

---

# Support Vector Machines as Probabilistic Models

---

**Vojtěch Franc**

Czech Technical University in Prague, Technická 2, 166 27 Praha 6, Czech Republic

XFRANCV@CMP.FELK.CVUT.CZ

**Alex Zien**

LIFE Biosystems GmbH, Belfortstr. 2, 69115 Heidelberg, Germany

ALEXANDER.ZIEN@LIFEBIOSYSTEMS.COM

**Bernhard Schölkopf**

Max Planck Institute for Intelligent Systems, Spemannstr. 38, 72076 Tübingen, Germany

BS@TUEBINGEN.MPG.DE

## Abstract

We show how the SVM can be viewed as a maximum likelihood estimate of a class of probabilistic models. This model class can be viewed as a reparametrization of the SVM in a similar vein to the  $\nu$ -SVM reparametrizing the classical ( $C$ -)SVM. It is not discriminative, but has a non-uniform marginal. We illustrate the benefits of this new view by re-deriving and re-investigating two established SVM-related algorithms.

## 1. Introduction

The SVM is one of the most used and best studied machine learning models. As such, many aspects of it have been assayed, including its links to learning theory and regularization, its geometry, the influence of kernels on its regularization, its consistency, and its efficient optimization (Vapnik, 1998; Schölkopf & Smola, 2002; Steinwart & Christmann, 2008). This understanding has laid the ground for significant developments including recent methods for efficient and effective structured output learning.

However, one major gap in our understanding of SVMs still persists: the attempts to place it in a probabilistic framework have remained unsatisfactory. For instance, (Sollich, 2002) interprets the hinge loss as  $-\log p(y|x)$ , which necessitates to artificially introduce a “don’t-know” class. Counter-intuitively its probability, a notion of predictive uncertainty, is minimal at the border of the margin (i.e., for  $f(\mathbf{x}) = \pm 1$ ) and increases when moving further away from the decision surface.

In (Grandvalet et al., 2005) the SVM objective is taken as an approximation to the negative log-likelihood such that the SVM outputs are translated into probability intervals. In a practical but also heuristic approach, (Platt, 2000) suggested to retrospectively fit a logit function to map (non-probabilistic) SVM outputs to probabilities. This works well and has become the standard, but fails to provide insight.

In fact, there are theoretical arguments indicating that the hinge loss used by the SVM does not lend itself well to the estimation of posterior class probabilities: as the number of datapoints goes to infinity, under certain conditions on the kernel and on the rate at which the regularization strength tends to zero, the real-valued discriminant function  $f$  returned by the SVM essentially converges to the optimal  $\pm 1$ -valued classifier,  $\text{sign}(\Pr(y = 1|x) - \frac{1}{2})$  (Steinwart & Christmann, 2008). This indicates that the SVM tries to approximate a function that does not retain information beyond the (optimal) class membership.<sup>1</sup> However, this does not mean that the estimation of conditional probabilities is necessarily impossible in the finite sample setting. In that case,  $f$  will typically have a much larger range (e.g., a non-empty interval of  $\mathbb{R}$ ). Moreover, for the widely used linear kernel, the RKHS is not rich enough to approximate the Bayes classifier, so the above result does not apply in the first place.

Why has the SVM evaded being cast as an ML or MAP estimate of a probabilistic model—especially given that the rather similar (penalized) kernel logistic regression (LR) is clear and simple? Two mental barriers had to be overcome. First, typically the SVM (analogous to LR) is taken to be a discriminative model, i.e., one that only specifies the conditional

---

Appearing in *Proceedings of the 28<sup>th</sup> International Conference on Machine Learning*, Bellevue, WA, USA, 2011. Copyright 2011 by the author(s)/owner(s).

---

<sup>1</sup>In a sense, this is the flipside of the SVM’s sparsity, as argued by (Bartlett & Tewari, 2007).

$p(y|\mathbf{x})$ . In contrast, we argue that the SVM implies a (non-uniform) marginal  $p(\mathbf{x})$ , giving it a generative touch. Second, it is alluring to expect a model whose ML/MAP solution exactly agrees with the standard SVM. While we do not know whether this is possible at all, we did succeed in recovering an alternative parameterization of the SVM as an ML estimate of a suitable model. This is analogous to the  $\nu$ -SVM reformulation of the standard SVM (Schölkopf et al., 2000), in which the rather unintuitive regularization parameter is replaced by a parameter controlling the number of SVs. In our model, the hyperparameter will be the length  $\|\mathbf{w}\|$  of the hyperplane normal.

After a brief review of the SVM (Section 2), we present our model (Section 3.1), which is generative and semi-parametric. For technical reasons we restrict our analysis to the SVM classifier without the bias term.<sup>2</sup> The core result, the equivalence of ML in our model with the SVM, is presented and proved in Section 4. In Section 5 we demonstrate how max-margin clustering drops out of our model; after this, we conclude.

## 2. Support Vector Machine classification

We are given a set of training examples  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in (\mathbb{R}^n \times \{+1, -1\})^m$  assumed to be i.i.d. from an unknown probability distribution function (p.d.f.)  $p^*(\mathbf{x}, y)$ . The goal is to learn a Bayes classifier  $q: \mathbb{R}^n \rightarrow \{+1, -1\}$  which minimizes the expected classification error  $\int_{\mathbf{x} \in \mathbb{R}^n} \sum_{y \in \{+1, -1\}} \mathbb{I}[y \neq q(\mathbf{x})] p^*(\mathbf{x}, y) d\mathbf{x}$  where  $\mathbb{I}[X] = 1$  if  $X$  is satisfied and 0 otherwise.

The SVM model without bias assumes that the Bayes classifier can be well approximated by a linear classifier  $q_{\text{SVM}}: \mathbb{R}^n \rightarrow \{+1, -1\}$  parametrized by a vector  $\mathbf{w} \in \mathbb{R}^n$  such that

$$q_{\text{SVM}}(\mathbf{x}; \mathbf{w}) = \begin{cases} +1 & \text{if } \langle \mathbf{x}, \mathbf{w} \rangle \geq 0, \\ -1 & \text{if } \langle \mathbf{x}, \mathbf{w} \rangle < 0. \end{cases} \quad (1)$$

The parameter vector  $\mathbf{w}$  is evaluated by a cost function

$$F(\mathbf{w}; \lambda, \omega) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + R(\mathbf{w}; \omega)$$

where  $R(\mathbf{w}; \omega) = \sum_{i=1}^m \omega^{y_i} \ell(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)$  is a convex approximation of the training (empirical) classification error,  $\ell(t) = \max\{0, 1 - t\}$  is the hinge-loss,  $\lambda \in (0, \infty) =: \mathbb{R}^{++}$  is a strictly positive regularization constant, and  $\omega \in (0, 1)$  is a scalar defining cost-factors

<sup>2</sup>Which can, by augmenting the feature space, be used to arbitrarily well approximate the solution of an SVM with bias.

