Discussion of Grace Wahba's Wald Lecture 3

JSM-2003, San Francisco

Acknowledgement: Xiaotong Shen

Formulation of 2-class SVM, from Wahba's lecture:

$$y = \begin{matrix} +1 = \mathcal{A} \\ -1 = \mathcal{B} \end{matrix} \text{ (note different coding)}$$

Find $f(t) = d + h(t)$ with $h \in \mathcal{H}_K$ to min

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(t_i))_+ + \lambda \|h\|^2_{\mathcal{H}_K} \qquad (**)$$

where $(\tau)_+ = \tau, \tau > 0, = 0$ otherwise.

Then

$$f_\lambda(t) = d + \sum_{i=1}^{n} c_i K(t, t_i). \qquad (*)$$

Substitute (*) into (**), choose λ, given λ, find c and d.
The classifier is

$$f_\lambda(t) > 0 \rightarrow \mathcal{A}$$

The SVM cost has two terms

-- hinge loss: sum of $h(\tau_i) = (1 - \tau_i)_+$
where $\tau = y\, f(t)$

-- penalty: $\lambda \|w\|^2$ in the linear case

Historically this is motivated by consideration of geometric margin in the separable case.
Y. Lin showed the expected value of the hinge loss is minimized by the Bayes rule.
However, it is fruitful to try other forms of the loss and the penalty terms.

