9}$$

In the next step we use Lemma **??** to write

$$\Delta_\psi(w, w_t) + \Delta_\psi(w_t, \hat{w}_{t+1}) - \Delta_\psi(w, \hat{w}_{t+1}) = \langle \nabla\psi(\hat{w}_{t+1}) - \nabla\psi(w_t), w - w_t \rangle .$$

Since $\nabla\psi^* = (\nabla\psi)^{-1}$, the update in step 3 of Algorithm 1.2 can equivalently be written as $\nabla\psi(\hat{w}_{t+1}) - \nabla\psi(w_t) = -\eta_t g_t$. Plugging this in the above equation and rearranging

$$\Delta_\psi(w, \hat{w}_{t+1}) = \Delta_\psi(w, w_t) - \eta_t \langle g_t, w_t - w \rangle + \Delta_\psi(w_t, \hat{w}_{t+1}). \tag{1.10}$$

Finally we upper bound $\Delta_\psi(w_t, \hat{w}_{t+1})$. For this we need two observations: First, $\langle x, y \rangle \leq \frac{1}{2\sigma} \|x\|^2 + \frac{\sigma}{2} \|y\|^2$ for all $x, y \in \mathbb{R}^n$ and $\sigma > 0$. Second, the σ strong convexity of ψ allows us to bound $\Delta_\psi(\hat{w}_{t+1}, w_t) \geq \frac{\sigma}{2} \|w_t - \hat{w}_{t+1}\|^2$.

Using these two observations

$$\begin{aligned}
\Delta_\psi(w_t, \hat{w}_{t+1}) &= \psi(w_t) - \psi(\hat{w}_{t+1}) - \langle \nabla\psi(\hat{w}_{t+1}), w_t - \hat{w}_{t+1}\rangle \\
&= -(\psi(\hat{w}_{t+1}) - \psi(w_t) - \langle \nabla\psi(w_t), \hat{w}_{t+1} - w_t\rangle) + \langle \eta_t g_t, w_t - \hat{w}_{t+1}\rangle \\
&= -\Delta_\psi(\hat{w}_{t+1}, w_t) + \langle \eta_t g_t, w_t - \hat{w}_{t+1}\rangle \\
&\leq -\frac{\sigma}{2}\left\|w_t - \hat{w}_{t+1}\right\|^2 + \frac{\eta_t^2}{2\sigma}\left\|g_t\right\|^2 + \frac{\sigma}{2}\left\|w_t - \hat{w}_{t+1}\right\|^2 \\
&= \frac{\eta_t^2}{2\sigma}\left\|g_t\right\|^2. \qquad (1.11)
\end{aligned}$$

Inequality (1.7) follows by putting together (1.9), (1.10), and (1.11), while (1.8) follows by using (1.6) with $f = f_t$ and $w' = w_t$ and substituting into (1.7). ■

Now we are ready to prove regret bounds.

Lemma 1.4 *Let $w^* \in \Omega^*$ denote the best parameter chosen in hindsight, and let $\|g_t\| \leq L$ for all t. Then the regret of Algorithm 1.2 can be bounded via*

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq F\left(\frac{1}{\eta_T} - T\lambda\right) + \frac{L^2}{2\sigma}\sum_{t=1}^{T}\eta_t. \qquad (1.12)$$

Proof Set $w = w^*$ and rearrange (1.8) to obtain

$$f_t(w_t) - f_t(w^*) \leq \frac{1}{\eta_t}\left((1 - \lambda\eta_t)\Delta_\psi(w^*, w_t) - \Delta_\psi(w^*, w_{t+1})\right) + \frac{\eta_t}{2\sigma}\left\|g_t\right\|^2.$$

Summing over t

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq \underbrace{\sum_{t=1}^{T}\frac{1}{\eta_t}\left((1 - \eta_t\lambda)\Delta_\psi(w^*, w_t) - \Delta_\psi(w^*, w_{t+1})\right)}_{T_1} + \underbrace{\sum_{t=1}^{T}\frac{\eta_t}{2\sigma}\left\|g_t\right\|^2}_{T_2}.$$