$\omega^+ = \omega$  and  $\omega^- = 1 - \omega$  for positive and negative class, respectively. For any fixed  $\lambda > 0$ ,  $\omega \geq 0$ , the function  $R(\mathbf{w}; \omega)$  is convex and the function  $F(\mathbf{w}; \lambda, \omega)$  is strictly convex.

For given  $\lambda$  and  $\omega$ , the SVM learning algorithm returns the parameter vector  $\mathbf{w}_{\text{SVM}}(\lambda, \omega)$  which is a unique minimum of  $F(\mathbf{w}; \lambda, \omega)$ , i.e.,

$$\mathbf{w}_{\text{SVM}}(\lambda, \omega) = \underset{\mathbf{w} \in \mathbb{R}^n}{\operatorname{argmin}} F(\mathbf{w}; \lambda, \omega). \quad (2)$$

The problem (2) is well understood and there exists a plethora of efficient optimization algorithms for its solution.

The SVM algorithm specifies how to learn the parameter vector  $\mathbf{w}$  while the hyper-parameters  $\lambda$  and  $\omega$  must be determined differently. The standard SVMs sets  $\omega = \frac{1}{2}$ . However, tuning of  $\omega$  is a routinely used heuristic in the case of unbalanced class distribution. A common practice is to selected the best combination of  $\lambda$  and  $\omega$  based on solving

$$(\lambda_{\text{best}}, \omega_{\text{best}}) = \underset{\lambda \in \Lambda, \omega \in \Omega}{\operatorname{argmin}} G[q_{\text{SVM}}(\cdot; \mathbf{w}_{\text{SVM}}(\lambda, \omega))] \quad (3)$$

where the sets  $\Lambda = \{\lambda_1, \dots, \lambda_p\}$  and  $\Omega = \{\omega_1, \dots, \omega_p\}$  are prescribed manually based on user's experience. The functional  $G[q_{\text{SVM}}(\cdot; \mathbf{w})]$  is an estimator of the expected classification error of the rule  $q_{\text{SVM}}(\cdot; \mathbf{w})$ . The classification error computed on an independent set of examples, the cross-validation or the leave-one-out are among the most typically used error estimators. The resulting classifier is then  $q_{\text{SVM}}(\mathbf{x}; \mathbf{w}_{\text{SVM}}(\lambda_{\text{best}}, \omega_{\text{best}}))$ .

## 3. Semi-parametric probabilistic model

### 3.1. The model

We consider the following semi-parametric p.d.f.

$$p(\mathbf{x}, y; \tau, \omega, \mathbf{u}) = Z(\tau, \omega) \cdot \exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle)) \cdot h(\mathbf{x}) \quad (4)$$

defined over  $\mathbb{R}^n \times \{+1, -1\}$ . The distribution (4) is parametrized by a unit vector  $\mathbf{u} \in \mathcal{U} = \{\mathbf{u}' \in \mathbb{R}^n \mid \|\mathbf{u}'\| = 1\}$ , a strictly positive scalar  $\tau \in \mathbb{R}^{++}$  and a scalar  $\omega \in (0, 1)$  defining  $\omega^+ = \omega$  and  $\omega^- = 1 - \omega$ . The distribution (4) is composed of three terms. The first term,  $Z(\tau, \omega)$ , is a normalization constant invariant to  $\mathbf{u}$ . The second term,  $\exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle))$ , is a function of all three parameters  $(\tau, \omega, \mathbf{u})$  and the input variables  $(\mathbf{x}, y)$ . Finally, the third term,  $h(\mathbf{x})$ , is a function which ensures that  $p(\mathbf{x}, y; \tau, \omega, \mathbf{u})$  is integrable and that the normalization constant  $Z(\tau, \omega)$  does not depend on  $\mathbf{u}$ . The properties of  $h$  are defined in Theorem 1.

**Theorem 1** Let  $h: \mathbb{R}^n \rightarrow \mathbb{R}$  be a piece-wise continuous function which satisfy the following assumptions:

1.  $0 \leq h(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^n$  (positive)
2.  $0 < \int_{\mathbf{x} \in \mathbb{R}^n} h(\mathbf{x}) < \infty$  (integrable)
3.  $h(\mathbf{x}) = h(\mathbf{x}'), \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$  such that  $\|\mathbf{x}\| = \|\mathbf{x}'\|$  (radial basis function)

Then for any  $\tau \in \mathbb{R}^{++}$ ,  $\omega \in (0, 1)$  and  $\mathbf{u} \in \mathcal{U}$  the integrals  $I^+(\tau, \omega, \mathbf{u})$  and  $I^-(\tau, \omega, \mathbf{u})$  defined by

$$I^y(\tau, \omega, \mathbf{u}) = \int_{\mathbf{x} \in \mathbb{R}^n} \exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle)) \cdot h(\mathbf{x}) \cdot d\mathbf{x} \quad (5)$$

satisfy the following properties

1.  $0 < I^y(\tau, \omega, \mathbf{u})$  (strictly positive)
2.  $I^y(\tau, \omega, \mathbf{u}) < \infty$  (finite)
3.  $I^y(\tau, \omega, \mathbf{u}) = I^y(\tau, \omega)$  (invariant to  $\mathbf{u}$ )

PROOF: For fixed  $\tau, \omega, \mathbf{u}$ , we introduce a short-hand  $g(\mathbf{x}) = \exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle))$  which simplifies (5) to  $I^y(\tau, \omega, \mathbf{u}) = \int_{\mathbf{x} \in \mathbb{R}^n} h(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}$ . It is seen that  $g: \mathbb{R}^n \rightarrow (0, 1]$ .

