Statistical Learning Theory and Support Vector Machines

Gert Cauwenberghs
Johns Hopkins University
gert@jhu.edu

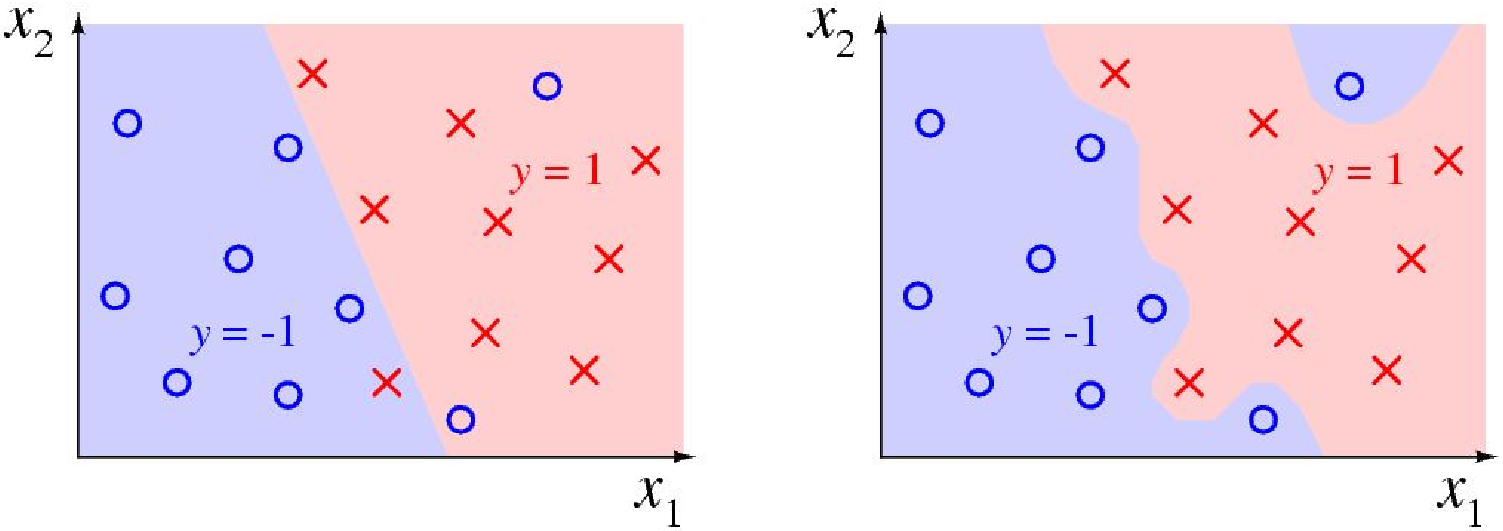
520.776 Learning on Silicon
http://bach.ece.jhu.edu/gert/courses/776

Statistical Learning Theory and Support Vector Machines

OUTLINE

- **Introduction to Statistical Learning Theory**
  - VC Dimension, Margin and Generalization
  - Support Vectors
  - Kernels
- **Cost Functions and Dual Formulation**
  - Classification
  - Regression
  - Probability Estimation
- **Implementation: Practical Considerations**
  - Sparsity
  - Incremental Learning
- **Hybrid SVM-HMM MAP Sequence Estimation**
  - Forward Decoding Kernel Machines (FDKM)
  - Phoneme Sequence Recognition (TIMIT)

Generalization and Complexity



- Generalization is the key to supervised learning, for classification or regression.
- *Statistical Learning Theory* offers a principled approach to understanding and controlling generalization performance.
  - *The complexity of the hypothesis class of functions determines generalization performance.*
  - *Complexity relates to the effective number of function parameters, but effective control of margin yields low complexity even for infinite number of parameters.*

VC Dimension and Generalization Performance

Vapnik and Chervonenkis, 1974

– For a *discrete* hypothesis space H of functions, with probability $1\text{-}\delta$:

$$E[y \neq f(\mathbf{x})] \;\leq\; \frac{1}{m}\sum_{i=1}^{m}(y_i \neq f(\mathbf{x}_i)) \;+\; \sqrt{\frac{2}{m}\ln\frac{2|H|}{\delta}}$$

Generalization error — Empirical (training) error — Complexity

where $f = \arg\min_{f\in H}\sum_{i=1}^{m}(y_i \neq f(\mathbf{x}_i))$ minimizes empirical error over m training samples $\{\mathbf{x}_i, y_i\}$, and $|H|$ is the cardinality of H.

– For a *continuous* hypothesis function space H, with probability $1\text{-}\delta$:

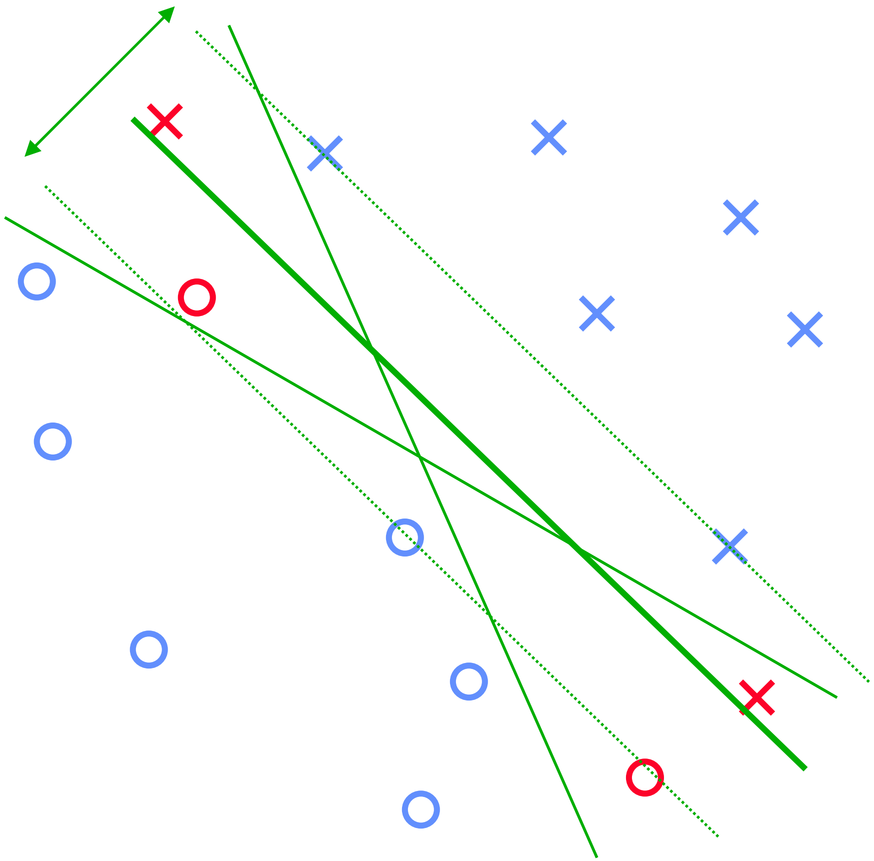
$$E[y \neq f(\mathbf{x})] \;\leq\; \frac{1}{m}\sum_{i=1}^{m}(y_i \neq f(\mathbf{x}_i)) \;+\; \sqrt{\frac{c}{m}\left(d + \ln\frac{1}{\delta}\right)}$$

where d is the *VC dimension* of H, the largest number of points $\mathbf{x}_i$ completely "shattered" (separated in all possible combinations) by elements of H.

– For linear classifiers $f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}\cdot\mathbf{X} + b)$ in N dimensions, the VC dimension is the number of parameters, $N + 1$.

– For linear classifiers with margin ρ over a domain contained within diameter D, the VC dimension is bounded by D/ρ.

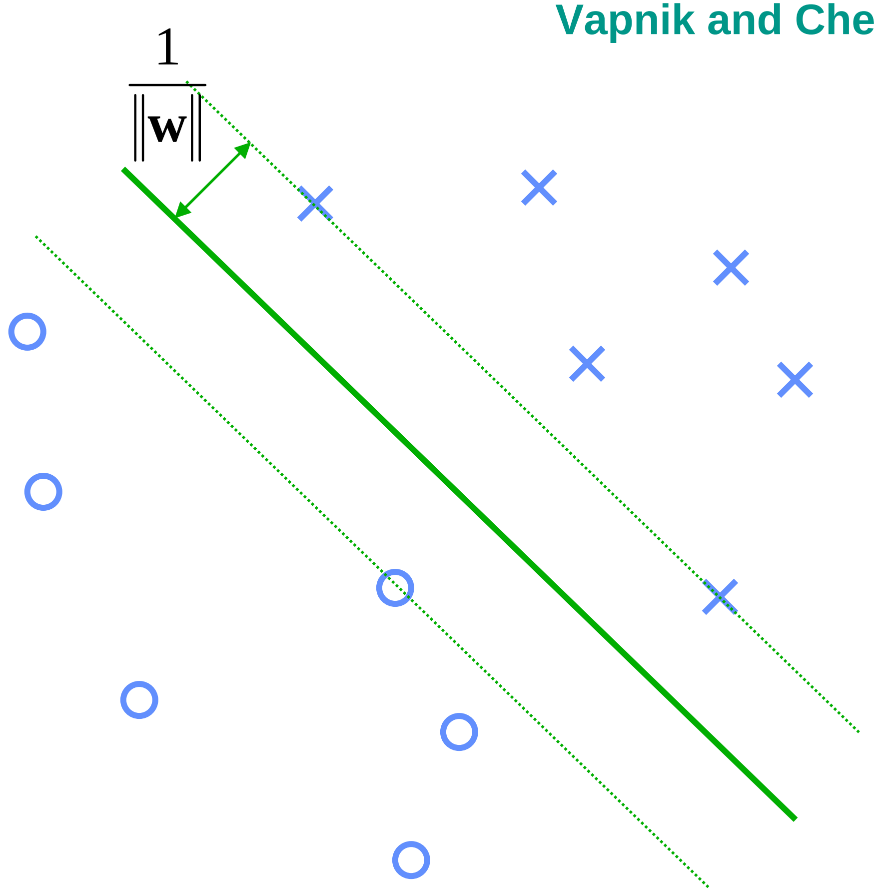
Learning to Classify Linearly Separable Data



