Low Bias Bagged Support Vector Machines

Giorgio Valentini
Dipartimento di Scienze dell Informazione
Università degli Studi di Milano, Italy
valentini@dsi.unimi.it

Thomas G. Dietterich
Department of Computer Science
Oregon State University
Corvallis, Oregon 97331 USA
http://www.cs.orst.edu/~tgd

Two Questions:

- Can bagging help SVMs?
- If so, how should SVMs be tuned to give the best bagged performance?

The Answers

- Can bagging help SVMs?
  - Yes
- If so, how should SVMs be tuned to give the best bagged performance?
  - Tune to minimize the bias of each SVM

SVMs

minimize: $||\mathbf{w}||^2 + C \sum_i \xi_i$

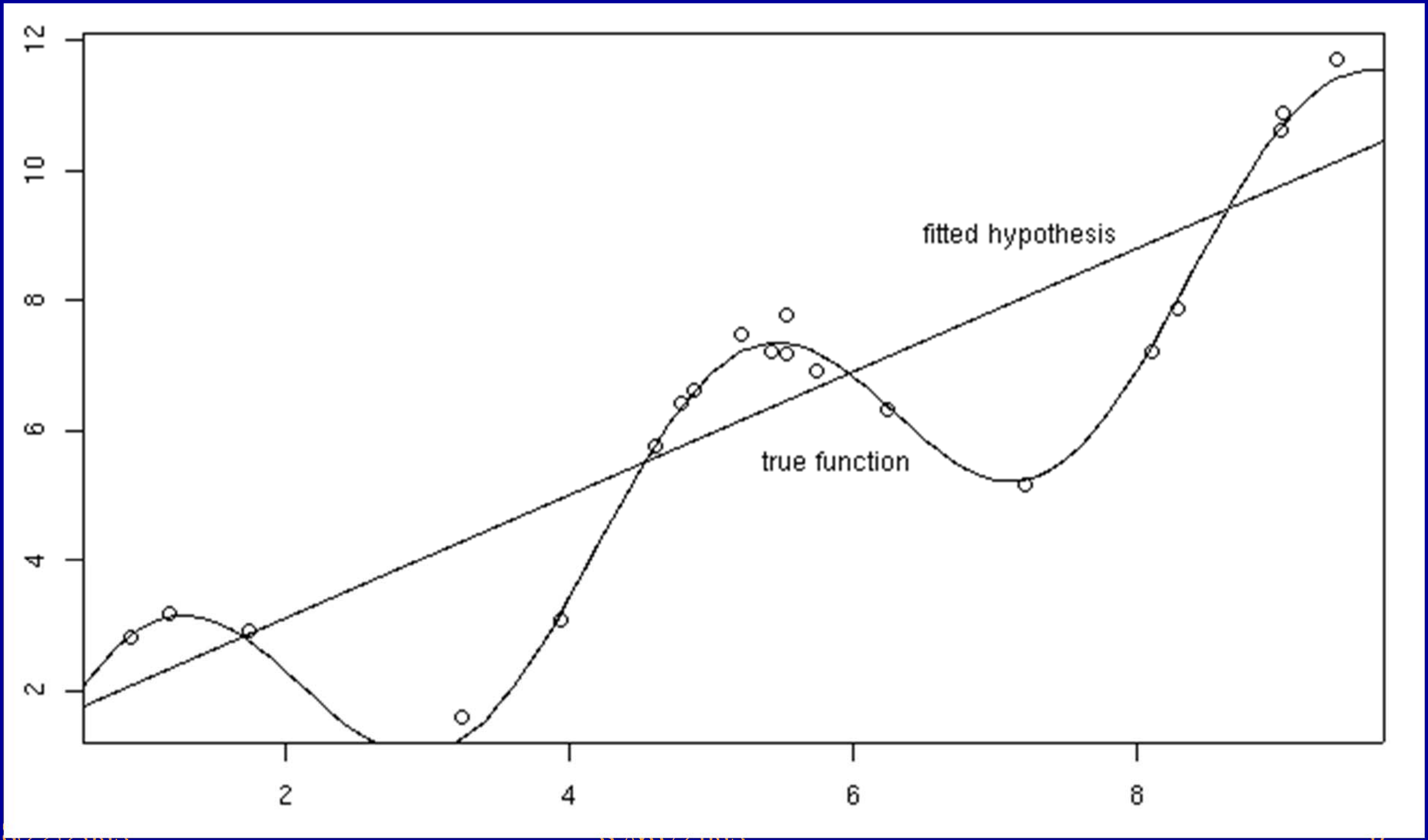
subject to: $y_i (\mathbf{w} \cdot x_i + b) + \xi_i \geq 1$

- Soft Margin Classifier
  - Maximizes VC dimension subject to soft separation of the training data
  - Dot product can be generalized using kernels $K(x_j,x_i;\sigma)$
  - Set C and σ using an internal validation set
- Excellent control of the bias/variance tradeoff: Is there any room for improvement?