Because  $h$  is a piece-wise continuous and its integral is strictly positive then there must exist  $\boldsymbol{\mu} \in \mathbb{R}^n, r > 0$ , and  $\varepsilon_1 > 0$  such that for all  $\mathbf{x}$  within the ball  $\mathcal{B}(\boldsymbol{\mu}, r) = \{\mathbf{x}' \in \mathbb{R}^n \mid \|\mathbf{x}' - \boldsymbol{\mu}\| \leq r\}$  the value of  $h(\mathbf{x})$  is not less than  $\varepsilon_1$ . The volume  $V$  of  $\mathcal{B}(\boldsymbol{\mu}, r)$  is greater than 0. As  $g$  is strictly positive everywhere there must exist  $\varepsilon_2 > 0$  such that  $g(\mathbf{x}) \geq \varepsilon_2, \forall \mathbf{x} \in \mathcal{B}(\boldsymbol{\mu}, r)$ . This implies that  $I^y(\tau, \omega, \mathbf{u}) \geq \varepsilon_1 \varepsilon_2 V > 0$  which proves the property 1.

The property 2 follows from integrability of  $h$  and boundedness of  $g$ .

Finally, we prove the property 3. Let  $\mathbf{u}$  and  $\mathbf{u}'$  be arbitrary unit vectors. Then, there exists a orthogonal matrix  $\mathbf{R} \in \mathbb{R}^{n \times n}$  with determinant +1 (i.e, rotation matrix) such that  $\mathbf{u}' = \mathbf{R}\mathbf{u}$ . Let  $\varphi: \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a vector-valued function defined by  $\varphi(\mathbf{v}) = \mathbf{R}\mathbf{v}$ . It is seen that the determinant of the Jacobian matrix of  $\varphi$  is  $D_\varphi(\mathbf{v}) = +1$ . We can write

$$\begin{aligned} I^y(\tau, \omega, \mathbf{u}) &\stackrel{(1)}{=} \int_{\mathbf{x} \in \mathbb{R}^n} h(\mathbf{x}) \exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle)) d\mathbf{x} \\ &\stackrel{(2)}{=} \int_{\mathbf{v} \in \mathbb{R}^n} h(\varphi(\mathbf{v})) \exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \varphi(\mathbf{v}) \rangle)) |D_\varphi(\mathbf{v})| d\mathbf{v} \\ &\stackrel{(3)}{=} \int_{\mathbf{v} \in \mathbb{R}^n} h(\mathbf{v}) \exp(-\omega^y \ell(y \langle \tau \mathbf{u}', \mathbf{v} \rangle)) d\mathbf{v} \stackrel{(4)}{=} I^y(\tau, \omega, \mathbf{u}') \end{aligned} \quad (6)$$

The second equality follows from the substitution theorem for multivariate integrals. The third equality

uses  $h(\mathbf{x}) = h(\|\mathbf{x}\|)$  ensured by the assumption 3, the fact that  $|D_\varphi(\mathbf{v})| = 1$ , and the equality  $\langle \tau \mathbf{u}, \varphi(\mathbf{v}) \rangle = \langle \tau \mathbf{R}\mathbf{u}, \mathbf{v} \rangle = \langle \tau \mathbf{u}', \mathbf{v} \rangle$ . The equalities 1 and 4 are due to definition (5) which completes the proof. ■

Two examples of functions satisfying assumptions of Theorem 1 are  $h_1(\mathbf{x}) = \exp(-\langle \mathbf{x}, c_1 \mathbf{E}\mathbf{x} \rangle)$  and  $h_2(\mathbf{x}) = c_2 \mathbb{I}[\|\mathbf{x}\| \leq c_3]$  where  $c_1, c_2$  and  $c_3$  are arbitrary strictly positive scalars and  $\mathbf{E}$  is the identity matrix.

**Corrolary 1** Let  $h$  be a function satisfying the assumptions of Theorem 1 and let us define

$$Z(\tau, \omega) = \frac{1}{I^+(\tau, \omega) + I^-(\tau, \omega)}. \quad (7)$$

Then for any fixed  $\tau \in \mathbb{R}^{++}, \omega \in (0, 1)$  and  $\mathbf{u} \in \mathcal{U}$ , the function  $p(\mathbf{x}, y; \tau, \omega, \mathbf{u})$  given by (4) is a proper p.d.f. defined over  $\mathbb{R}^n \times \{+1, -1\}$ , that is,

$$\begin{aligned} p(\mathbf{x}, y; \tau, \omega, \mathbf{u}) &\geq 0, \quad \forall \mathbf{x} \in \mathbb{R}^n, y \in \{+1, -1\}, \\ \int_{\mathbf{x} \in \mathbb{R}^n} \sum_{y \in \{+1, -1\}} p(\mathbf{x}, y; \tau, \omega, \mathbf{u}) &= 1. \end{aligned}$$

### 3.2. Prior probability

It follows from (7) that the priory probability of the class label  $y$  under the model (4) is given by

$$p(y; \tau, \omega) = \frac{I^y(\tau, \omega)}{I^+(\tau, \omega) + I^-(\tau, \omega)}. \quad (8)$$

The prior probability does not depend on the parameter  $\mathbf{u}$ . Moreover, we have the following theorem.

**Theorem 2** For any  $\tau \in \mathbb{R}^{++}$  it holds that

$$\begin{aligned} p(y = +1; \tau, \omega) &= 0.5 \quad \text{for } \omega = 0.5, \\ p(y = +1; \tau, \omega) &< 0.5 \quad \text{for } \omega > 0.5, \\ p(y = +1; \tau, \omega) &> 0.5 \quad \text{for } \omega < 0.5. \end{aligned}$$

PROOF: It follows from the equality (6) that

$$I^y(\tau, \omega) = \int_{\mathbf{x} \in \mathbb{R}^n} f^y(\mathbf{x}) \cdot h(\mathbf{x}) \cdot d\mathbf{x}, \quad (9)$$

where  $f^y(\mathbf{x}) = \exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle))$  is a function which is for  $\omega = 0.5$  (i.e.,  $\omega^+ = \omega^- = 0.5$ ) invariant to  $y \in \{+1, -1\}$ . In turn,  $\omega = 0.5$  implies that  $I^+(\tau, 0.5) = I^-(\tau, 0.5)$  which after substituting to (8) yields  $p(y = 1; \tau, 0.5) = 0.5$ . For  $\omega > 0.5$  we have that  $f^+(\mathbf{x}) < f^-(\mathbf{x})$  on the whole subspace  $\{\mathbf{x} \in \mathbb{R}^n \mid \langle \tau \mathbf{u}, \mathbf{x} \rangle < 1\}$  which implies that  $I^+(\tau, \omega) < I^-(\tau, \omega)$  and thus also  $p(y = +1; \tau, \omega) < 0.5$ . The same reasoning can be used to prove that  $\omega < 0.5$  implies  $p(y = +1; \tau, \omega) > 0.5$ . ■

Theorem 2 establishes correspondence between the hyper-parameter  $\omega$  and the prior probability  $p(y; \tau, \omega)$ . Provided the prior is uniform we know the value of hyper-parameter  $\omega$  exactly, namely,  $\omega = 0.5$ . In the case of an unbalanced priors, we only know whether  $\omega$  is greater or less than 0.5. Note that to make the correspondence exact we would need to compute the integral (9), which involves the function  $h$ .