- vectors $\mathbf{X}_i$
- labels $y_i = \pm 1$

Optimal Margin Separating Hyperplane

Vapnik and Lerner, 1963
Vapnik and Chervonenkis, 1974



- vectors $\mathbf{X}_i$
- labels $y_i = \pm 1$

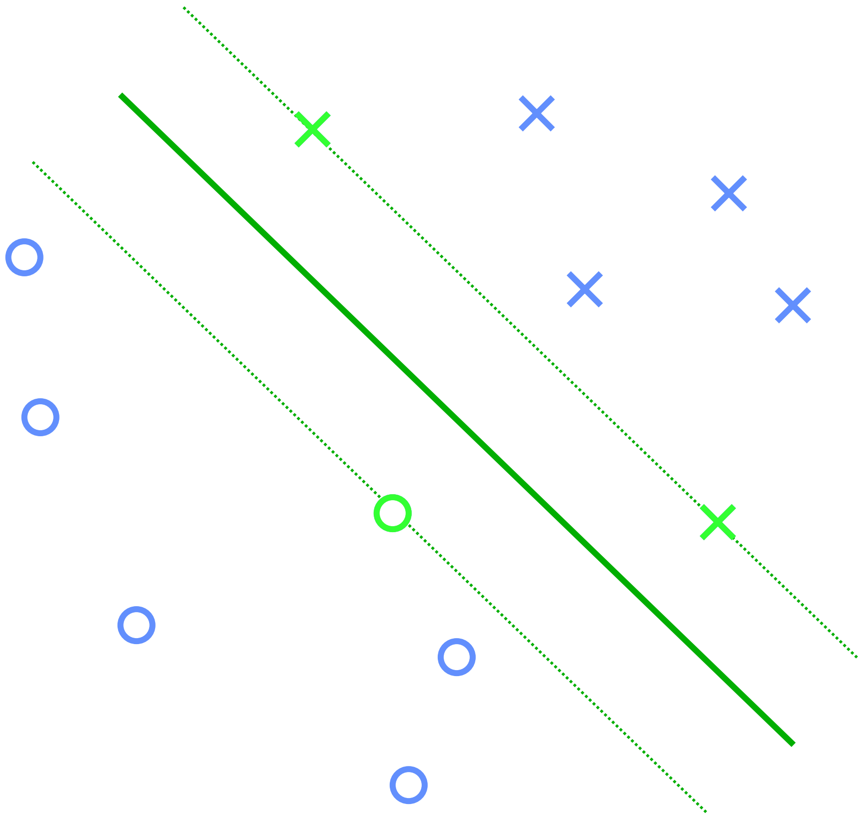
$$y = \mathrm{sign}(\mathbf{w} \cdot \mathbf{X} + b)$$

$$y_i(\mathbf{w} \cdot \mathbf{X}_i + b) \geq 1$$

$$\min_{\mathbf{w},b} : \|\mathbf{w}\|$$

Support Vectors

Boser, Guyon and Vapnik, 1992



- vectors $\mathbf{X}_i$
- labels $y_i = \pm 1$

$$y = \mathrm{sign}(\mathbf{w} \cdot \mathbf{X} + b)$$

$$y_i(\mathbf{w} \cdot \mathbf{X}_i + b) \geq 1$$

$$\min_{\mathbf{w},b} : \|\mathbf{w}\|$$

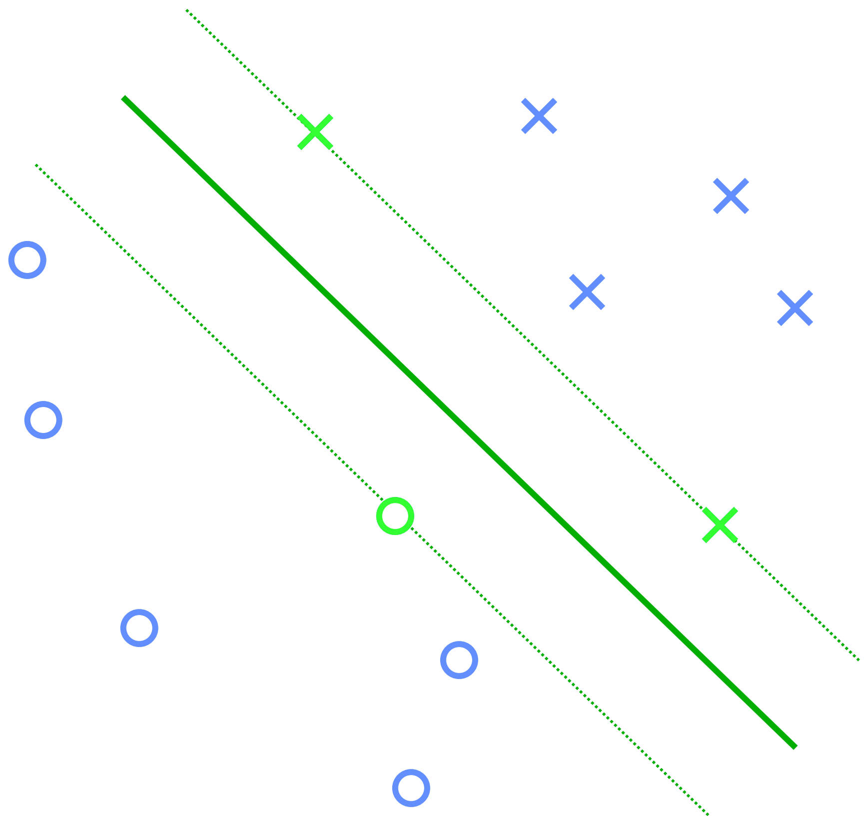
- support vectors:

$$y_i(\mathbf{w} \cdot \mathbf{X}_i + b) = 1, \quad i \in S$$

$$\mathbf{w} = \sum_{i \in S} \alpha_i y_i \mathbf{X}_i$$

Support Vector Machine (SVM)

Boser, Guyon and Vapnik, 1992



- vectors $\mathbf{X}_i$
- labels $y_i = \pm 1$

$$y = \mathrm{sign}(\mathbf{w}\cdot\mathbf{X} + b)$$

$$y_i(\mathbf{w}\cdot\mathbf{X}_i + b) \geq 1$$

$$\min_{\mathbf{w},b} : \|\mathbf{w}\|$$

- support vectors:

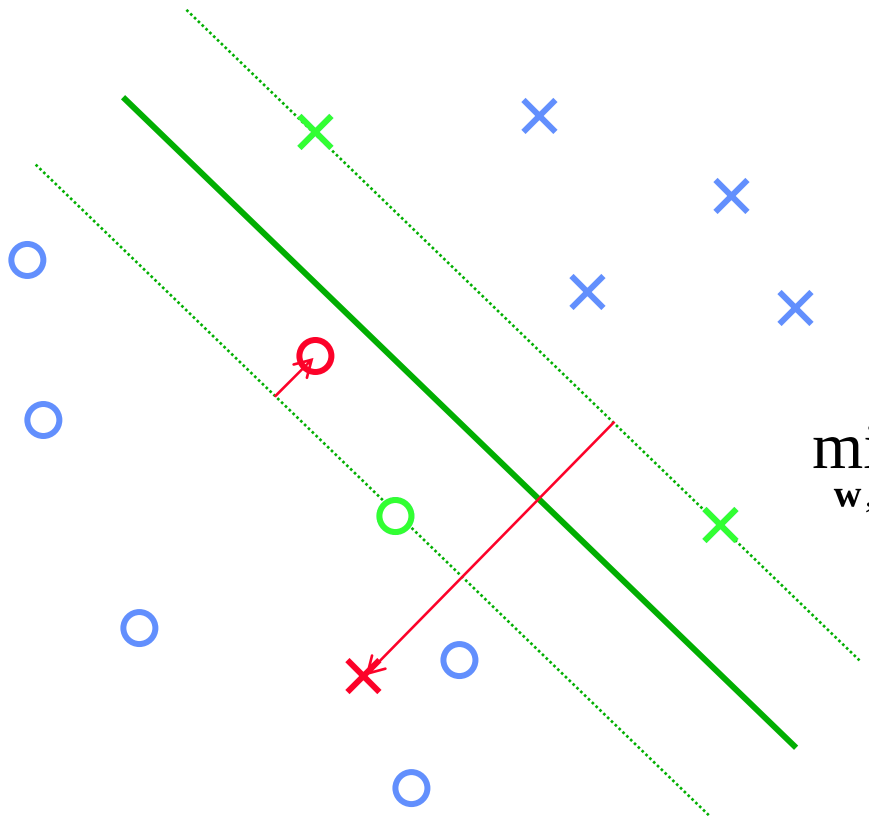
$$y_i(\mathbf{w}\cdot\mathbf{X}_i + b) = 1, \quad i \in S$$

$$\mathbf{w} = \sum_{i\in S} \alpha_i y_i \mathbf{X}_i$$

$$y = \mathrm{sign}(\sum_{i\in S} \alpha_i y_i \mathbf{X}_i \cdot \mathbf{X} + b)$$

Soft Margin SVM

Cortes and Vapnik, 1995



- vectors $\mathbf{X}_i$
- labels $y_i = \pm 1$

$$y = \mathrm{sign}(\mathbf{w} \cdot \mathbf{X} + b)$$

$$\min_{\mathbf{w},b} : \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \left[1 - y_i(\mathbf{w}\cdot\mathbf{X}_i + b)\right]^+$$

- support vectors:

  (margin and error vectors)

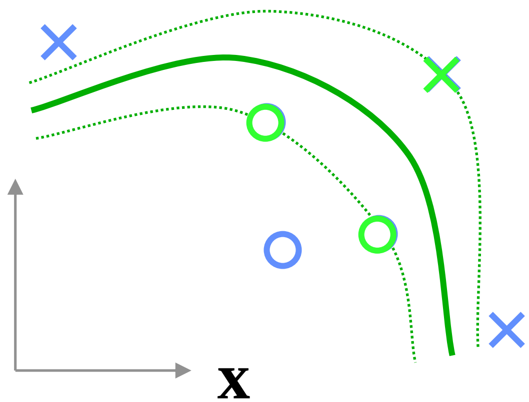
$$y_i(\mathbf{w}\cdot\mathbf{X}_i + b) \le 1, \quad i \in S$$

$$\mathbf{w} = \sum_{i\in S} \alpha_i y_i \mathbf{X}_i$$

$$y = \mathrm{sign}\left(\sum_{i\in S} \alpha_i y_i \mathbf{X}_i \cdot \mathbf{X} + b\right)$$

Kernel Machines

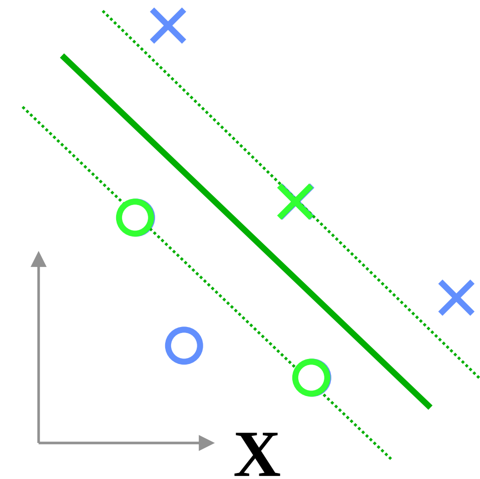
Mercer, 1909; Aizerman et al., 1964
Boser, Guyon and Vapnik, 1992



$$\Phi(\cdot)$$

$$\mathbf{X}_i = \Phi(\mathbf{x}_i)$$

$$\mathbf{X} = \Phi(\mathbf{x})$$



$$\mathbf{X}_i \cdot \mathbf{X} = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})$$

$$y = \mathrm{sign}(\sum_{i \in S} \alpha_i y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b)$$

$$y = \mathrm{sign}(\sum_{i \in S} \alpha_i y_i \mathbf{X}_i \cdot \mathbf{X} + b)$$

$$K(\cdot,\cdot)$$

$$\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) = K(\mathbf{x}_i, \mathbf{x})$$

Mercer's Condition

$$y = \mathrm{sign}(\sum_{i \in S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b)$$

Some Valid Kernels

Boser, Guyon and Vapnik, 1992

- Polynomial (Splines etc.)