Since the diameter of Ω is bounded by F and Δ_ψ is non-negative

$$\begin{aligned}
T_1 &= \left(\frac{1}{\eta_1} - \lambda\right)\Delta_\psi(w^*, w_1) - \frac{1}{\eta_T}\Delta_\psi(w^*, w_{T+1}) + \sum_{t=2}^{T}\Delta_\psi(w^*, w_t)\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \lambda\right) \\
&\leq \left(\frac{1}{\eta_1} - \lambda\right)\Delta_\psi(w^*, w_1) + \sum_{t=2}^{T}\Delta_\psi(w^*, w_t)\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \lambda\right) \\
&\leq \left(\frac{1}{\eta_1} - \lambda\right)F + \sum_{t=2}^{T}F\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \lambda\right) = F\left(\frac{1}{\eta_T} - T\lambda\right).
\end{aligned}$$

On the other hand, since the subgradients are Lipschitz continuous with constant L it follows that

$$T_2 \leq \frac{L^2}{2\sigma} \sum_{t=1}^{T} \eta_t.$$

Putting together the bounds for T_1 and T_2 yields (1.12). ■

Corollary 1.5 *If $\lambda > 0$ and we set $\eta_t = \frac{1}{\lambda t}$ then*

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \frac{L^2}{2\sigma\lambda}(1 + \log(T)),$$

On the other hand, when $\lambda = 0$, if we set $\eta_t = \frac{1}{\sqrt{t}}$ then

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \left(F + \frac{L^2}{\sigma}\right)\sqrt{T}.$$

Proof First consider $\lambda > 0$ with $\eta_t = \frac{1}{\lambda t}$. In this case $\frac{1}{\eta_T} = T\lambda$, and consequently (1.12) specializes to

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq \frac{L^2}{2\sigma\lambda} \sum_{t=1}^{T} \frac{1}{t} \leq \frac{L^2}{2\sigma\lambda}(1 + \log(T)).$$

When $\lambda = 0$, and we set $\eta_t = \frac{1}{\sqrt{t}}$ and use problem 1.2 to rewrite (1.12) as

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq F\sqrt{T} + \frac{L^2}{\sigma} \sum_{t=1}^{T} \frac{1}{2\sqrt{t}} \leq F\sqrt{T} + \frac{L^2}{\sigma}\sqrt{T}.$$

■

1.3.1 Dealing with Composite Objective Functions

Next we consider algorithms for solving the following so-called composite problem:

$$\min_{w \in \Omega} J(w) + r(w) \text{ where } J(w) = \sum_{t=1}^{T} f_t(w), \tag{1.13}$$

and $r(w)$ is a simple to evaluate regularizer. For instance, $r(w) = \|w\|^2$ or $r(w) = \|w\|_1^2$ etc. We will operate under the same assumptions as in

Algorithm 1.3 Stochastic Mirror Descent for Composite Functions

1: **Input:** Initial point w_1, maximum iterations T
2: **for** $t = 1, \ldots, T$ **do**
3: Compute $\hat{w}_{t+1} = \operatorname{argmin}_w \eta_t \langle g_t, w \rangle + \eta r(w) + \Delta_\psi(w, w_t)$ with $g_t := \partial_w f_t(w_t)$
4: Set $w_{t+1} = P_{\psi,\Omega}(\hat{w}_{t+1})$
5: **end for**
6: **Return:** w_{T+1}

the previous sub-section. The algorithm that we will employ is given in Algorithm 1.3. Note that Algorithm 1.2 is recovered as a special case when $r(w) = 0$. Now we prove the analog of Lemma 1.3 for composite functions.