### 3.3. Posterior and marginal probabilities and relation to LR

The class posterior probability derived from the model (4) reads

$$p(y | \mathbf{x}; \tau, \omega, \mathbf{u}) = \frac{\exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle))}{\sum_{y \in \{+1, -1\}} \exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle))}. \quad (10)$$

It is seen that the posterior probability does not depend on the function  $h$ . The marginal p.d.f. reads

$$p(\mathbf{x}; \tau, \omega, \mathbf{u}) = Z(\tau, \omega) \cdot h(\mathbf{x}) \cdot f(\mathbf{x}; \tau, \omega, \mathbf{u})$$

where  $f(\mathbf{x}; \tau, \omega, \mathbf{u}) = \sum_{y \in \{+1, -1\}} \exp(-\omega^y \ell(y \langle \tau \mathbf{u}, \mathbf{x} \rangle))$  denotes its parametric part.

Figure 1 shows the posterior and the parametric part of the marginal p.d.f. in 1-d (i.e.  $\mathbf{x} \in \mathbb{R}^1$ ) for different values of the hyper-parameter  $\tau$  and the other parameters set to  $\omega = 0.5$ ,  $\mathbf{u} = 1$ . For comparison, we also plot the posterior probability of the LR model

$$p_{\text{LR}}(y | \mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-y \langle \mathbf{w}, \mathbf{x} \rangle)} \quad \text{with} \quad \mathbf{w} = \tau \mathbf{u}.$$

It is seen that the posterior probability of our model is very close to that of LR model. In fact, both are exactly the same in the margin band.

The crucial difference between the LR model and our SVM model is the marginal p.d.f. The LR imposes no assumption about the shape of the marginal at all. By contrast, our model defines its shape up to the non-parametric part  $h$ . It is obvious that the marginal of our model is non-uniform. Its shape is consistent with its well-known property to be margin maximizing. The width of the margin band is inversely proportional to the hyper-parameter  $\tau$ .

### 3.4. Bayes classifier

The (optimal) Bayes classifier minimizing the expected classification error is based on the log-likelihood ratio  $q(\mathbf{x}) = \frac{p(y=+1|\mathbf{x})}{p(y=-1|\mathbf{x})}$ ; the input  $\mathbf{x}$  is assigned the label  $y = +1$  if  $q(\mathbf{x}) \geq 0$  and the label  $y = -1$  otherwise.

Using (10) we can show after a little algebra that the

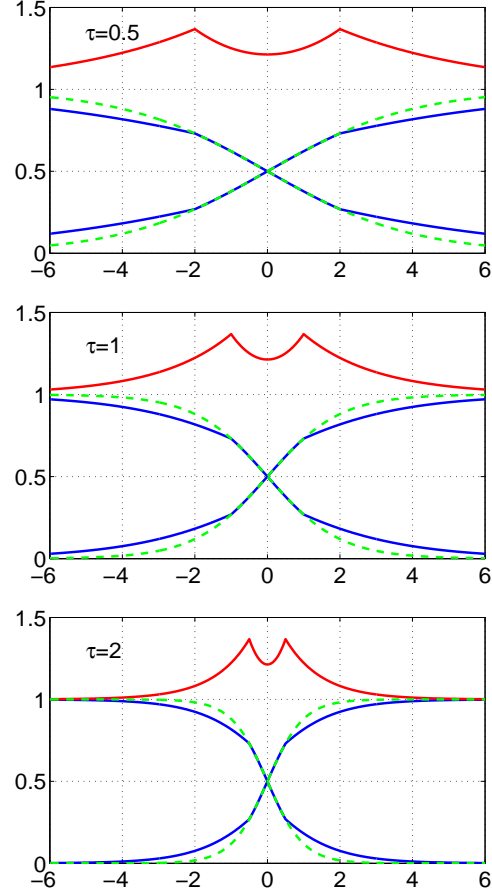


Figure 1. The figures show the posterior probability  $p(y | \mathbf{x}; \tau, \omega, \mathbf{u})$  (blue) and the parametric part  $f(\mathbf{x}; \tau, \omega, \mathbf{u})$  of the marginal p.d.f. (red) for three different values of the hyper-parameter  $\tau$  which is reciprocal to the margin width. The input variable corresponding to the x-axis is univariate,  $\mathbf{x} \in \mathbb{R}^1$ , and the parameters are set to  $\omega = 0.5$  and  $\mathbf{u} = 1$ . For comparison, the figures also show the posterior probability  $p_{\text{LR}}(y | \mathbf{x}; \mathbf{w})$  (dashed green) of the Logistic Regression model with  $\mathbf{w} = \tau \mathbf{u}$ .

log-likelihood ratio is a piece-wise linear function

$$\begin{aligned} q(\mathbf{x}; \tau, \omega, \mathbf{u}) &= \log \frac{p(y = +1 | \mathbf{x}; \tau, \omega, \mathbf{u})}{p(y = -1 | \mathbf{x}; \tau, \omega, \mathbf{u})} \\ &= \begin{cases} (\langle \tau \mathbf{u}, \mathbf{x} \rangle - 1)\omega^+ & \text{if } \langle \tau \mathbf{u}, \mathbf{x} \rangle \in (-\infty, -1], \\ \langle \tau \mathbf{u}, \mathbf{x} \rangle + 1 - 2\omega & \text{if } \langle \tau \mathbf{u}, \mathbf{x} \rangle \in [-1, 1], \\ (\langle \tau \mathbf{u}, \mathbf{x} \rangle + 1)\omega^- & \text{if } \langle \tau \mathbf{u}, \mathbf{x} \rangle \in [1, \infty). \end{cases} \end{aligned} \quad (11)$$

Using the log-likelihood ratio (11), we can derive that the Bayes classifier is the linear classification rule

$$q_{\text{Bayes}}(\mathbf{x}; \tau, \omega, \mathbf{u}) = \begin{cases} +1 & \text{if } \langle \tau \mathbf{u}, \mathbf{x} \rangle \geq b, \\ -1 & \text{if } \langle \tau \mathbf{u}, \mathbf{x} \rangle < b, \end{cases} \quad (12)$$

where  $b = 2\omega - 1$ . Note that the classifier (12) becomes for  $\omega = 0.5$  unbiased just like the SVM classifier (1).