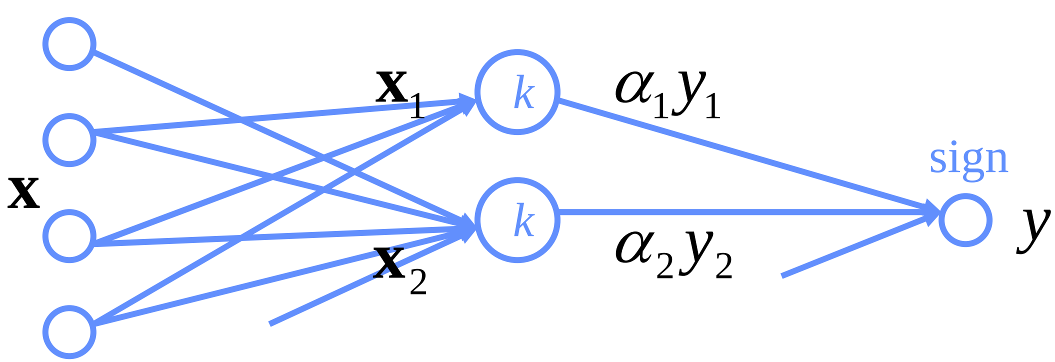
$$K(\mathbf{x}_i, \mathbf{x}) = (1 + \mathbf{x}_i \cdot \mathbf{x})^{\nu}$$

- Gaussian (Radial Basis Function Networks)

$$K(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}\|^2}{2\sigma^2}\right)$$

- Sigmoid (Two-Layer Perceptron)

$$K(\mathbf{x}_i, \mathbf{x}) = \tanh(L + \mathbf{x}_i \cdot \mathbf{x})$$ *only for certain L*



Other Ways to Arrive at Kernels…

- **Smoothness constraints in non-parametric regression** *[Wahba <<1999]*
  - Splines are radially symmetric kernels.
  - Smoothness constraint in the Fourier domain relates directly to (Fourier transform of) kernel.
- **Reproducing Kernel Hilbert Spaces (RKHS)** *[Poggio 1990]*
  - The class of functions $f(\mathbf{x}) = \sum_i c_i \varphi_i(\mathbf{x})$ with orthogonal basis $\varphi_i(\mathbf{x})$ forms a reproducing Hilbert space.
  - Regularization by minimizing the norm over Hilbert space yields a similar kernel expansion as SVMs.
- **Gaussian processes** *[MacKay 1998]*
  - Gaussian prior on Hilbert coefficients yields Gaussian posterior on the output, with covariance given by kernels in input space.
  - Bayesian inference predicts the output label distribution for a new input vector given old (training) input vectors and output labels.

Gaussian Processes

Neal, 1994
MacKay, 1998
Opper and Winther, 2000

– Bayes:

Posterior *Evidence* *Prior*

$$P(\mathbf{w}\,|\,y,\mathbf{x}) \propto P(y\,|\,\mathbf{x},\mathbf{w})P(\mathbf{w})$$

– Hilbert space expansion, with additive white noise:

$$y = f(\mathbf{x}) + n = \sum_i w_i \varphi_i(\mathbf{x}) + n$$

– Uniform Gaussian prior on Hilbert coefficients:

$$P(\mathbf{w}) = N(0, \sigma_w{}^2\mathbf{I})$$

yields Gaussian posterior on output:

$$P(y\,|\,\mathbf{x},\mathbf{w}) = N(0,\mathbf{C})$$

$$\mathbf{C} = \mathbf{Q} + \sigma_v{}^2\mathbf{I}$$

with kernel covariance

$$Q_{nm} = \sigma_w{}^2 \sum_i \varphi_i(\mathbf{x}_n)\varphi_i(\mathbf{x}_m) = k(\mathbf{x}_n,\mathbf{x}_m).$$

– Incremental learning can proceed directly through recursive computation of the inverse covariance (using a matrix inversion lemma).

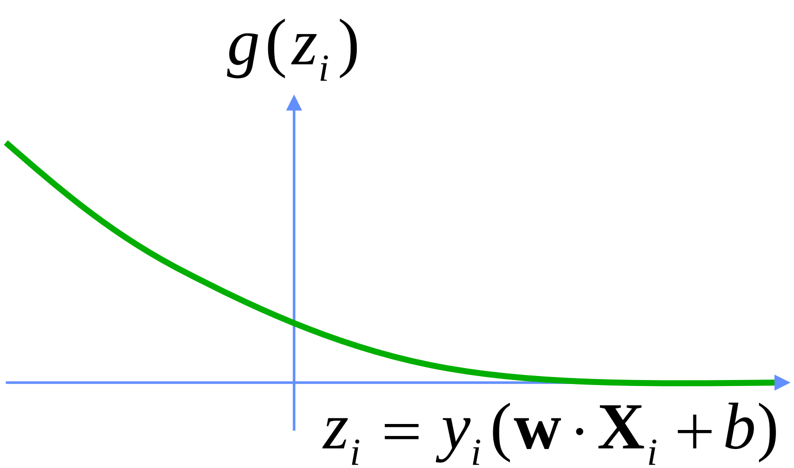
Kernel Machines: A General Framework

$$y = f(\mathbf{w}\cdot\Phi(\mathbf{x})+b) = f(\mathbf{w}\cdot\mathbf{X}+b)$$

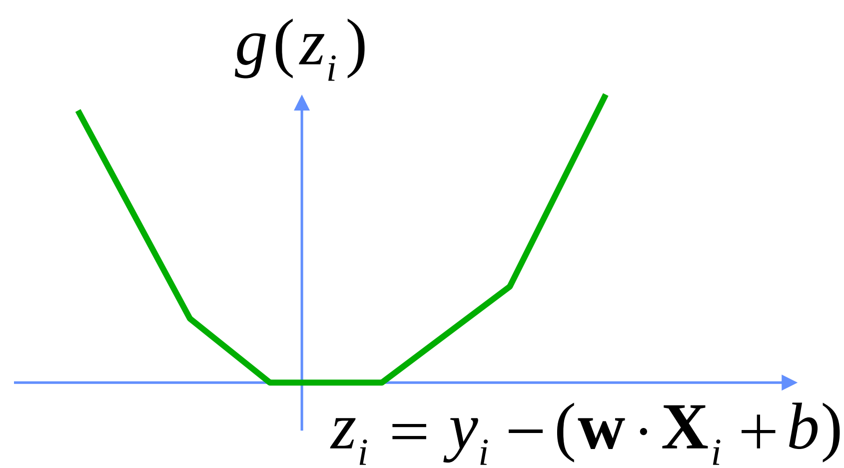
$$\min_{\mathbf{w},b} : \mathcal{E} = \tfrac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_i g(z_i)$$

Structural Risk	Empirical Risk	(SVMs)
Smoothness	*Fidelity*	(Regularization Networks)
Log Prior	*Log Evidence*	(Gaussian Processes)

- $g(.)$: convex cost function
- z_i: "margin" of each datapoint



Classification



Regression

Optimality Conditions

$$\min_{\mathbf{w},b} : \mathcal{E} = \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_i g(z_i)$$

$$z_i = y_i(\mathbf{w}\cdot\mathbf{X}_i + b)$$ *(Classification)*

– First-Order Conditions:

$$\frac{d\mathcal{E}}{d\mathbf{w}} \equiv 0 : \quad \mathbf{w} = -C\sum_i g'(z_i)y_i\mathbf{X}_i = \sum_i \alpha_i y_i \mathbf{X}_i$$

$$\frac{d\mathcal{E}}{db} \equiv 0 : \quad 0 = -C\sum_i g'(z_i)y_i = \sum_i \alpha_i y_i$$

with:

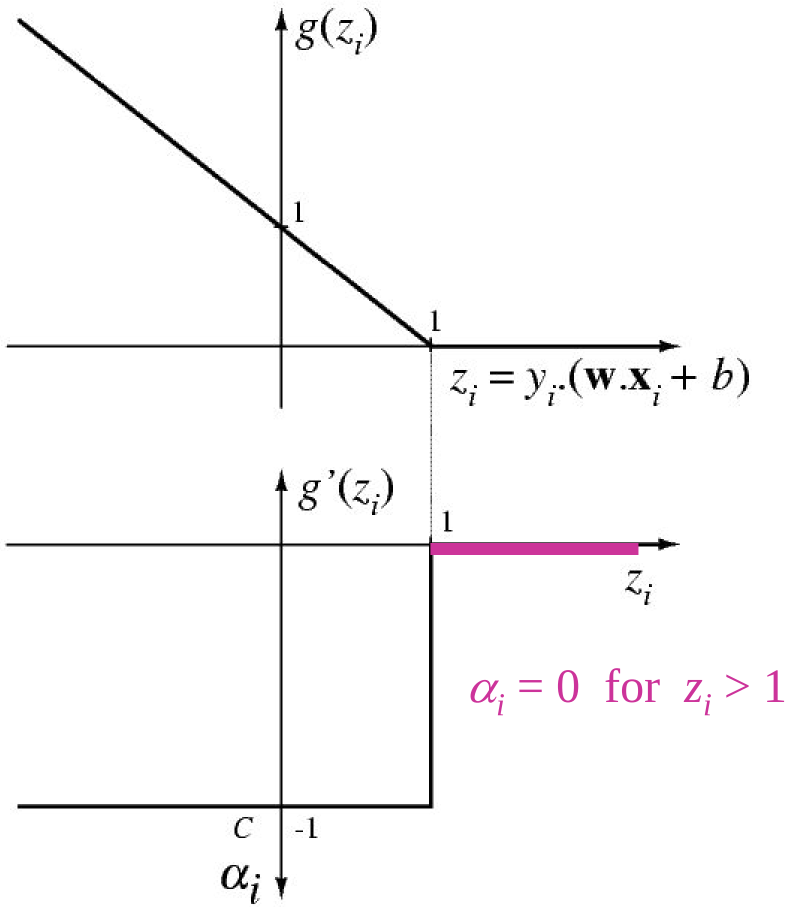
$$\alpha_i = -Cg'(z_i)$$

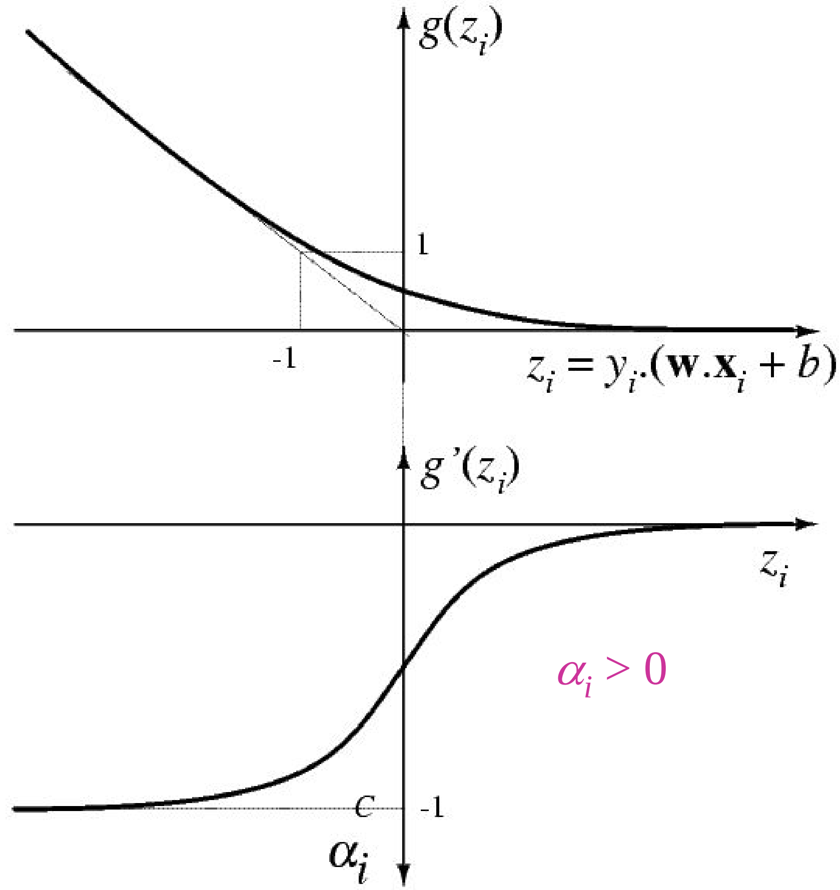
$$z_i = \sum_j Q_{ij}\alpha_j + by_i$$

$$Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

– Sparsity: $\alpha_i = 0$ requires $g'(z_i) = 0$

Sparsity





Soft-Margin SVM Classification

$$y = \mathrm{sign}(\mathbf{w} \cdot \mathbf{X} + b)$$

$$g(z_i) = \left[1 - y_i(\mathbf{w} \cdot \mathbf{X}_i + b)\right]^{+}$$

Logistic Probability Regression

$$\Pr(y \,|\, \mathbf{X}) = (1 + e^{-y(\mathbf{w} \cdot \mathbf{X} + b)})^{-1}$$

$$g(z_i) = \log(1 + e^{-y_i(\mathbf{w} \cdot \mathbf{X}_i + b)})$$

Dual Formulation

(Legendre transformation)