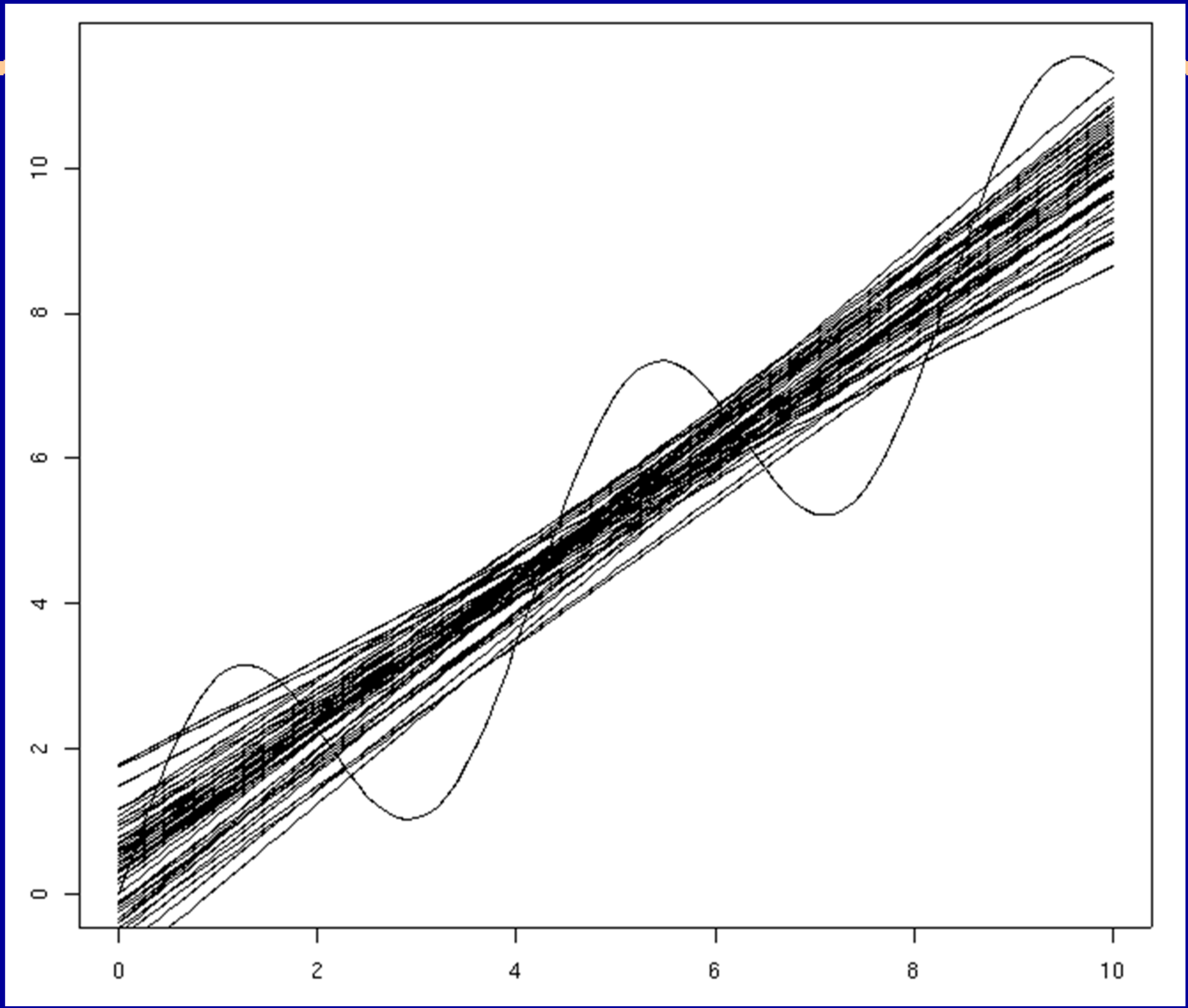
Bias/Variance Error Decomposition for Squared Loss