### 3.5. The Maximum-Likelihood estimator

Given training examples  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in (\mathbb{R}^n \times \{+1, -1\})^m$  assumed to be i.i.d. from the distribution (4), the negative log-likelihood (NLL) of the parameters  $(\tau, \omega, \mathbf{u})$  reads

$$\begin{aligned} L(\tau, \omega, \mathbf{u}) &= - \sum_{i=1}^m \log p(\mathbf{x}_i, y_i; \tau, \omega, \mathbf{u}) \\ &= \sum_{i=1}^m \omega^{y_i} \ell(y_i \langle \tau \mathbf{u}, \mathbf{x}_i \rangle) - m \log Z(\tau, \omega) - \sum_{i=1}^m \log h(\mathbf{x}_i). \end{aligned} \quad (13)$$

The key observation is as follows. To compute the ML estimate of all the three parameters  $(\tau, \omega, \mathbf{u})$  we need to know the function  $h$ . However, under the assumption that  $\tau$  and  $\omega$  are given, the ML estimator of the parameter vector  $\mathbf{u}$  does not depend on  $h$ . Let  $\mathbf{u}_{\text{ML}}(\tau, \omega)$  denote the ML estimator of  $\mathbf{u}$  provided the hyper-parameters  $\tau$  and  $\omega$  are known, then we have

$$\mathbf{u}_{\text{ML}}(\tau, \omega) \in \underset{\mathbf{u} \in \mathcal{U}}{\operatorname{argmin}} L(\tau, \omega, \mathbf{u}) = \underset{\mathbf{u} \in \mathcal{U}}{\operatorname{argmin}} R(\tau \mathbf{u}; \omega), \quad (14)$$

where  $R(\tau \mathbf{u}; \omega)$  is the SVM risk term (see Section 2).

Having the ML estimator, one can implement the ML learning of the plug-in Bayes classifier (12), i.e., we plug-in the ML estimate of the parameters for the real ones. Following a common practice, the hyper-parameters  $(\tau, \omega)$  can be found by solving

$$(\tau_{\text{best}}, \omega_{\text{best}}) = \underset{\tau \in \mathcal{T}, \omega \in \Omega}{\operatorname{argmin}} G[q_{\text{Bayes}}(\cdot; \tau, \omega, \mathbf{u}_{\text{ML}}(\tau, \omega))] \quad (15)$$

where the sets  $\mathcal{T} = \{\tau_1, \dots, \tau_p\}$  and  $\Omega = \{\omega_1, \dots, \omega_p\}$  are prescribed manually based on user's experience. The functional  $G[q_{\text{Bayes}}(\cdot; \tau, \omega, \mathbf{u}_{\text{ML}}(\tau, \omega))]$  is an estimator of the expected classification error of the rule  $q_{\text{Bayes}}(\cdot; \tau, \omega, \mathbf{u})$ . The resulting classifier is then  $q_{\text{Bayes}}(\mathbf{x}; \tau_{\text{best}}, \omega_{\text{best}}, \mathbf{u}_{\text{ML}}(\tau_{\text{best}}, \omega_{\text{best}}))$ .

## 4. Equivalence between SVM and ML learning

We first prove a theorem which establishes equivalence between the optimization problems appearing in the SVM and ML learning.

**Theorem 3** *Let us consider the following optimization problems*

$$\mathbf{w}(\lambda) = \underset{\mathbf{w} \in \mathbb{R}^n}{\operatorname{argmin}} \left[ \frac{\lambda}{2} \|\mathbf{w}\|^2 + R(\mathbf{w}) \right] \quad (16)$$

$$\mathbf{u}(\tau) = \underset{\mathbf{u} \in \mathcal{U}}{\operatorname{argmin}} R(\tau \mathbf{u}) \quad (17)$$

where  $R: \mathbb{R}^n \rightarrow \mathbb{R}$  is a convex function and assume that the minimum of (16) exists. Then, there exists

a monotonically decreasing mapping  $\theta: \mathbb{R}^{++} \rightarrow \mathbb{R}^+$  which for any  $\lambda \in \mathbb{R}^{++}$  returns  $\tau = \theta(\lambda) = \|\mathbf{w}(\lambda)\|$  such that the following equality holds:

$$\tau \mathbf{u}(\tau) = \mathbf{w}(\lambda). \quad (18)$$

At first sight it may be puzzling that an ML estimate (hence lacking an explicit prior) can be equivalent to the SVM objective, for its regularizer is commonly interpreted as a log prior (in analogy to penalized LR). The resolving insight is simple: the regularizer (prior) acts only on  $\|\mathbf{w}\|$ , which is kept fixed in our model, serving as a hyperparameter substitute for SVM- $\lambda$ .

Before proving Theorem 3 we introduce the following auxiliary lemma.

**Lemma 1** *For any  $\lambda_1 \in \mathbb{R}^{++}$  and  $\lambda_2 \in \mathbb{R}^{++}$  such that  $\lambda_1 > \lambda_2$  the inequality  $\|\mathbf{w}_1\| \leq \|\mathbf{w}_2\|$  holds where  $\mathbf{w}_1 \in \mathbf{w}(\lambda_1)$  and  $\mathbf{w}_2 \in \mathbf{w}(\lambda_2)$  are solutions of the problem (16) for  $\lambda_1$  and  $\lambda_2$ , respectively.*

PROOF: Since  $\mathbf{w}_1$  and  $\mathbf{w}_2$  are minimizers of (16), we have that the following inequalities

$$\frac{\lambda_1}{2} \|\mathbf{w}_1\|^2 + R(\mathbf{w}_1) \leq \frac{\lambda_1}{2} \|\mathbf{w}\|^2 + R(\mathbf{w}), \quad (19)$$

$$\frac{\lambda_2}{2} \|\mathbf{w}_2\|^2 + R(\mathbf{w}_2) \leq \frac{\lambda_2}{2} \|\mathbf{w}\|^2 + R(\mathbf{w}). \quad (20)$$

hold  $\forall \mathbf{w} \in \mathbb{R}^n$ . Substituting  $\mathbf{w} = \mathbf{w}_2$  to (19) and  $\mathbf{w} = \mathbf{w}_1$  to (20) yields

$$\frac{\lambda_1}{2} \|\mathbf{w}_1\|^2 + R(\mathbf{w}_1) \leq \frac{\lambda_1}{2} \|\mathbf{w}_2\|^2 + R(\mathbf{w}_2), \quad (21)$$

$$\frac{\lambda_2}{2} \|\mathbf{w}_2\|^2 + R(\mathbf{w}_2) \leq \frac{\lambda_2}{2} \|\mathbf{w}_1\|^2 + R(\mathbf{w}_1). \quad (22)$$