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{X}_i \qquad \alpha_i = -Cg'(z_i)$$

$$0 = \sum_i \alpha_i y_i \qquad z_i = \sum_j Q_{ij}\alpha_j + by_i$$

Eliminating the unknowns z_i:

$$z_i = \sum_j Q_{ij}\alpha_j + by_i = g'^{-1}(-\tfrac{\alpha_i}{C})$$

yields the equivalent of the first-order conditions of a "dual" functional $\mathcal{E}_2$ to be minimized in α_i:

$$\min_{\mathbf{w},b} : \mathcal{E}_2 = \tfrac{1}{2}\sum_i \sum_j \alpha_i Q_{ij} \alpha_j - C \sum_i G(\tfrac{\alpha_i}{C})$$

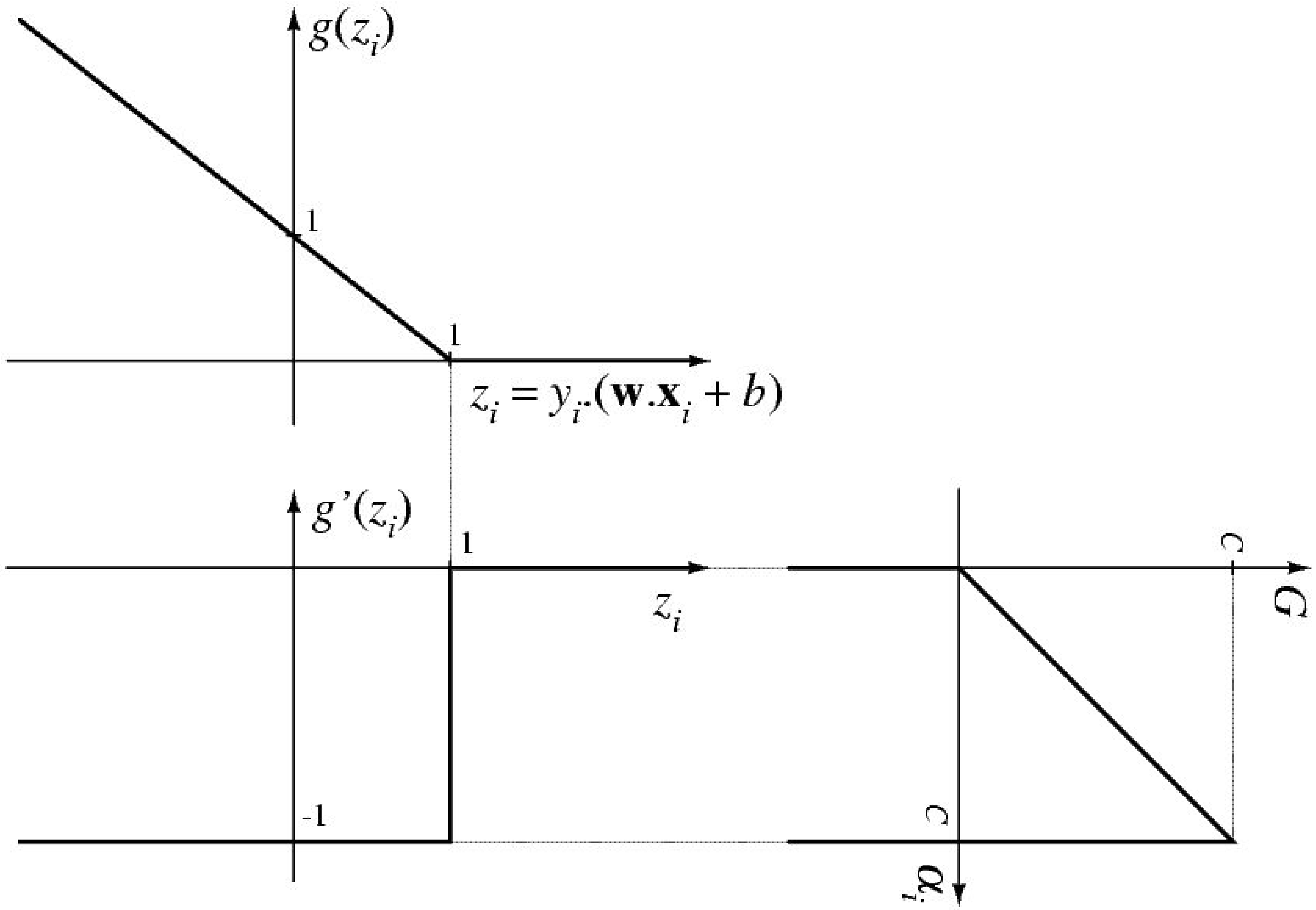
$$\text{subject to} : \sum_j y_i \alpha_i \equiv 0$$

with Lagrange parameter b, and "potential function"

$$G(u) = \int^u g'^{-1}(-v)\,dv$$

Soft-Margin SVM Classification

Cortes and Vapnik, 1995

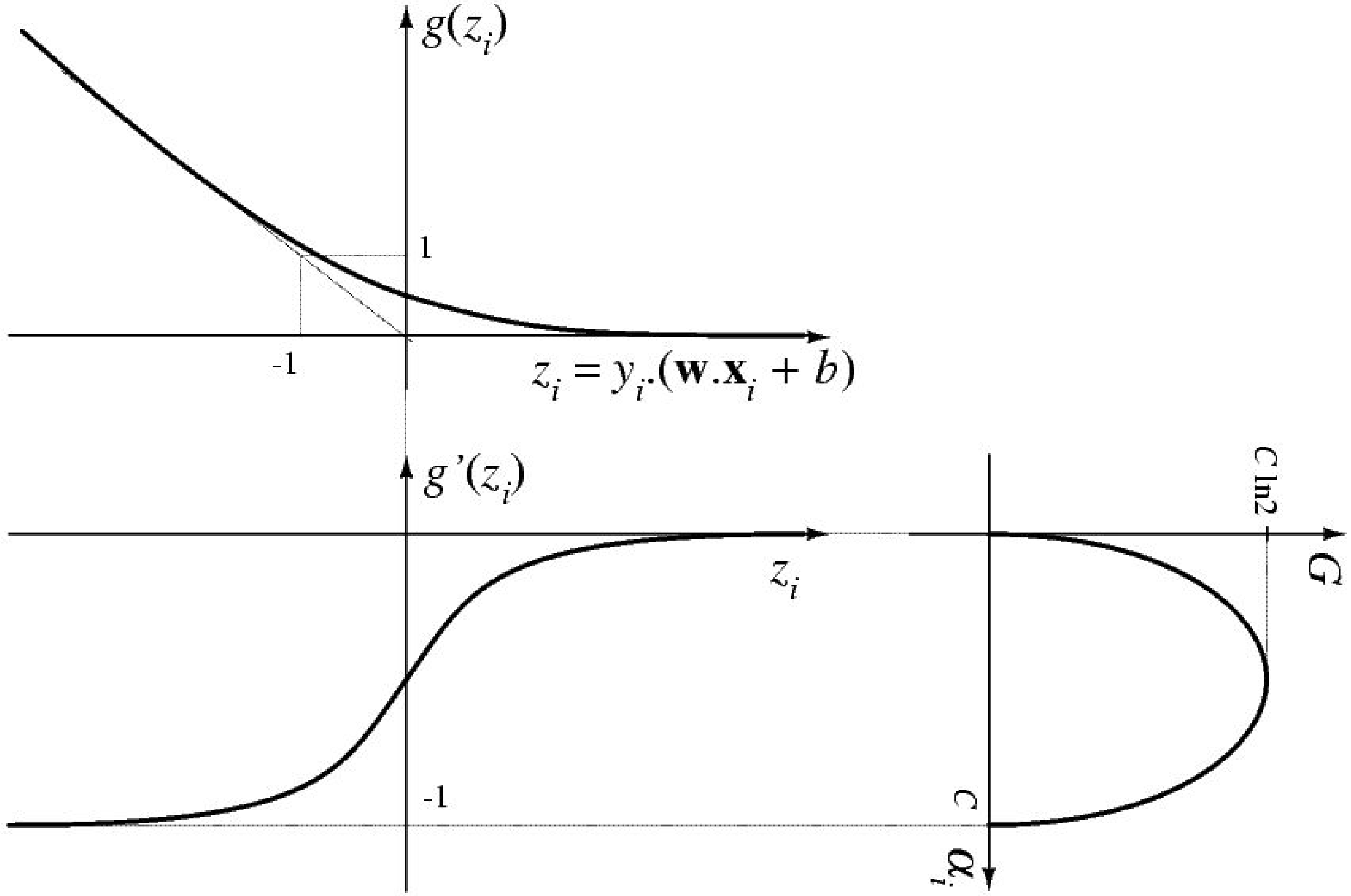


$$\min_{\mathbf{w},b} : \varepsilon_2^{SVcM} = \tfrac{1}{2}\sum_i \sum_j \alpha_i Q_{ij} \alpha_j - \sum_i \alpha_i$$

$$\text{subject to} : \sum_i y_i \alpha_i \equiv 0 \ \text{ and } \ 0 \le \alpha_i \le C, \forall i$$