Lemma 1.6 *Let f_t be strongly convex with respect to ψ with modulus $\lambda \geq 0$ for all t. For any $w \in \Omega$ the sequences generated by Algorithm 1.2 satisfy*

$$\Delta_\psi(w, w_{t+1}) \leq \Delta_\psi(w, w_t) - \eta_t \langle g_t, w_t - w \rangle - \eta_t \langle \nabla r(w_{t+1}), w_{t+1} - w \rangle + \frac{\eta_t^2}{2\sigma} \|g_t\|^2 \tag{1.14}$$

$$\leq (1 - \eta_t \lambda)\Delta_\psi(w, w_t) - \eta_t(f_t(w_t) - f_t(w)) - \eta_t(r(w_{t+1}) - r(w)) + \frac{\eta_t^2}{2\sigma} \|g_t\|^2 . \tag{1.15}$$

Proof We prove the result in three steps. First we upper bound $\Delta_\psi(w, w_{t+1})$ by $\Delta_\psi(w, \hat{w}_{t+1})$. This is a consequence of (1.4) and the non-negativity of the Bregman divergence which allows us to write

$$\Delta_\psi(w, w_{t+1}) \leq \Delta_\psi(w, \hat{w}_{t+1}). \tag{1.16}$$

In the next step we use Lemma **??** to write

$$\Delta_\psi(w, w_t) + \Delta_\psi(w_t, \hat{w}_{t+1}) - \Delta_\psi(w, \hat{w}_{t+1}) = \langle \nabla\psi(\hat{w}_{t+1}) - \nabla\psi(w_t), w - w_t \rangle .$$

Since $\nabla\psi^* = (\nabla\psi)^{-1}$, the update in step 3 of Algorithm 1.3 can equivalently be written as $\nabla\psi(\hat{w}_{t+1}) - \nabla\psi(w_t) = -\eta_t g_t - \eta_t \nabla r(w_{t+1})$. Plugging this in the above equation and rearranging

$$\Delta_\psi(w, \hat{w}_{t+1}) = \Delta_\psi(w, w_t) - \eta_t \langle g_t, w_t - w \rangle - \eta_t \langle \nabla r(w_{t+1}), w_t - w \rangle + \Delta_\psi(w_t, \hat{w}_{t+1}). \tag{1.17}$$

Finally we upper bound $\Delta_\psi(w_t, \hat{w}_{t+1})$. For this we need two observations: First, $\langle x, y \rangle \leq \frac{1}{2\sigma} \|x\|^2 + \frac{\sigma}{2} \|y\|^2$ for all $x, y \in \mathbb{R}^n$ and $\sigma > 0$. Second, the σ strong convexity of ψ allows us to bound $\Delta_\psi(\hat{w}_{t+1}, w_t) \geq \frac{\sigma}{2} \|w_t - \hat{w}_{t+1}\|^2$.

Using these two observations

$$\begin{aligned}
\Delta_\psi(w_t, \hat{w}_{t+1}) &= \psi(w_t) - \psi(\hat{w}_{t+1}) - \langle \nabla\psi(\hat{w}_{t+1}), w_t - \hat{w}_{t+1}\rangle \\
&= -(\psi(\hat{w}_{t+1}) - \psi(w_t) - \langle \nabla\psi(w_t), \hat{w}_{t+1} - w_t\rangle) \\
&\quad + \langle \eta_t g_t, w_t - \hat{w}_{t+1}\rangle + \eta_t \langle \nabla r(w_{t+1}), w_t - \hat{w}_{t+1}\rangle \\
&= -\Delta_\psi(\hat{w}_{t+1}, w_t) + \langle \eta_t g_t, w_t - \hat{w}_{t+1}\rangle + \eta_t \langle \nabla r(w_{t+1}), w_t - \hat{w}_{t+1}\rangle \\
&\leq -\frac{\sigma}{2}\|w_t - \hat{w}_{t+1}\|^2 + \frac{\eta_t^2}{2\sigma}\|g_t\|^2 + \frac{\sigma}{2}\|w_t - \hat{w}_{t+1}\|^2 + \eta_t \langle \nabla r(w_{t+1}), w_t - \hat{w}_{t+1}\rangle \\
&= \frac{\eta_t^2}{2\sigma}\|g_t\|^2 + \eta_t \langle \nabla r(w_{t+1}), w_t - \hat{w}_{t+1}\rangle . \qquad (1.18)
\end{aligned}$$

Inequality (1.14) follows by putting together (1.16), (1.17), (1.18), and simplifying while (1.15) follows by using (1.6) with $f = f_t$ and $w' = w_t$ and substituting into (1.14). ■

Problems

Problem 1.1 (Generalized Cauchy-Schwartz {1}) *Show that* $\langle x, y\rangle \leq \frac{1}{2\sigma}\|x\|^2 + \frac{\sigma}{2}\|y\|^2$ *for all* $x, y \in \mathbb{R}^n$ *and* $\sigma > 0$.