- For regression problems, loss is $(\hat{y} - y)^2$
  - $\text{error}^2 = \text{bias}^2 + \text{variance} + \text{noise}$
  - $E_S[(\hat{y}-y)^2] = (E_S[\hat{y}] - f(x))^2 + E_S[(\hat{y} - E_S[\hat{y}])^2] + E[(y - f(x))^2]$
- Bias: Systematic error at data point *x* averaged over all training sets S of size N
- Variance: Variation around the average
- Noise: Errors in the observed labels of *x*

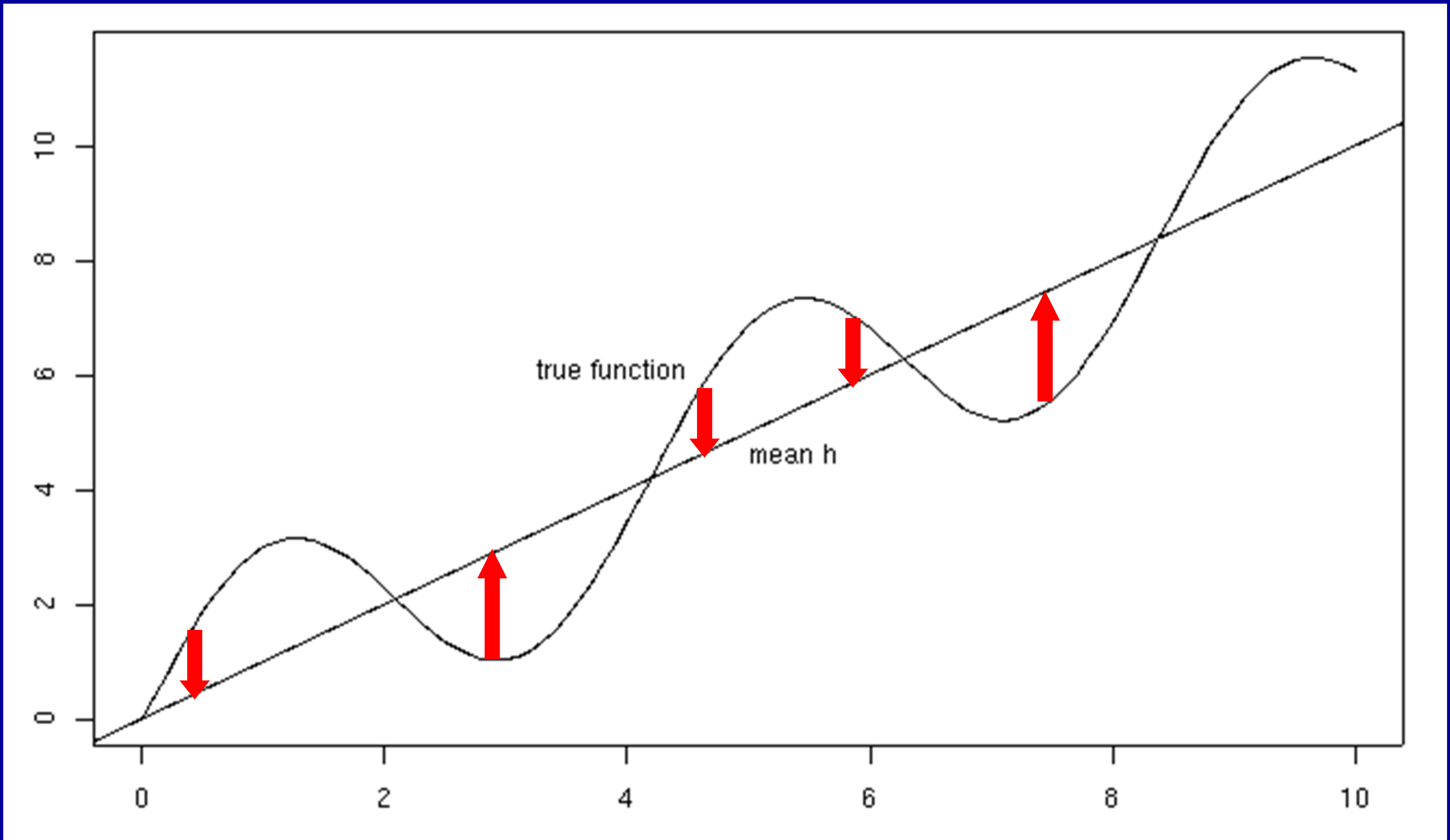
Example: 20 points
$y = x + 2\sin(1.5x) + N(0,0.2)$