Kernel Logistic Probability Regression

Jaakkola and Haussler, 1999

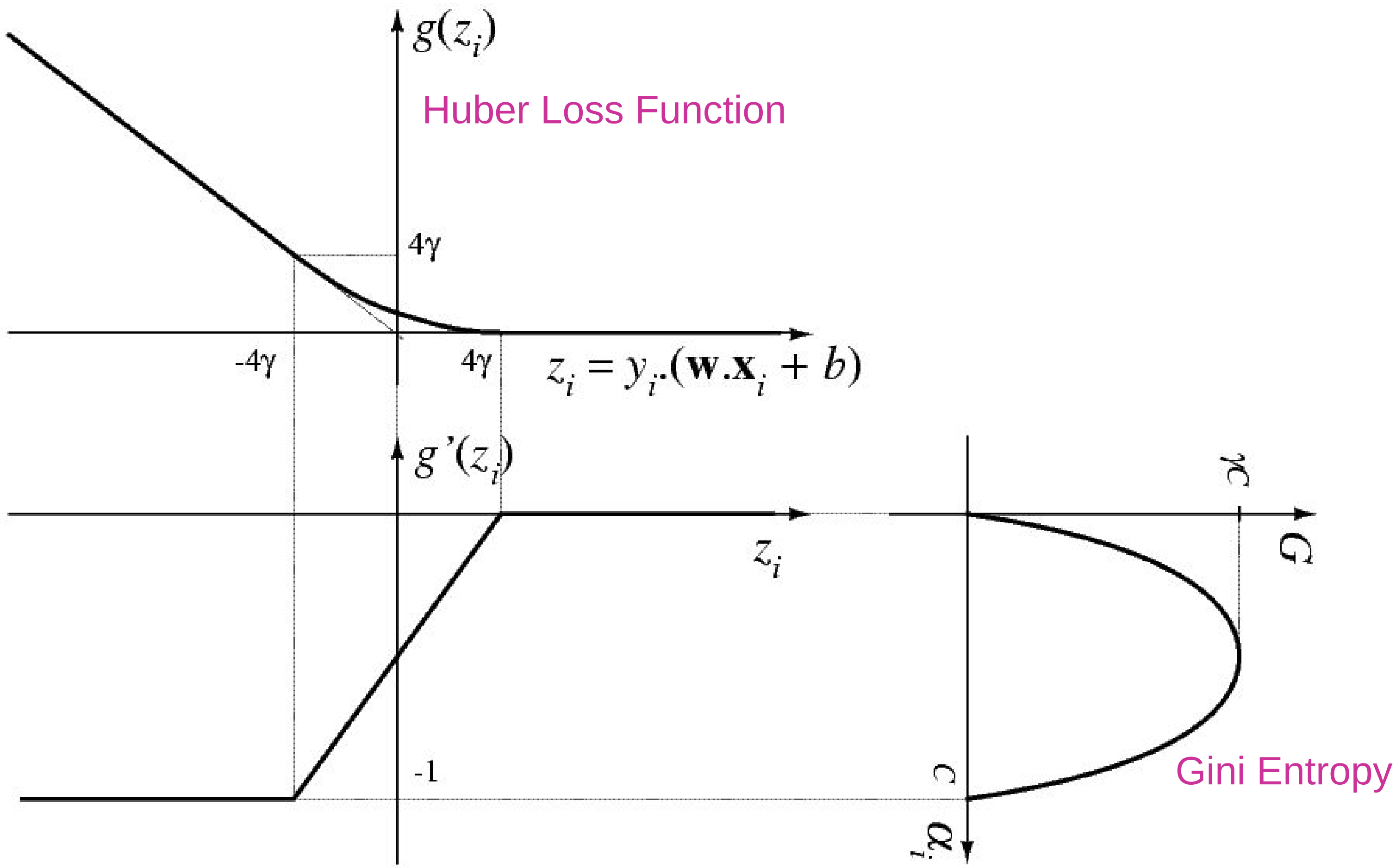


$$\min_{\mathbf{w},b} : \mathcal{E}_2^{kLR} = \tfrac{1}{2}\sum_i \sum_j \alpha_i Q_{ij} \alpha_j - C\sum_i H(\tfrac{\alpha_i}{C})$$

$$\text{subject to} : \sum_i y_i \alpha_i \equiv 0 \ \text{, with} \ \ H(a) = -a \ln a - (1-a)\ln(1-a)$$

*Gini*SVM Sparse Probability Regression

Chakrabartty and Cauwenberghs, 2002

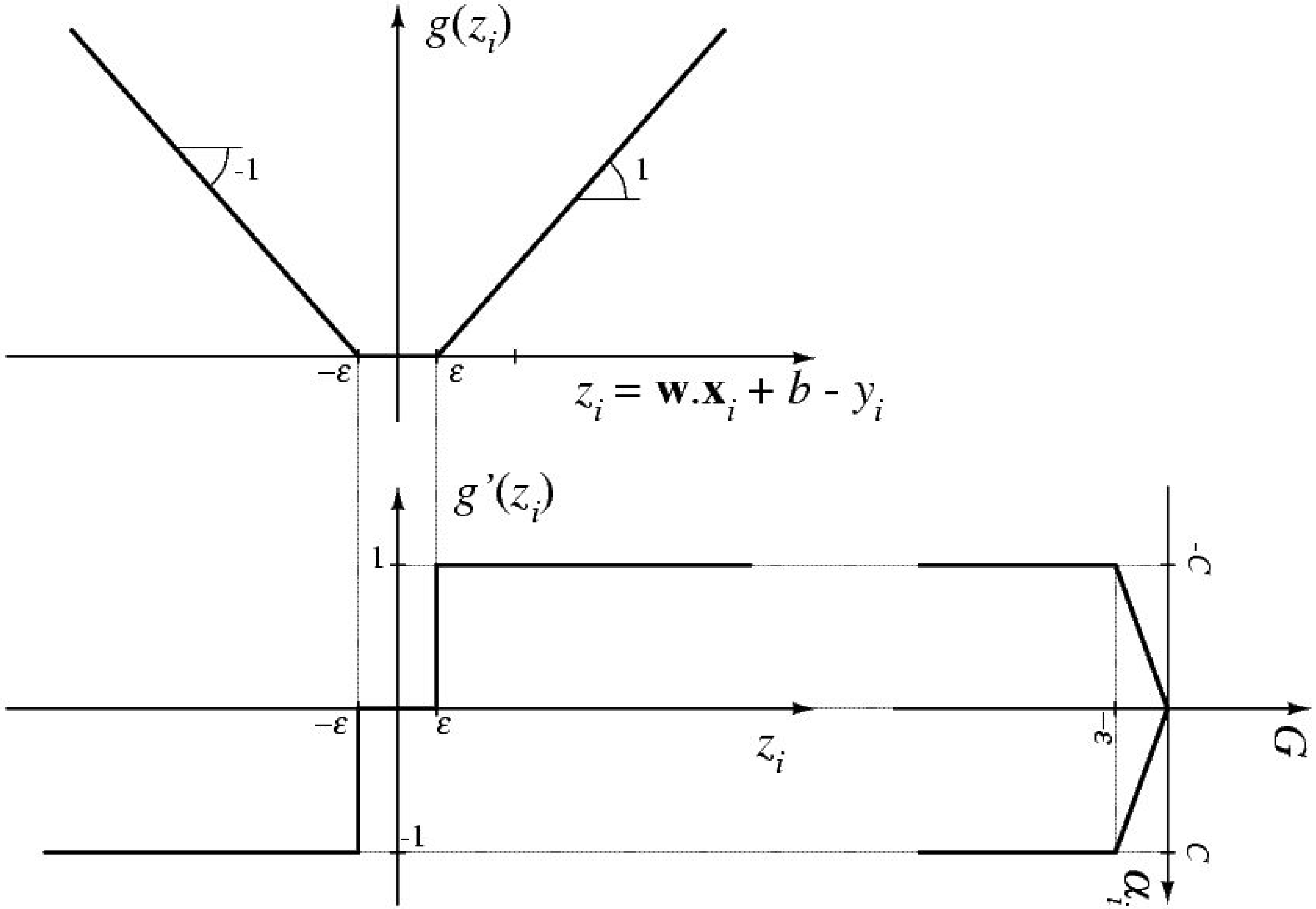


$$\min_{\mathbf{w},b} : \mathcal{E}_2^{kGini} = \tfrac{1}{2}\sum_i \sum_j \alpha_i Q_{ij} \alpha_j - C\sum_i H_{Gini}(\tfrac{\alpha_i}{C})$$

$$\text{subject to} : \sum_i y_i \alpha_i \equiv 0 \;\text{ and } 0 \le \alpha_i \le \mathrm{C}, \text{ with } \; H_{Gini}(a) = 4\gamma(1-a)a$$

Soft-Margin SVM Regression

Vapnik, 1995 ; Girosi, 1998



$$\min_{\mathbf{w},b} : \mathcal{E}_2^{SVrM} = \tfrac{1}{2}\sum_i \sum_j \alpha_i Q_{ij} \alpha_j + \varepsilon \sum_i |\alpha_i|$$

$$\text{subject to} : \sum_i y_i \alpha_i \equiv 0 \;\text{ and }\; 0 \le \alpha_i \le C, \forall i$$

Sparsity Reconsidered

Osuna and Girosi, 1999
Burges and Schölkopf, 1997
Cauwenberghs, 2000

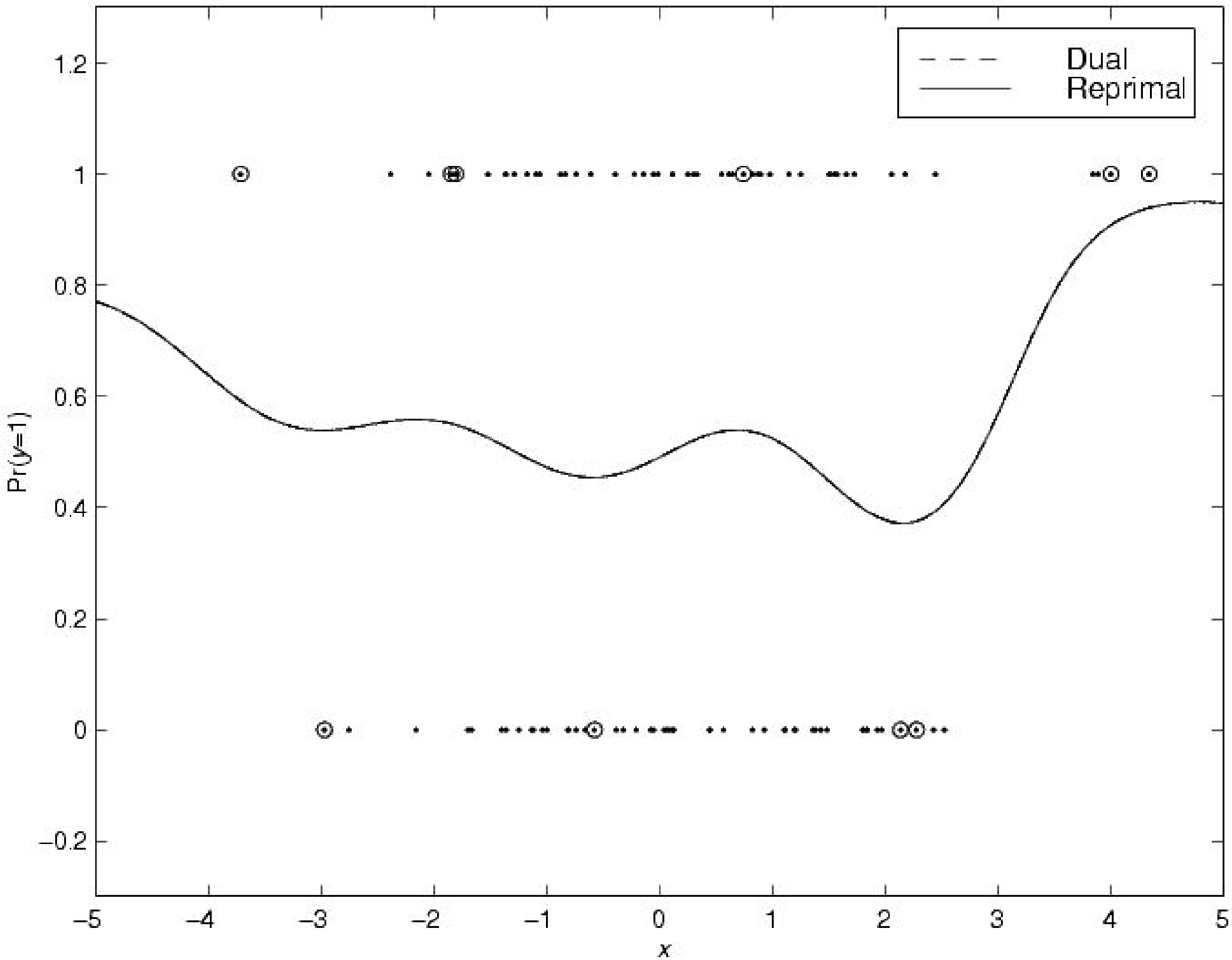
- The dual formulation gives a unique solution; however primal (re-) formulation may yield functionally equivalent solutions that are sparser, *i.e.* that obtain the same representation with fewer 'support vectors' (fewer kernels in the expansion).

$$\text{Dual } \alpha_j \text{ and (re-)primal } \alpha_j^* \text{ coefficients are equivalent}$$