Problem 1.2 (Bounding sum of a series {1}) *Show that* $\sum_{t=a}^{b} \frac{1}{2\sqrt{t}} \leq \sqrt{b - a + 1}$. ***Hint:*** *Upper bound the sum by an integral.*

Bibliography

[ABB+00] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, *Gene ontology: tool for the unification of biology. the gene ontology consortium*, Nat Genet **25** (2000), 25–29.

[Ach03] D. Achlioptas, *Database-friendly random projections: Johnson-lindenstrauss with binary coins*, J. Comput. Syst. Sci **66** (2003), no. 4, 671–687.

[BH04] J. Basilico and T. Hofmann, *Unifying collaborative and content-based filtering*, Proceedings of the International Conference on Machine Learning (New York, NY), ACM Press, 2004, pp. 65–72.

[BHK98] J. S. Breese, D. Heckerman, and C. Kardie, *Empirical analysis of predictive algorithms for collaborative filtering*, Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998, pp. 43–52.

[BHS+07] G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. V. N. Vishwanathan, *Predicting structured data*, MIT Press, Cambridge, Massachusetts, 2007.

[BM92] K. P. Bennett and O. L. Mangasarian, *Robust linear programming discrimination of two linearly inseparable sets*, Optim. Methods Softw. **1** (1992), 23–34.

[BT03] Amir Beck and Marc Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), no. 3, 167–175.

[CDLS99] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter, *Probabilistic networks and expert sytems*, Springer, New York, 1999.

[CH04] Lijuan Cai and T. Hofmann, *Hierarchical document categorization with support vector machines*, Proceedings of the Thirteenth ACM conference on Information and knowledge management (New York, NY, USA), ACM Press, 2004, pp. 78–87.

[Cre93] N. A. C. Cressie, *Statistics for spatial data*, John Wiley and Sons, New York, 1993.

[CS03] K. Crammer and Y. Singer, *Ultraconservative online algorithms for multiclass problems*, Journal of Machine Learning Research **3** (2003), 951–991.

[CSS00] M. Collins, R. E. Schapire, and Y. Singer, *Logistic regression, AdaBoost and Bregman distances*, Proc. 13th Annu. Conference on Comput. Learning Theory, Morgan Kaufmann, San Francisco, 2000, pp. 158–169.

[CV95] C. Cortes and V. Vapnik, *Support vector networks*, Machine Learning **20** (1995), no. 3, 273–297.

[JK02] K. Jarvelin and J. Kekalainen, *IR evaluation methods for retrieving highly relevant documents*, ACM Special Interest Group in Information Retrieval (SIGIR), New York: ACM, 2002, pp. 41–48.

[Joa05] Thorsten Joachims, *A support vector method for multivariate performance measures*, Proc. Intl. Conf. Machine Learning (San Francisco, California), Mor-