By summing up the inequalities (21) and (22) we get

$$\begin{aligned} \frac{\lambda_1}{2} \|\mathbf{w}_1\|^2 + \frac{\lambda_2}{2} \|\mathbf{w}_2\|^2 + R(\mathbf{w}_1) + R(\mathbf{w}_2) &\leq \\ \frac{\lambda_1}{2} \|\mathbf{w}_2\|^2 + \frac{\lambda_2}{2} \|\mathbf{w}_1\|^2 + R(\mathbf{w}_1) + R(\mathbf{w}_2) & \end{aligned}$$

which after a little algebra yields

$$(\lambda_1 - \lambda_2)(\|\mathbf{w}_2\|^2 - \|\mathbf{w}_1\|^2) \geq 0.$$

and so  $\lambda_1 > \lambda_2$  implies  $\|\mathbf{w}_1\| \leq \|\mathbf{w}_2\|$ .  $\blacksquare$

Now, we prove Theorem 3.

PROOF: Let  $\mathbf{w}(\lambda)$  be the minimizer of (16) for some  $\lambda \in \mathbb{R}^{++}$ . Note that  $\mathbf{w}(\lambda)$  is unique as the objective of (16) is strictly convex. Let us denote  $\tau = \|\mathbf{w}(\lambda)\|$  and  $\mathcal{W}_\tau = \{\mathbf{w} \in \mathbb{R}^n \mid \|\mathbf{w}\| = \tau\}$ . Then, we can write

$$\begin{aligned} \mathbf{w}(\lambda) &\stackrel{(1)}{=} \underset{\mathbf{w} \in \mathcal{W}_\tau}{\operatorname{argmin}} \left[ \frac{\lambda}{2} \|\mathbf{w}\|^2 + R(\mathbf{w}) \right] \stackrel{(2)}{=} \underset{\mathbf{w} \in \mathcal{W}_\tau}{\operatorname{argmin}} R(\mathbf{w}) \\ &\stackrel{(3)}{=} \tau \mathbf{u}(\tau) \quad \text{where} \quad \mathbf{u}(\tau) = \underset{\mathbf{u} \in \mathcal{U}}{\operatorname{argmin}} R(\tau \mathbf{u}). \end{aligned}$$



The first equality follows from the fact that  $\mathcal{W}_\tau$  is a subset of  $\mathbb{R}^n$  containing the minimizer  $\mathbf{w}(\lambda)$ . Since all vectors in  $\mathcal{W}_\tau$  have the same norm the second equality holds true. The third equality results from the variable substitution  $\mathbf{w} = \tau\mathbf{u}$ . This proves that for any  $\lambda \in \mathbb{R}^{++}$  and  $\tau = \theta(\lambda) = \|\mathbf{w}(\lambda)\|$  the equality (18) holds.

It remains to prove monotonicity of  $\theta$ , i.e.,  $\lambda_1 > \lambda_2$  implies  $\theta(\lambda_1) \leq \theta(\lambda_2)$ . However, this is a direct consequence of Lemma 1 and the fact that  $\theta(\lambda) = \|\mathbf{w}(\lambda)\|$ . ■

#### 4.1. Standard SVM classifier

The formulation of the standard SVMs assumes that the cost-factors for both classes are equal, i.e.,  $\omega = 0.5$ . Recall that by Theorem 2 the value  $\omega = 0.5$  implies that the posterior probability of our model is uniform, i.e.  $p(y; \tau, \omega, \mathbf{u}) = 0.5$ .

**Theorem 4** *Let  $\omega = 0.5$ . Then, for any linear SVM classifier there exists an equivalent plug-in Bayes classifier derived from the model (4) whose parameters are estimated by the ML principle, i.e., the equality*

$$q_{\text{SVM}}(\mathbf{x}; \mathbf{w}_{\text{SVM}}(\lambda, \omega)) = q_{\text{Bayes}}(\mathbf{x}; \tau, \omega, \mathbf{u}_{\text{ML}}(\tau, \omega))$$

holds for any  $\mathbf{x} \in \mathbb{R}^n$ ,  $\lambda \in \mathbb{R}^{++}$  and  $\tau = \theta(\lambda) = \|\mathbf{w}_{\text{SVM}}(\lambda, \omega)\|$ . Moreover, the mapping  $\theta: \mathbb{R}^{++} \rightarrow \mathbb{R}^+$  is monotonically decreasing.

PROOF: The proof follows trivially from Theorem 3 and the formulas for the linear SVM classifier (1) and the Bayes classifier (12). ■

Note that the standard SVM is theoretically linked to our model in much the same way the  $\nu$ -SVM is linked to the standard SVM. In our case, training an SVM using  $\lambda$  would tell us which  $\tau$  to use to get the same result; in the other case, training a  $\nu$ -SVM using an a priori chosen  $\nu$  would tell us which  $\lambda$  to use in the standard SVM to get the same result (see Proposition 6 in (Schölkopf et al., 2000)).

Now let us compare the SVM and the ML learning from a more practical point of view. We still assume the standard setting  $\omega = 0.5$ . The SVM learning requires a user to supply a tuning set  $\Lambda = \{\lambda_1, \dots, \lambda_p\}$  for the hyper-parameter  $\lambda$ . The resulting SVM classifier is obtained by the procedure (3) which selects the best parameter vector from  $\mathbf{w}_{\text{SVM}}(\lambda, 0.5)$ ,  $\lambda \in \Lambda$ , based on a validation criterion. The ML learning requires the user to supply a tuning set  $\mathcal{T} = \{\tau_1, \dots, \tau_p\}$  for the hyper-parameter  $\tau$ . The result-

ing plug-in Bayes classifier is obtained by the procedure (15) which selects the best parameter vector from  $\mathbf{u}_{\text{ML}}(\tau, 0.5)$ ,  $\tau \in \mathcal{T}$ , based on a validation criterion. Theorem (4) guarantees that both procedures will return exactly the same linear classifier provided we use  $\mathcal{T} = \{\|\mathbf{w}_{\text{SVM}}(\lambda_1, 0.5)\|, \dots, \|\mathbf{w}_{\text{SVM}}(\lambda_p, 0.5)\|\}$  and the same validation criterion in (3) and (15). Hence an empirical comparison is unnecessary.