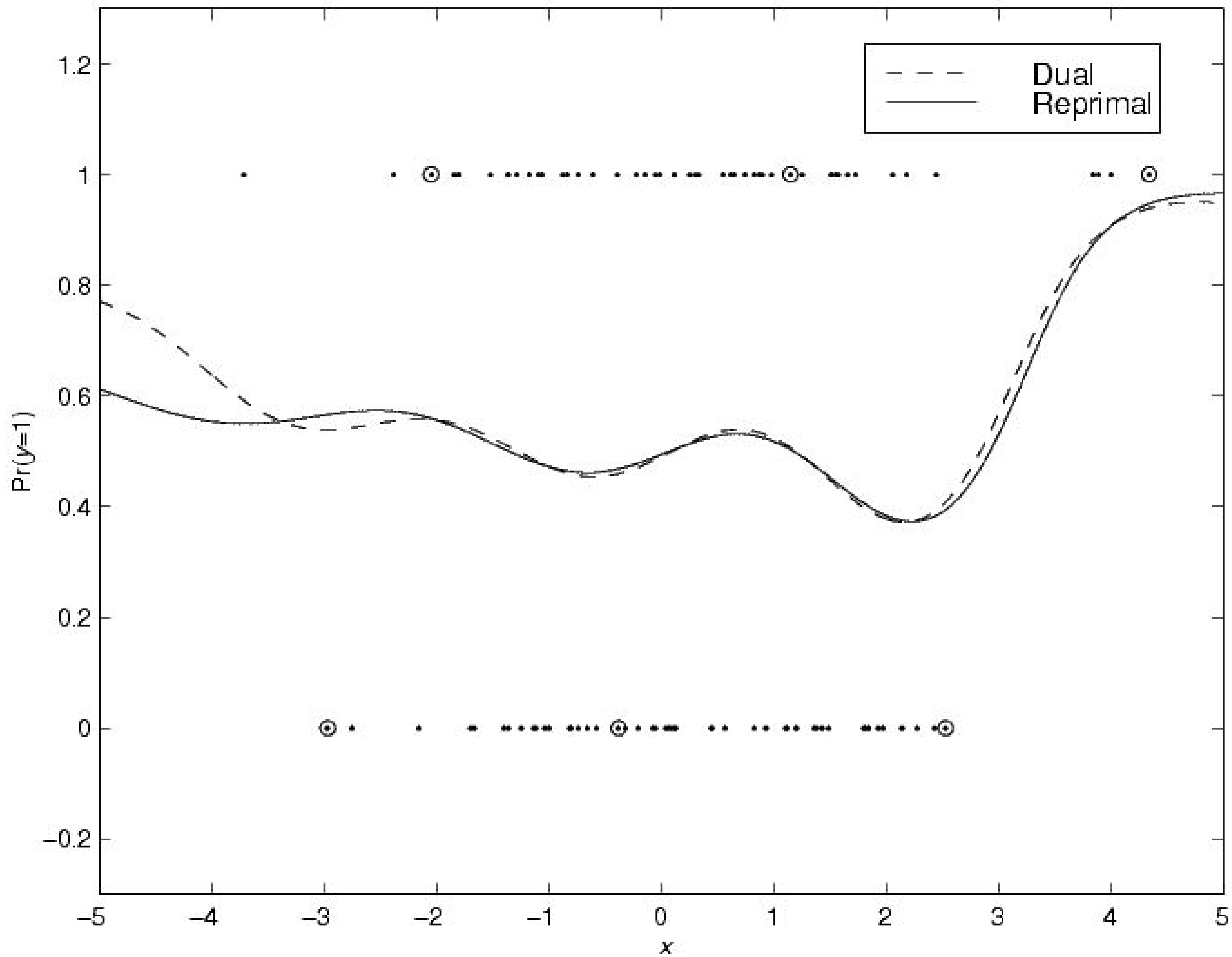
$$\Updownarrow$$

$$\sum_j Q_{ij}(\alpha_j^* - \alpha_j) \equiv 0 \quad \forall i$$

- The degree of (optimal) sparseness in the primal representation depends on the distribution of the input data in feature space. The tendency to sparseness is greatest when the kernel matrix $\boldsymbol{Q}$ is near to singular, i.e. the data points are highly redundant and consistent.



Logistic probability regression in one dimension, for a Gaussian kernel. Full dual solution (with 100 kernels), and approximate 10-kernel “reprimal” solution, obtained by truncating the kernel eigenspectrum to a 10^5 spread.

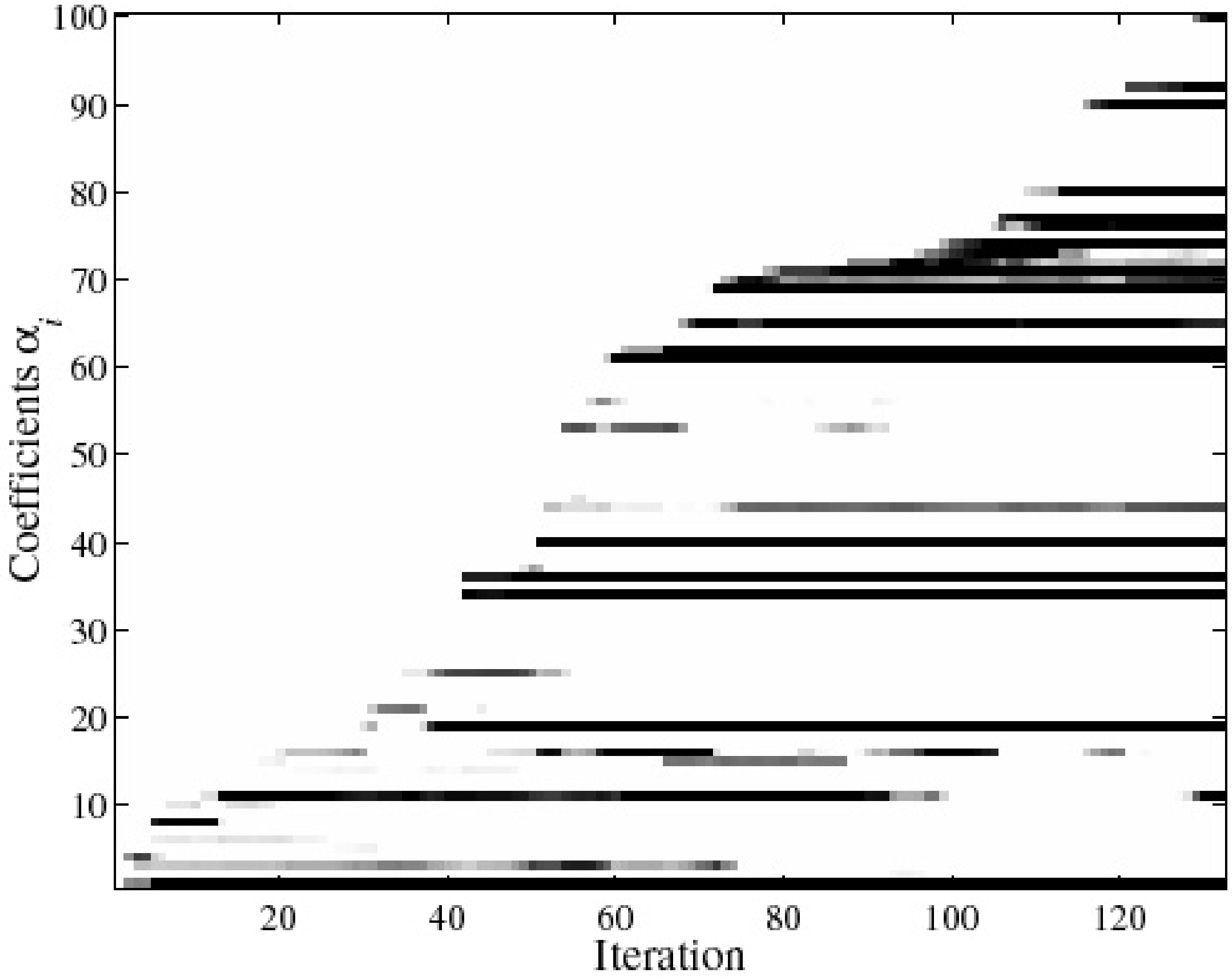


Logistic probability regression in one dimension, for the same Gaussian kernel. A less accurate, 6-kernel "reprimal" solution now truncates the kernel eigenspectrum to a spread of 100.

Incremental Learning

Cauwenberghs and Poggio, 2001

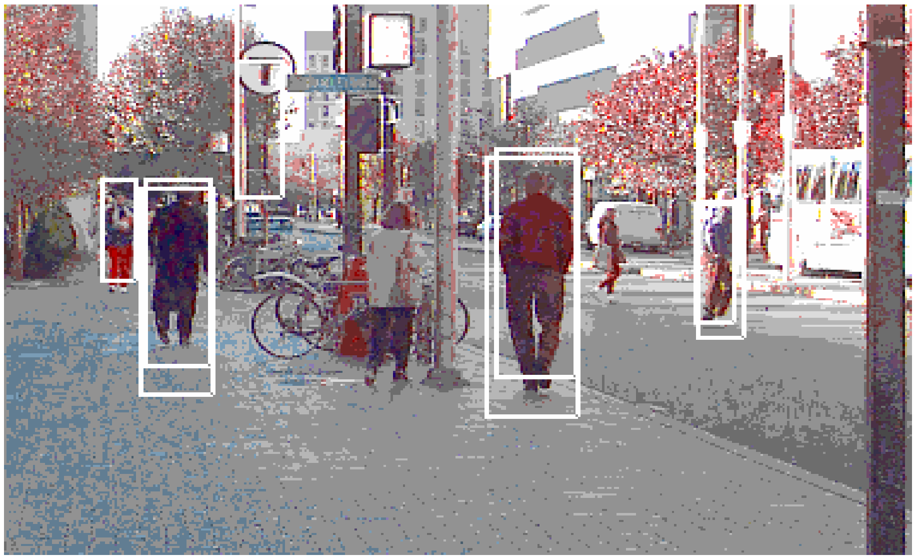
- Support Vector Machine training requires solving a linearly constrained quadratic programming problem in a number of coefficients equal to the number of data points.
- An incremental version, training one data point at at time, is obtained by solving the QP problem in recursive fashion, without the need for QP steps or inverting a matrix.
  - *On-line learning is thus feasible, with no more than L^2 state variables, where L is the number of margin (support) vectors.*
  - *Training time scales approximately linearly with data size for large, low-dimensional data sets.*
- Decremental learning (adiabatic reversal of incremental learning) allows to directly evaluate the exact leave-one-out generalization performance on the training data.
- When the incremental inverse jacobian is (near) ill-conditioned, a direct L1-norm minimization of the α coefficients yields an optimally sparse solution.

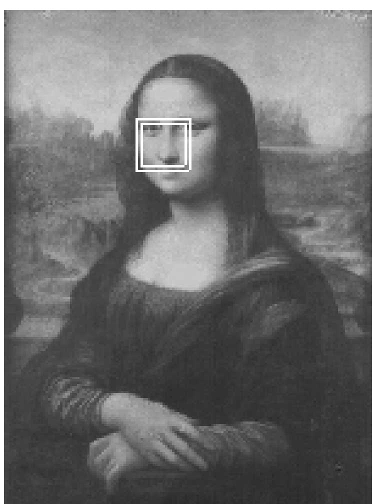


Trajectory of coefficients a as a function of time during incremental learning, for 100 data points in the non-separable case, and using a Gaussian kernel.