gan Kaufmann Publishers, 2005, pp. 377–384.
[Joa06] ______, *Training linear SVMs in linear time*, Proc. ACM Conf. Knowledge Discovery and Data Mining (KDD), ACM, 2006.
[Jor08] M. I. Jordan, *An introduction to probabilistic graphical models*, MIT Press, 2008, To Appear.
[JV87] R. Jonker and A. Volgenant, *A shortest augmenting path algorithm for dense and sparse linear assignment problems*, Computing **38** (1987), 325–340.
[Kar80] R.M. Karp, *An algorithm to solve the* $m \times n$ *assignment problem in expected time* $O(mn \log n)$, Networks **10** (1980), no. 2, 143–152.
[KD05] S. S. Keerthi and D. DeCoste, *A modified finite Newton method for fast solution of large scale linear SVMs*, J. Mach. Learn. Res. **6** (2005), 341–361.
[Koe05] R. Koenker, *Quantile regression*, Cambridge University Press, 2005.
[Kuh55] H.W. Kuhn, *The Hungarian method for the assignment problem*, Naval Research Logistics Quarterly **2** (1955), 83–97.
[LMP01] J. D. Lafferty, A. McCallum, and F. Pereira, *Conditional random fields: Probabilistic modeling for segmenting and labeling sequence data*, Proceedings of International Conference on Machine Learning (San Francisco, CA), vol. 18, Morgan Kaufmann, 2001, pp. 282–289.
[LS07] Quoc V. Le and Alexander J. Smola, *Direct optimization of ranking measures*, Tech. Report 0704.3359, arXiv, April 2007, http://arxiv.org/abs/0704.3359.
[McA07] David McAllester, *Generalization bounds and consistency for structured labeling*, Predicting Structured Data (Cambridge, Massachusetts), MIT Press, 2007.
[MSR+97] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, *Predicting time series with support vector machines*, Artificial Neural Networks ICANN'97 (Berlin) (W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, eds.), Lecture Notes in Comput. Sci., vol. 1327, Springer-Verlag, 1997, pp. 999–1004.
[Mun57] J. Munkres, *Algorithms for the assignment and transportation problems*, Journal of SIAM **5** (1957), no. 1, 32–38.
[Ned02] Angelia Nedic, *Subgradient methods for convex minimization*, Ph.D. thesis, MIT, 2002.
[OL93] J.B. Orlin and Y. Lee, *Quickmatch: A very fast algorithm for the assignment problem*, Working Paper 3547-93, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, March 1993.
[RBZ06] N. Ratliff, J. Bagnell, and M. Zinkevich, *Maximum margin planning*, International Conference on Machine Learning, July 2006.
[RSS+07] G. Rätsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.-R. Müller, R. J. Sommer, and B. Schölkopf, *Improving the Caenorhabditis elegans genome annotation using machine learning*, PLoS Computational Biology **3** (2007), no. 2, e20 doi:10.1371/journal.pcbi.0030020.
[SPST+01] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, *Estimating the support of a high-dimensional distribution*, Neural Computation **13** (2001), no. 7, 1443–1471.
[SSS07] S. Shalev-Shwartz and Y. Singer, *Logarithmic regret algorithms for strongly convex repeated games*, Tech. report, School of Computer Science, Hebrew University, 2007.
[TGK04] B. Taskar, C. Guestrin, and D. Koller, *Max-margin Markov networks*, Advances in Neural Information Processing Systems 16 (Cambridge, MA)

(S. Thrun, L. Saul, and B. Schölkopf, eds.), MIT Press, 2004, pp. 25–32.

[TJHA05] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, *Large margin methods for structured and interdependent output variables*, J. Mach. Learn. Res. **6** (2005), 1453–1484.

[VGS97] V. Vapnik, S. Golowich, and A. J. Smola, *Support vector method for function approximation, regression estimation, and signal processing*, Advances in Neural Information Processing Systems 9 (Cambridge, MA) (M. C. Mozer, M. I. Jordan, and T. Petsche, eds.), MIT Press, 1997, pp. 281–287.

[Voo01] E. Voorhees, *Overview of the TRECT 2001 question answering track*, TREC, 2001.

[Wah97] G. Wahba, *Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV*, Tech. Report 984, Department of Statistics, University of Wisconsin, Madison, 1997.

[Wil98] C. K. I. Williams, *Prediction with Gaussian processes: From linear regression to linear prediction and beyond*, Learning and Inference in Graphical Models (M. I. Jordan, ed.), Kluwer Academic, 1998, pp. 599–621.

[WJ03] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*, Tech. Report 649, UC Berkeley, Department of Statistics, September 2003.

[Zin03] M. Zinkevich, *Online convex programming and generalised infinitesimal gradient ascent*, Proceedings of the International Conference on Machine Learning, 2003, pp. 928–936.