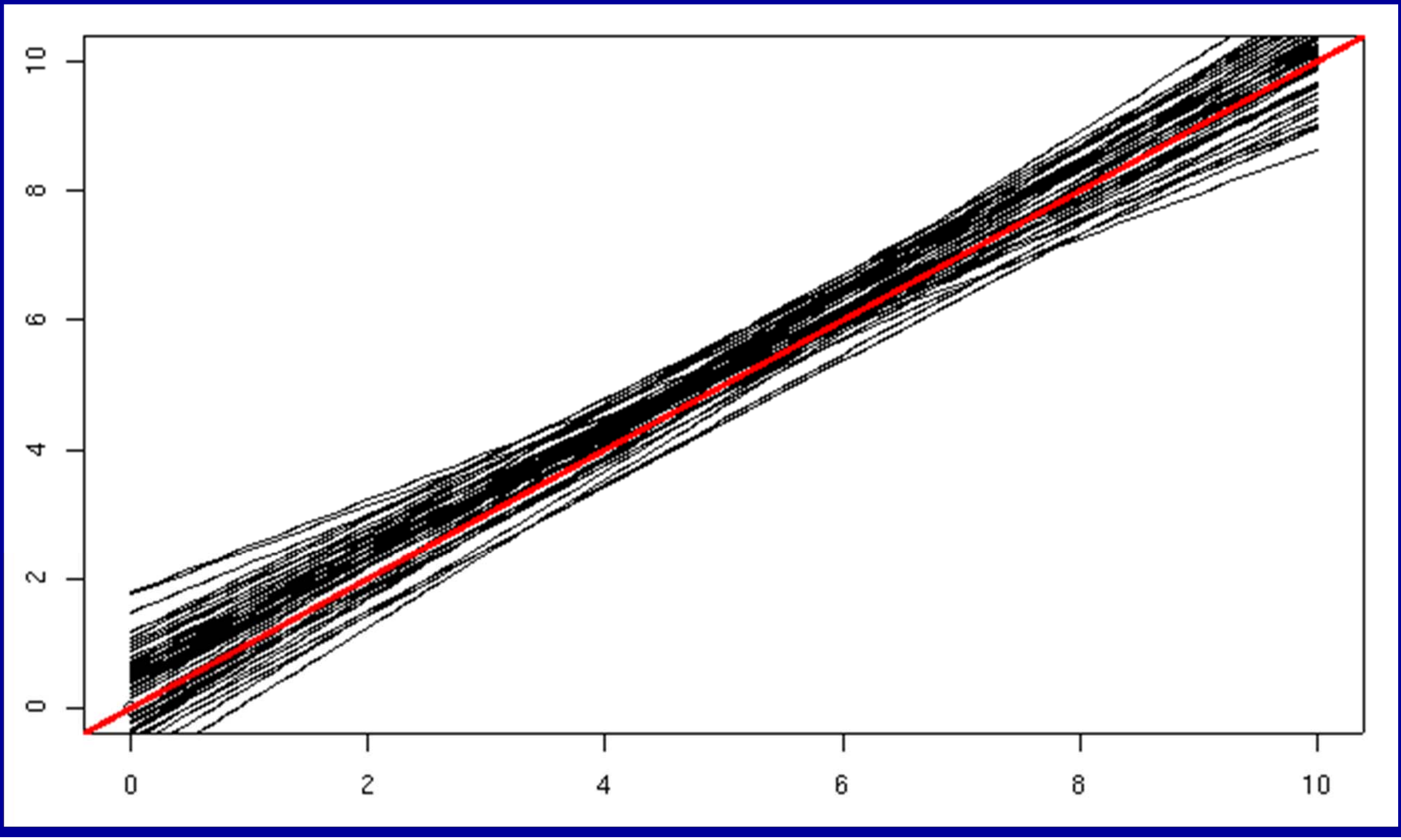
Example: 50 fits (20 examples each)