Note that the tuning sets are heuristic in both cases. While this is outside the scope of this paper, we imagine that the direct geometric interpretation of  $\tau$  as the reciprocal of the margin width will facilitate heuristics for finding reasonable settings, which could provide a practical advantage.

#### 4.2. SVM classifier with different cost factors

It is common knowledge that SVMs often do not work well if the class distribution of the training examples is highly unbalanced. In view of the previous section this is not surprising, as the standard SVM classifier is equivalent to a Bayes classifier which assumes uniform prior probabilities. To cope with unbalanced classes, SVM practitioners routinely use two heuristics:

1. Set a higher cost-factor for the class which is less represented in the training data. For example, if the first class is the smaller class, then set  $\omega > 0.5$ , i.e.,  $\omega^+ > \omega^-$ . This option is supported by all major SVM solvers like the SVM<sup>light</sup> (Joachims, 1999). A proper setting of the cost-factor  $\omega$  is then tuned as an additional hyper-parameter.
2. After the linear SVM classifier is trained, tune only the bias of the classifier to achieve desired error rate. Recall that the standard SVMs classifier (1) is unbiased.

Let us confront these heuristics with the plug-in Bayes classifier (12) whose parameters are obtained by the ML estimator (14). Theorem 2 shows that the hyper-parameter  $\omega$  is proportional to the prior probability  $p(y; \tau, \omega)$ . This is perfectly consistent with the first heuristic. For example, if the first class is less frequent in the training data then, according to Theorem 2, we should set  $\omega > 0.5$  to guarantee that  $p(y = +1; \tau, \omega) < p(y = -1; \tau, \omega)$  holds and vice-versa.

Regarding the second heuristic, we showed that plug-in Bayes classifier under the model (4) is a linear classification rule (12) with a bias term  $b = 2\omega - 1$ . That is, for uniform priors the bias is set to 0, while for unbalanced classes the bias is negative if the first class is more probable (and positive otherwise). Note that the probabilistic model exactly specifies the value of

the bias term given the hyper-parameter  $\omega$ . By contrast, in classical SVM training the bias must be tuned as an additional hyper-parameter.

## 5. Equivalence between Maximum Margin Clustering and Classification Maximum Likelihood approach

The Maximum Margin Clustering (MMC) (Xu et al., 2004) is a popular heuristic that transfers the maximum margin principle from supervised SVM learning to the unsupervised setting. In this section we show how our model (4) theoretically justifies MMC. In particular, we demonstrate how MMC emerges from our model by applying the Classification Maximum Likelihood (CML) approach (Scott & Symons, 1971), which is a statistically motivated and theoretically well-understood clustering method.

We consider the clustering problem defined as follows. Let  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in (\mathbb{R}^n \times \{+1, -1\})^m$  be i.i.d. from an underlying distribution  $p^*(\mathbf{x}, y)$ . Let us assume that we are given only the observations  $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$  and our goal is to estimate the corresponding hidden labels  $\{y_1, \dots, y_m\}$ .

The MMC finds labels by solving

$$\mathbf{y}_{\text{MMC}}(\lambda) = \underset{\mathbf{y} \in \{+1, -1\}^m}{\operatorname{argmin}} \min_{\mathbf{w} \in \mathbb{R}^n} \left[ \frac{\lambda}{2} \|\mathbf{w}\|^2 + R_{\text{MMC}}(\mathbf{y}, \mathbf{w}) \right] \quad (23)$$

where  $R_{\text{MMC}}(\mathbf{y}, \mathbf{w}) = \sum_{i=1}^m \ell(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)$ . Thus the MMC searches for the labels which allow best separation of the data by the SVM classifier. The problem (23) may not have a unique minimizer and hence  $\mathbf{y}_{\text{MMC}}(\lambda)$  denotes a set. The problem (23) is difficult for optimization due to the integer variables  $\mathbf{y} = (y_1, \dots, y_m)$ . A plethora of algorithms have been proposed to solve (23) approximately.

As before, we assume an unbiased linear classifier (i.e. the hyperplane passes through the origin). With this, (23) is a well-posed problem—unlike the variant with a biased classifier, which has a trivial solution assigning all observations to just one class. This complication is can also be solved by introducing an additional balance constraint enforcing solutions with prescribed number of labels in each class. Note that all derivations below can easily be repeated with the balance constraint to recover the biased variant of the MMC.

### 5.1. Classification Maximum Likelihood approach to clustering

Assume a conditional density  $p(\mathbf{x} | y; \boldsymbol{\theta})$  parametrized by  $\boldsymbol{\theta} \in \Theta$ . Given the observations  $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ , the

NLL of labels  $\mathbf{y} \in \{+1, -1\}^m$  and a parameter  $\boldsymbol{\theta}$  is

$$L(\mathbf{y}, \boldsymbol{\theta}) = - \sum_{i=1}^m \log p(\mathbf{x}_i | y_i; \boldsymbol{\theta}).$$

The CML approach finds labels by solving

$$\mathbf{y}_{\text{CML}} \in \underset{\mathbf{y} \in \{+1, -1\}^m}{\operatorname{argmin}} \min_{\boldsymbol{\theta} \in \Theta} L(\mathbf{y}, \boldsymbol{\theta}).$$

The CML assumes that both the vector  $\boldsymbol{\theta}$  and the labels  $\mathbf{y}$  are the parameters to be estimated.

Let us instantiate the CML approach for our model. We assume that the hyper-parameter  $\omega = 0.5$  and  $\tau$  is fixed otherwise (e.g. tuned on validation set). Recall that  $\omega = 0.5$  implies the uniform prior  $p(y; \tau, 0.5) = 0.5$ . Then, the NLL under the model (4) reads

$$L(\mathbf{y}, \mathbf{u}; \tau) = \sum_{i=1}^m \left[ \log h(\mathbf{x}_i) - \omega^{y_i} \langle \tau \mathbf{u}, \mathbf{x}_i \rangle - I^{y_i}(\tau, \omega) \right].$$

and the labels are found by solving

$$\begin{aligned} \mathbf{y}_{\text{CML}}(\tau) &= \underset{\mathbf{y} \in \{+1, -1\}^m}{\operatorname{argmin}} \min_{\mathbf{u} \in \mathcal{U}} L(\mathbf{y}, \mathbf{u}; \tau) \\ &= \underset{\mathbf{y} \in \{+1, -1\}^m}{\operatorname{argmin}} \min_{\mathbf{u} \in \mathcal{U}} R_{\text{MMC}}(\mathbf{y}, \tau \mathbf{u}). \end{aligned} \quad (24)$$

### 5.2. Equivalence between MMC and CML approach

Now, we show that any clustering returned by the MMC can be found by the CML approach.