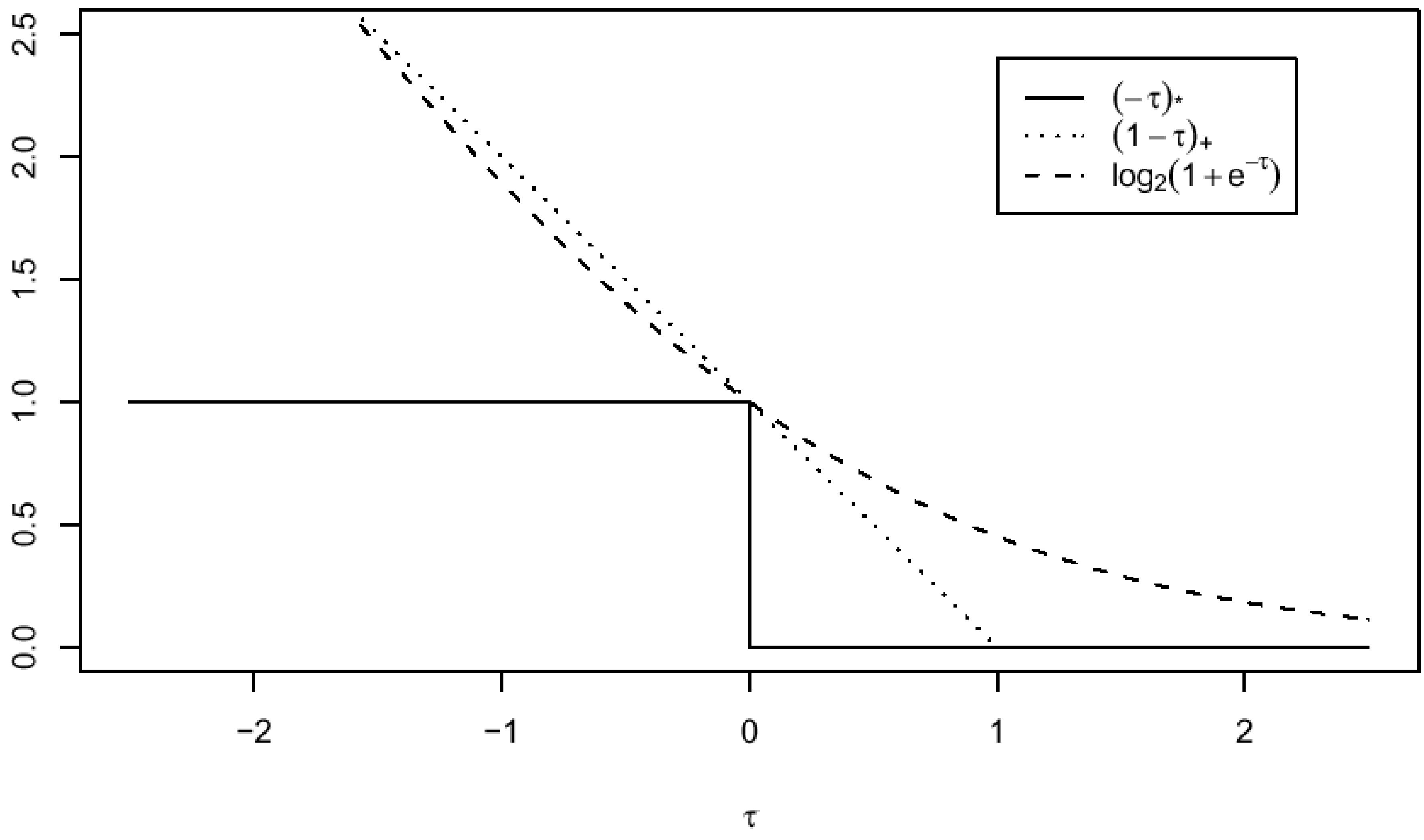


Figure 1. Let $\mathcal{C}(y_i, f(t_i)) = c(y_i f(t_i)) = c(\tau)$.

What we really care about is the

true error rate = $(1/2)E(1-\mathrm{sign}(y\, f(t)))$

The hinge loss is an convex upper bound of the

training error = $(1/2)\Sigma\, (1-\mathrm{sign}(y_i\, f(t_i)))$

What if we replace the hinge loss by the training error?