Trainable Modular Vision Systems: The SVM Approach

Papageorgiou, Oren, Osuna and Poggio, 1998





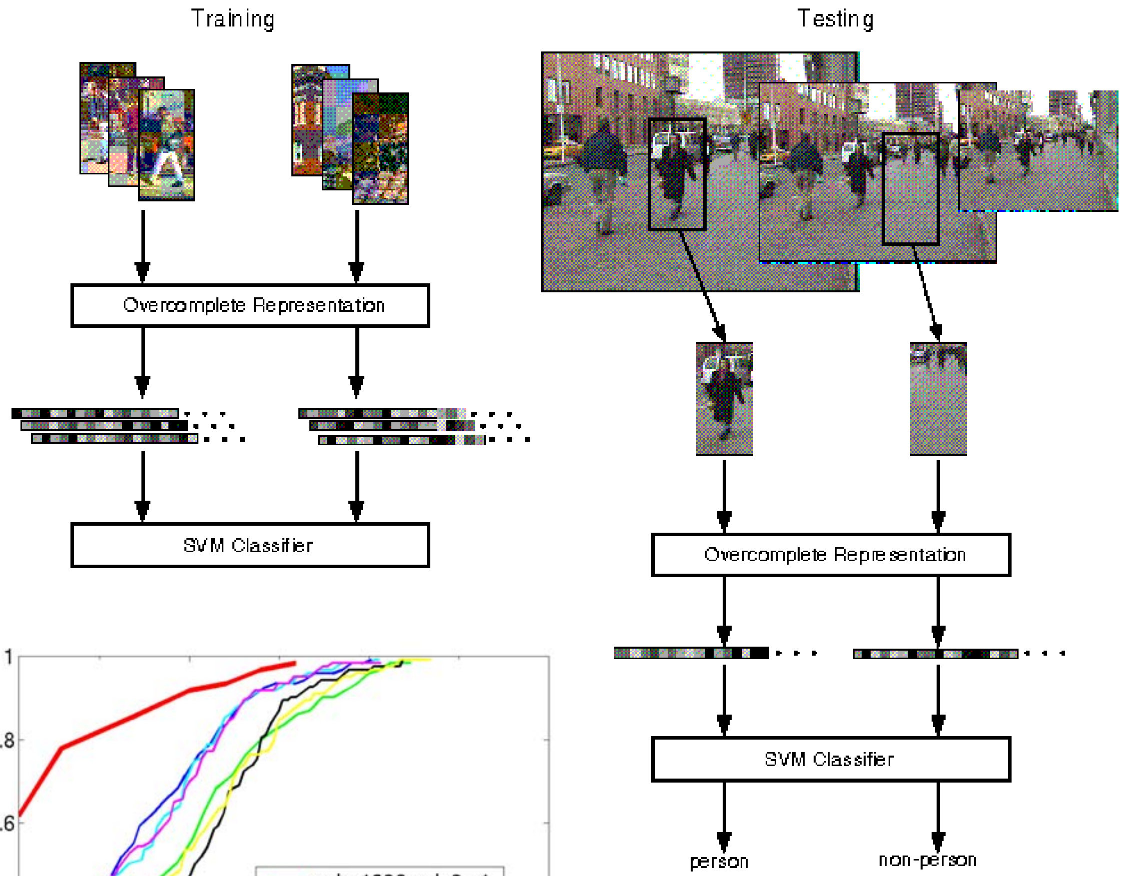
SVM classification for pedestrian and face object detection

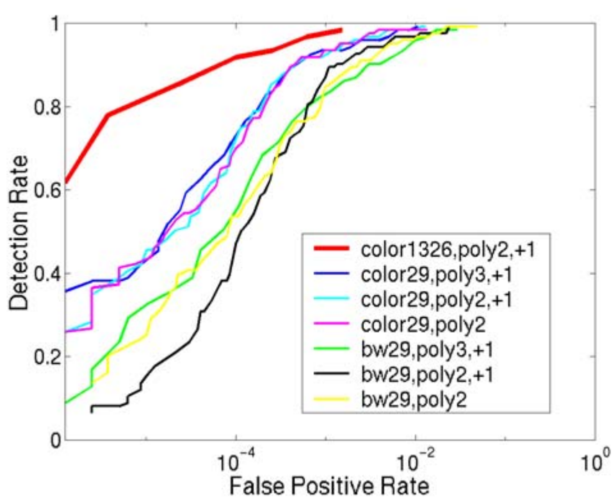
- Strong mathematical foundations in *Statistical Learning Theory* (Vapnik, 1995)

- The training process selects a small fraction of prototype *support vectors* from the data set, located at the *margin* on both sides of the classification boundary (e.g., barely faces vs. barely non-faces)

Trainable Modular Vision Systems: The SVM Approach

Papageorgiou, Oren, Osuna and Poggio, 1998



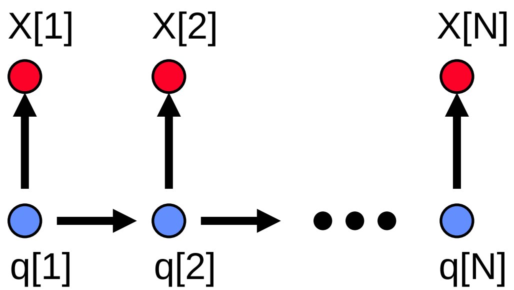


ROC curve for various image representations and dimensions

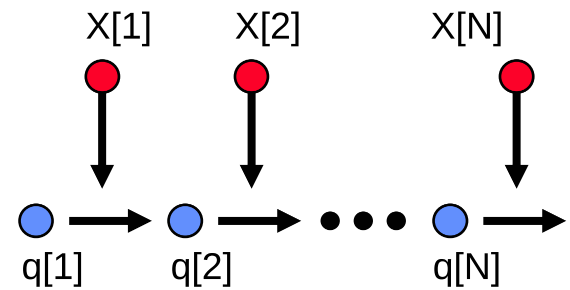
– The number of support vectors and their dimensions, in relation to the available data, determine the generalization performance

– Both training and run-time performance are severely limited by the computational complexity of evaluating kernel functions

Dynamic Pattern Recognition



Generative: ***HMM***

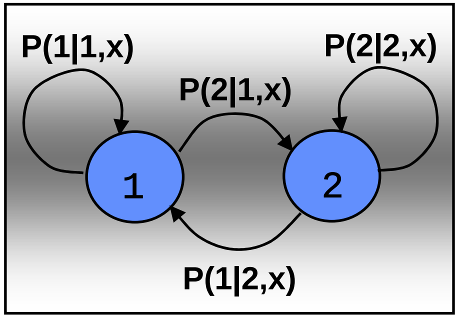


Discriminative: ***MEMM, CRF, FDKM***

Density models (such as mixtures of Gaussians) require vast amounts of training data to reliably estimate parameters.

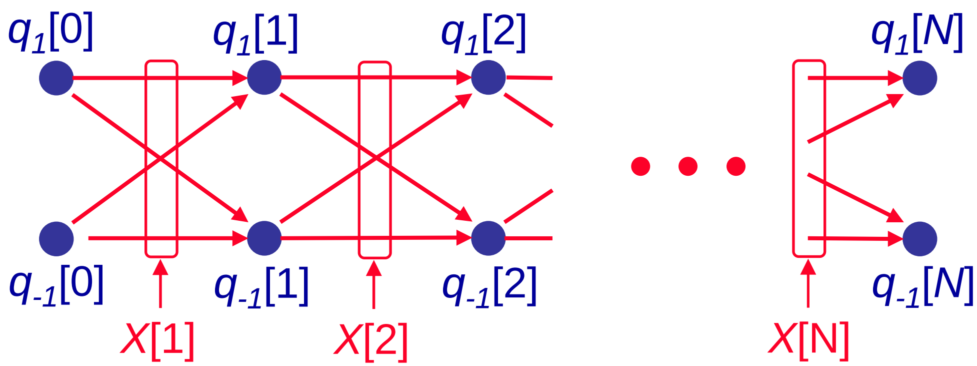
Transition-based speech recognition
(H. Bourlard and N. Morgan, 1994)

MAP forward decoding



Transition probabilities generated by large margin probability regressor

MAP Decoding Formulation



- States $$q_k[n]$$
- Posterior Probabilities (Forward) $$\alpha_k[n] = P(q_k[n] \mid \mathbf{X}[n], W)$$ $$\mathbf{X}[n] = (X[1], \ldots X[n])$$
- Transition Probabilities $$\boxed{P_{jk}[n] = P(q_k[n] \mid q_j[n-1], X[n], W)}$$ Large-Margin Probability Regression
- Forward Recursion $$\alpha_k[n] = \sum_j \alpha_j[n-1] P_{kj}[n]$$
- MAP Forward Decoding $$q^{est}[n] = \arg\max_i \alpha_i[n]$$

FDKM Training Formulation

Chakrabartty and Cauwenberghs, 2002

– Large-margin training of state transition probabilities, using regularized cross-entropy on the posterior state probabilities:

$$H = C\sum_{n=0}^{N-1}\sum_{i=0}^{S-1} y_i[n]\log\alpha_i[n] - \tfrac{1}{2}\sum_{j=0}^{S-1}\sum_{i=0}^{S-1}|w_{ij}|^2$$

– Forward Decoding Kernel Machines (FDKM) decompose an upper bound of the regularized cross-entropy (by expressing concavity of the logarithm in forward recursion on the previous state):

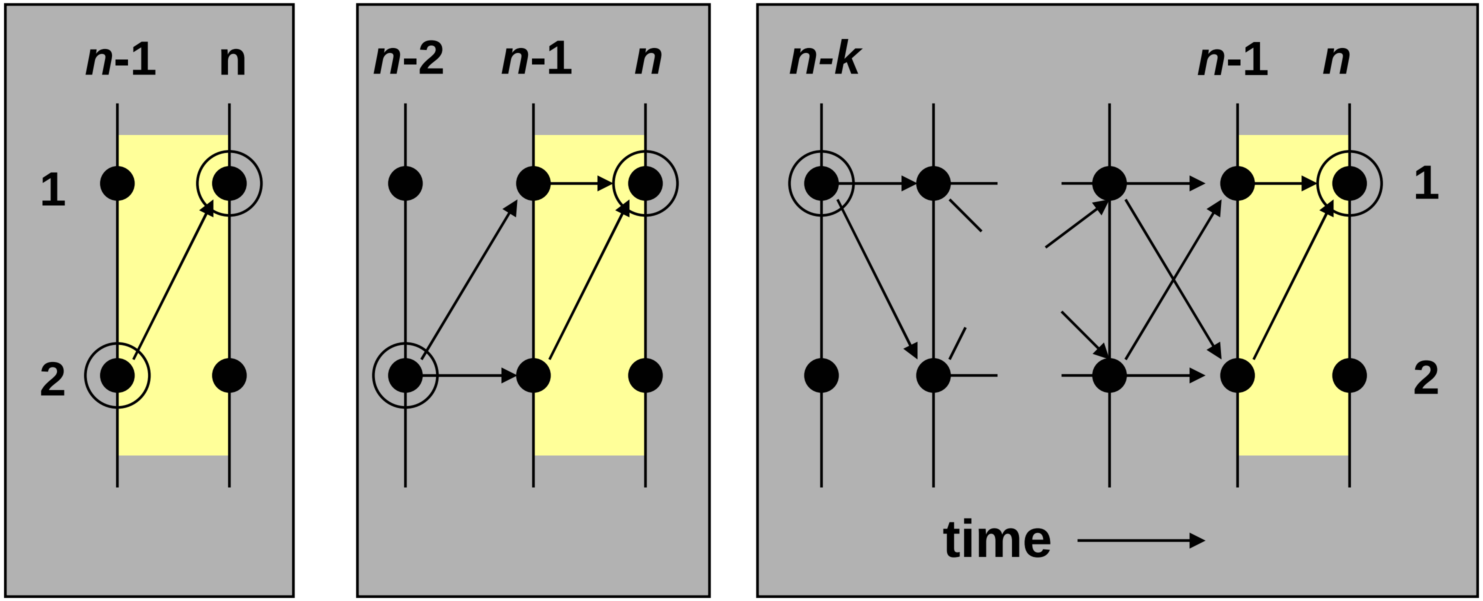
$$H \geq \sum_{j=0}^{S-1} H_j$$

which then reduces to S independent regressions of conditional probabilities, one for each outgoing state:

$$H_j = \sum_{n=0}^{N-1} C_j[n]\sum_{i=0}^{S-1} y_i[n]\log P_{ij}[n] - \tfrac{1}{2}\sum_{i=0}^{S-1}|w_{ij}|^2$$

$$C_j[n] = C\alpha_j[n-1]$$

Recursive MAP Training of FDKM

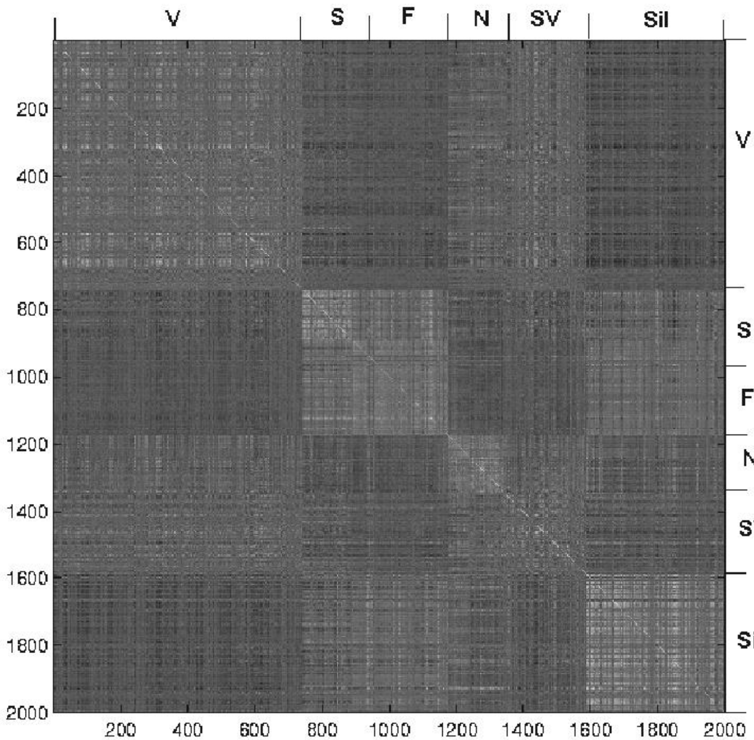


Epoch 1 Epoch 2 Epoch K

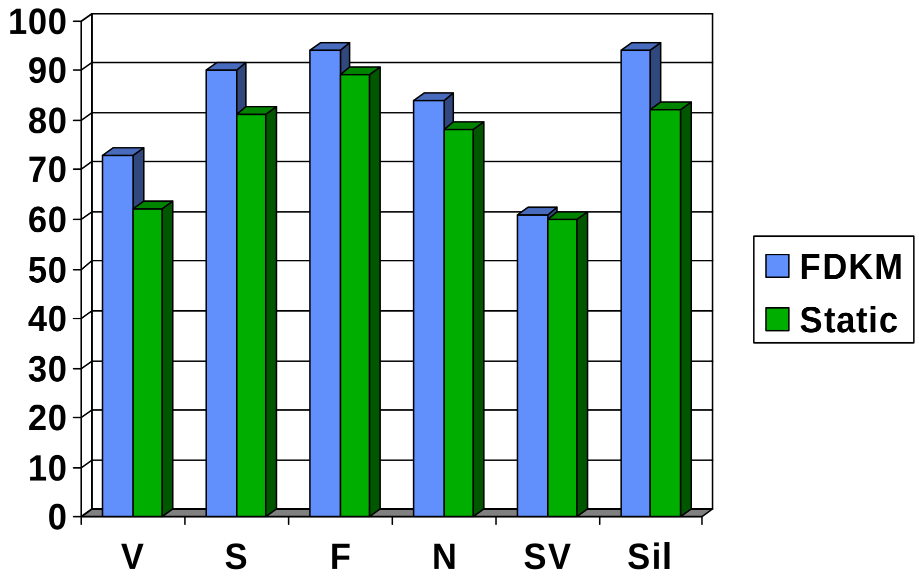
Phonetic Experiments (TIMIT)

Chakrabartty and Cauwenberghs, 2002

Features: cepstral coefficients for *Vowels*, *Stops, Fricatives*, *Semi-Vowels,* and *Silence*



Kernel Map



Recognition Rate

Conclusions

- **Kernel learning machines combine the universality of neural computation with mathematical foundations of statistical learning theory.**
  - Unified framework covers classification, regression, and probability estimation.
  - Incremental sparse learning reduces complexity of implementation and supports on-line learning.
- **Forward decoding kernel machines and GiniSVM probability regression combine the advantages of large-margin classification and Hidden Markov Models.**
  - Adaptive MAP sequence estimation in speech recognition and communication
  - EM-like recursive training fills in noisy and missing training labels.
- **Parallel charge-mode VLSI technology offers efficient implementation of high-dimensional kernel machines.**
  - Computational throughput is a factor 100-10,000 higher than presently available from a high-end workstation or DSP.
- **Applications include real-time vision and speech recognition.**

References

http://www.kernel-machines.org

Books:

[1] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd Ed., Springer, 2000.

[2] B. Schölkopf, C.J.C. Burges and A.J. Smola, Eds., *Advances in Kernel Methods*, Cambridge MA: MIT Press, 1999.

[3] A.J. Smola, P.L. Bartlett, B. Schölkopf and D. Schuurmans, Eds., *Advances in Large Margin Classifiers*, Cambridge MA: MIT Press, 2000.

[4] M. Anthony and P.L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, 1999.

[5] G. Wahba, *Spline Models for Observational Data*, Series in Applied Mathematics, vol. **59**, SIAM, Philadelphia, 1990.

Articles:

[6] M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. **25**, pp. 821-837, 1964.

[7] P. Bartlett and J. Shawe-Taylor, "Generalization performance of support vector machines and other pattern classifiers," in Schölkopf, Burges, Smola, Eds., *Advances in Kernel Methods — Support Vector Learning*, Cambridge MA: MIT Press, pp. 43-54, 1999.

[8] B.E. Boser, I.M. Guyon and V.N. Vapnik, "A training algorithm for optimal margin classifiers," *Proc. 5th ACM Workshop on Computational Learning Theory (COLT)*, ACM Press, pp. 144-152, July 1992.

[9] C.J.C. Burges and B. Schölkopf, "Improving the accuracy and speed of support vector learning machines," *Adv. Neural Information Processing Systems (NIPS*96)*, Cambridge MA: MIT Press, vol. **9**, pp. 375-381, 1997.

[10] G. Cauwenberghs and V. Pedroni, "A low-power CMOS analog vector quantizer," *IEEE Journal of Solid-State Circuits*, vol. **32** (8), pp. 1278-1283, 1997.

[11] G. Cauwenberghs and T. Poggio, "Incremental and decrementral support vector machine learning," *Adv. Neural Information Processing Systems (NIPS*2000)*, Cambridge, MA: MIT Press, vol. **13**, 2001.

[12] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. **20**, pp. 273-297, 1995.

[13] T. Evgeniou, M. Pontil and T. Poggio, "Regularization networks and support vector machines," *Adv. Computational Mathematics (ACM)*, vol. **13**, pp. 1-50, 2000.

[14] M. Girolami, "Mercer kernel based clustering in feature space," *IEEE Trans. Neural Networks*, 2001.

[15] F. Girosi, M. Jones and T. Poggio, "Regularization theory and neural network architectures," *Neural Computation*, vol. **7**, pp 219-269, 1995.

[16] F. Girosi, "An equivalence between sparse approximation and Support Vector Machines," *Neural Computation*, vol. **10** (6), pp. 1455-1480, 1998.

[17] R. Genov and G. Cauwenberghs, "Charge-Mode Parallel Architecture for Matrix-Vector Multiplication," submitted to *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Processing*, 2001.

[18] T.S. Jaakkola and D. Haussler, "Probabilistic kernel regression models," *Proc. 1999 Conf. on AI and Statistics*, 1999.

[19] T.S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," *Adv. Neural Information Processing Systems (NIPS*98)*, vol. **11**, Cambridge MA: MIT Press, 1999.

[20] D.J.C. MacKay, "Introduction to Gaussian Processes," Cambridge University, http://wol.ra.phy.cam.ac.uk/mackay/, 1998.

[21] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philos. Trans. Royal Society London*, A, vol. **209**, pp. 415-446, 1909.

[22] S. Mika, G. R¨ atsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," *Neural Networks for Signal Processing IX*, IEEE, pp 41-48, 1999.

[23] M. Opper and O. Winther, "Gaussian processes and SVM: mean field and leave-one-out," in Smola, Bartlett, Schölkopf and Schuurmans, Eds., *Advances in Large Margin Classifiers*, Cambridge MA: MIT Press, pp. 311-326, 2000.

[24] E. Osuna and F. Girosi, "Reducing the run-time complexity in support vector regression," in Schölkopf, Burges, Smola, Eds., *Advances in Kernel Methods — Support Vector Learning*, Cambridge MA: MIT Press, pp. 271-284, 1999.

[25] C.P. Papageorgiou, M. Oren and T. Poggio, "A general framework for object detection," in *Proceedings of International Conference on Computer Vision*, 1998.

[26] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proc. IEEE*, vol. **78** (9), 1990.

[27] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. **10**, pp. 1299-1319, 1998.

[28] A.J. Smola and B. Schölkopf, "On a kernel-based method for pattern recognition, regression, approximation and operator inversion," *Algorithmica*, vol. **22**, pp. 211-231, 1998.

[29] V. Vapnik and A. Lerner, "Pattern recognition using generalized portrait method," *Automation and Remote Control*, vol. **24**, 1963.

[30] V. Vapnik and A. Chervonenkis, "Theory of Pattern Recognition," *Nauk*a, Moscow, 1974.

[31] G.S. Kimeldorf and G. Wahba, "A correspondence between Bayesan estimation on stochastic processes and smoothing by splines," *Ann. Math. Statist*., vol. **2**, pp. 495-502, 1971.

[32] G. Wahba, "Support Vector Machines, Reproducing Kernel Hilbert Spaces and the randomized GACV," in Schölkopf, Burges, and Smola, Eds., *Advances in Kernel Methods — Support Vector Learning*, Cambridge MA, MIT Press, pp. 69-88, 1999.

References
(FDKM & *Gini*SVM)

- *Bourlard H. and Morgan, N., "**Connectionist Speech Recognition: A Hybrid Approach**", Kluwer Academic, 1994.*
- *Breiman, L. Friedman, J.H. et al. "**Classification and Regression Trees**", Wadsworth and Brooks, Pacific Grove, CA 1984.*
- *Chakrabartty, S. and Cauwenberghs, G. "**Forward Decoding Kernel Machines: A Hybrid HMM/SVM Approach to Sequence Recognition**," IEEE Int Conf. On Pattern Recognition: SVM workshop, Niagara Falls, Canada 2002.*
- *Chakrabartty, S. and Cauwenberghs, G. "**Forward Decoding Kernel-Based Phone Sequence Recognition**," Adv. Neural Information Processing Systems (http://nips.cc), Vancouver, Canada 2002.*
- *Clark, P. and Moreno M.J. "**On the Use of Support Vector Machines for Phonetic Classification**," IEEE Conf Proc, 1999.*
- *Jaakkola, T. and Haussler, D. "**Probabilistic Kernel Regression Models**," Proceedings of Seventh International Workshop on Artificial Intelligence and Statistics, 1999.*
- *Vapnik, V. **The Nature of Statistical Learning Theory**, New York: Springer-Verlag, 1995.*