**Theorem 5** *Let  $\mathbf{y}_{\text{MMC}}(\lambda)$  be a set of minimizers of the Maximum Margin Clustering problem (23) for some  $\lambda \in \mathbb{R}^{++}$ . Then, for any labeling  $\mathbf{y}^* \in \mathbf{y}_{\text{MMC}}(\lambda)$  there exists  $\tau \in \mathbb{R}^+$  such that  $\mathbf{y}^*$  is a minimizer of the Classification Maximum Likelihood problem (24), i.e.,  $\mathbf{y}^* \in \mathbf{y}_{\text{CML}}(\tau)$  holds.*

PROOF: Because  $\mathbf{y}^* \in \mathbf{y}_{\text{MMC}}(\lambda)$  is a minimizer of the problem (23) the inequality

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^n} \left[ \frac{\lambda}{2} \|\mathbf{w}\|^2 + R_{\text{MMC}}(\mathbf{y}^*, \mathbf{w}) \right] \\ \leq \min_{\mathbf{w} \in \mathbb{R}^n} \left[ \frac{\lambda}{2} \|\mathbf{w}\|^2 + R_{\text{MMC}}(\mathbf{y}, \mathbf{w}) \right], \end{aligned} \quad (25)$$

holds  $\forall \mathbf{y} \in \{+1, -1\}^m$ . Let us denote the minimizer of the left hand side of (25) as  $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^n} \left[ \frac{\lambda}{2} \|\mathbf{w}\|^2 + R_{\text{MMC}}(\mathbf{y}^*, \mathbf{w}) \right]$  which is unique as the objective is strictly convex for fixed  $\mathbf{y}^*$ . Let us denote  $\tau = \|\mathbf{w}^*\|$  and  $\mathcal{W}_\tau = \{\mathbf{w} \in \mathbb{R}^n \mid \|\mathbf{w}\| = \tau\}$ . Then, we can derive from (25) that the inequality

$$\min_{\mathbf{w} \in \mathcal{W}_\tau} R_{\text{MMC}}(\mathbf{y}^*, \mathbf{w}) \leq \min_{\mathbf{w} \in \mathcal{W}_\tau} R_{\text{MMC}}(\mathbf{y}, \mathbf{w}) \quad (26)$$

holds  $\forall \mathbf{y} \in \{+1, -1\}^m$ . To get from (25) to (26), we used the fact that all vectors in  $\mathcal{W}_\tau$  have the same norm and one of them is the minimizer  $\mathbf{w}^*$ . The inequality (26) implies that  $\mathbf{y}^*$  is a minimizer of

$$\begin{aligned} \mathbf{y}^* &\in \operatorname{argmin}_{\mathbf{y} \in \{+1, -1\}^m} \min_{\mathbf{w} \in \mathcal{W}_\tau} R_{\text{MMC}}(\mathbf{y}, \mathbf{w}) \\ &= \operatorname{argmin}_{\mathbf{y} \in \{+1, -1\}^m} \min_{\mathbf{u} \in \mathcal{U}} R_{\text{MMC}}(\mathbf{y}, \tau \mathbf{u}) \end{aligned}$$

where the letter equality, obtained by the variable substitution  $\mathbf{w} = \tau \mathbf{u}$ , is just the definition of the CML problem (24) which was to be proved. ■

The established correspondence between the MMC and the CML not only provide a theoretical justification of the MMC but it also opens ways for its extension. First, to cope with the unbalanced data one can tune the hyper-parameter  $\omega$  which would correspond to changing the prior probability  $p(y; \tau, \omega)$ . Second, the hard problem (23) required by the MMC can be attacked by algorithms routinely used for minimization of the CML criterion. Namely, the Classification Expectation Algorithm (CME) (Celeux & Covaert, 1992) is a simple iterative procedure which transforms the hard unsupervised problem to a series of much simpler supervised problems. In turn, the existing SVMs solvers can be readily recycled for solving the MMC.

## 6. Conclusion

The received wisdom in machine learning has so far been that maximum margin SVM learning and probabilistic models constitute two separate sub-domains of machine learning. Our work has been motivated by the unsettling fact that SVM-like methods, albeit being rooting in learning theory and being powerful and efficient in practice, do not enjoy the principled view on modeling offered by probabilistic methods. In this contribution, we heal this rupture by setting up a probabilistic model that is equivalent to the SVM. So far, however, this work is limited to linear SVMs without bias; whether and how kernelization can be incorporated remains to be investigated.

Apart from the theoretical satisfaction of unification, the probabilistic understanding of the SVM can also lead to further insight. As an example, we demonstrate how a common and empirically successful heuristic for dealing with unbalanced class sizes can be understood in terms of biased priors, and how maximum margin clustering is naturally linked to the generic CML (Classification Maximum Likelihood) principle. Further work on semi-supervised SVMs is in progress, and we anticipate that many more such relationships will be discovered.

## Acknowledgments

VF was supported by the Czech Ministry of Education project 1M0567 and by EC projects FP7-ICT-247525 HUMAVIPS, PERG04-GA-2008-239455 SEMISOL.

## References

- Bartlett, P. L. and Tewari, A. Sparseness vs estimating conditional probabilities: Some asymptotic results. *Journal of Machine Learning Research*, 8:775–790, 2007.
- Celeux, G. and Covaert, G. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3), 1992.
- Grandvalet, Y., Mariéthoz, J., and Bengio, S. Interpretation of SVMs with an application to unbalanced classification. *Advances in Neural Information Processing Systems*, NIPS 18, 2005.
- Joachims, T. Making large-scale SVM learning practical. In Schölkopf, B., Burges, C. J. C., and Smola, A. J. (eds.), *Advances in Kernel Methods—Support Vector Learning*, pp. 169–184. MIT Press, Cambridge, MA, 1999.
- Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola et al. (ed.), *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 2000.
- Schölkopf, B. and Smola, A. J. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- Schölkopf, B, Smola, AJ, Williamson, RC, and Bartlett, PL. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.
- Scott, A.J. and Symons, M.J. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27: 387–389, 1971.
- Sollich, P. Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine learning*, 46(1):21–52, 2002.
- Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer, New York, 2008.
- Vapnik, V. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.
- Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. Maximum margin clustering. In *Proc. of Neural Information Processing Systems (NIPS)*, 2004.