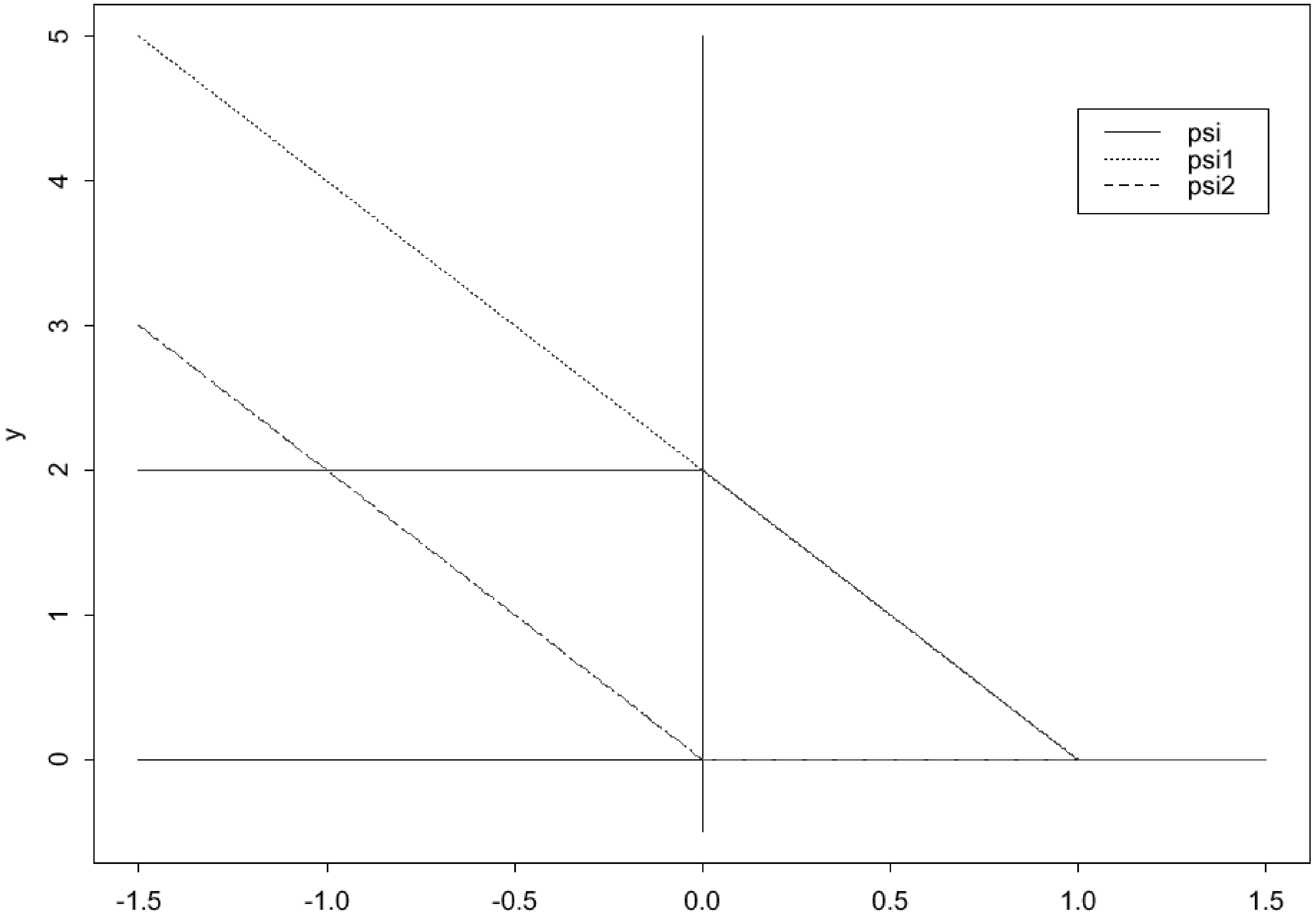
Difficulties with training error loss

- Scaling problem: $\mathrm{sign}(y\, f(t))$ is unchanged if we multiply $f(t)$ by a positive constant. This pushes the solution towards zero.

- Non-convexity: optimization is difficult.

ψ-learning: use $\psi_{dc}(\tau) = \psi_1(\tau) - \psi_2(\tau)$ instead of $(1\text{-sign}(\tau))$

Shen, Tseng, Zhang, Wong (2003, JASA)



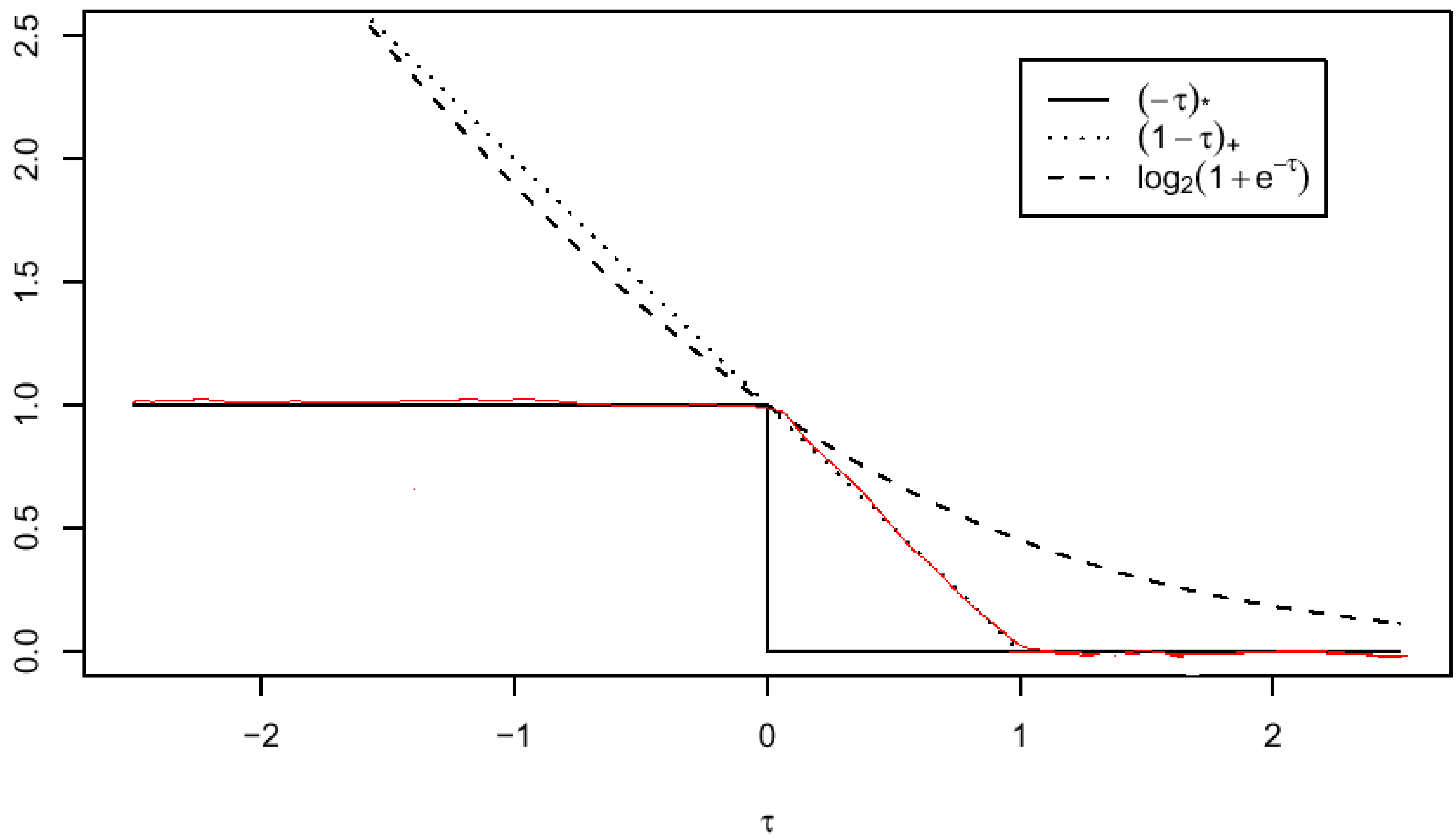


Figure 1. Let $\mathcal{C}(y_i, f(t_i)) = c(y_i f(t_i)) = c(\tau)$.

Linear:

1. Decision functions: $f(x) = \sum_{i=1}^{d} w_i x_i + b$.
2. Find (w, b) to minimize

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \psi(y_i f(x_i)).$$

3. Tuning parameter: $C > 0$.

Nonlinear:

1. Decision functions: $f(x) = g(x) + b$ with $g(x) = \sum_{i=1}^{n} w_i K(x_i, x)$.
2. Kernel: K (satisfy some assumptions).
3. Find (w, b) to minimize

$$\frac{1}{2}\|g\|_K^2 + C \sum_{i=1}^{n} \psi(y_i f(x_i))$$

where $w = (w_1, \cdots, w_n)$.

Theory

- $E(\psi_{dc}(y\, f(x)))$ is also minimized by the Bayes classifier.
- Convergence rate in terms of excess true error rate is available. It depends on the entropy and approximation rate of the sieve decision set space, the continuity property of the class densities, etc.
- The rate is typically faster than the rate of the SVM.

Computation

- Cost = $S_1 - S_2$
- $S_1 = \lambda\|w\|^2 + \Sigma\, \psi_1(y_i f(x_i))$
- $S_2 = \Sigma\, \psi_2(y_i f(x_i))$
- Thus, cost is a DC function. The powerful Difference of Convex Algorithm (DCA, An and Tao, 1997) to handle the optimization. (Shen et al, draft.)

Comments on MSVM

It is very nice that the MSVM cost

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k} L_{cat(i)r}(f^r(t_i) - y_{ir})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}$$

targets the right function.

It will be useful to also establish its relation to the training error rate. Is it an upper bound (as in the 2 class case)? **The challenge in multi-category problems lies partly our inability to design a loss that approximate the error counting function.**

Other issues

- Stability of SVM in high dimension, small sample cases
- Model selection and variable selection problems

Resampling are often useful for these problems.