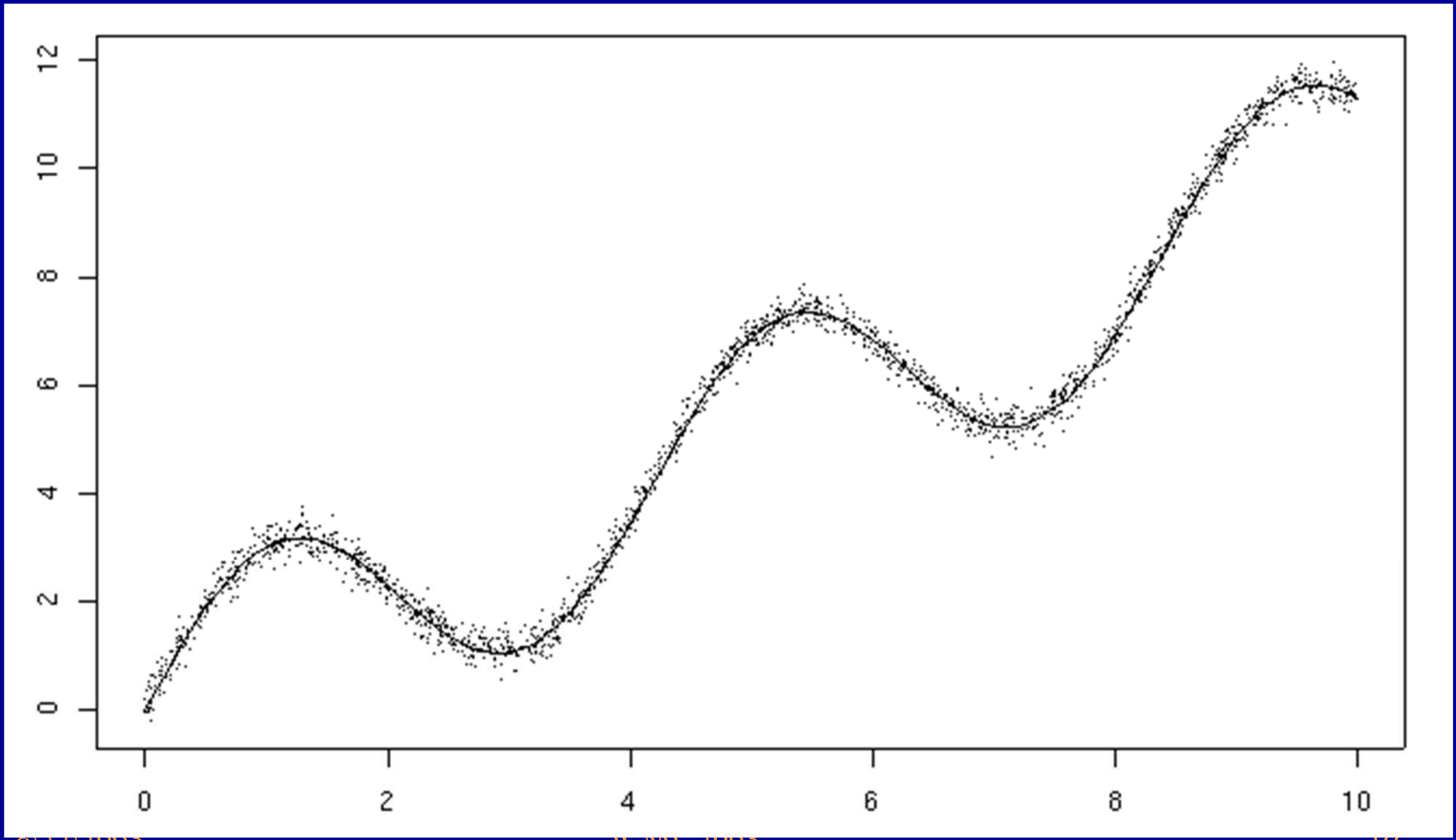
Bias



Variance



Noise



Variance Reduction and Bagging

- Bagging attempts to simulate a large number of training sets and compute the average prediction y_m of those training sets
- It then predicts y_m
- If the simulation is good enough, this eliminates all of the variance

Bias and Variance for 0/1 Loss
(Domingos, 2000)

- At each test point *x*, we have 100 estimates: $\hat{y}_1, \ldots, \hat{y}_{100} \in \{-1,+1\}$
- Main prediction: y_m = majority vote
- Bias(*x*) = 0 if y_m is correct and 1 otherwise
- Variance(*x*) = probability that $\hat{y} \neq y_m$
  - Unbiased variance $V_U(x)$: variance when Bias = 0
  - Biased variance $V_B(x)$: variance when Bias = 1
- Error rate(*x*) = Bias(*x*) + $V_U(x) - V_B(x)$
- Noise is assumed to be zero

Good Variance and Bad Variance

- Error rate(x) = Bias(x) + $V_U(x)$ – $V_B(x)$
- $V_B(x)$ is "good" variance, but only when the bias is high
- $V_U(x)$ is "bad" variance
- Bagging will reduce both types of variance. This gives good results if Bias(x) is small.
- Goal: <u>Tune classifiers to have small bias and rely on bagging to reduce variance</u>

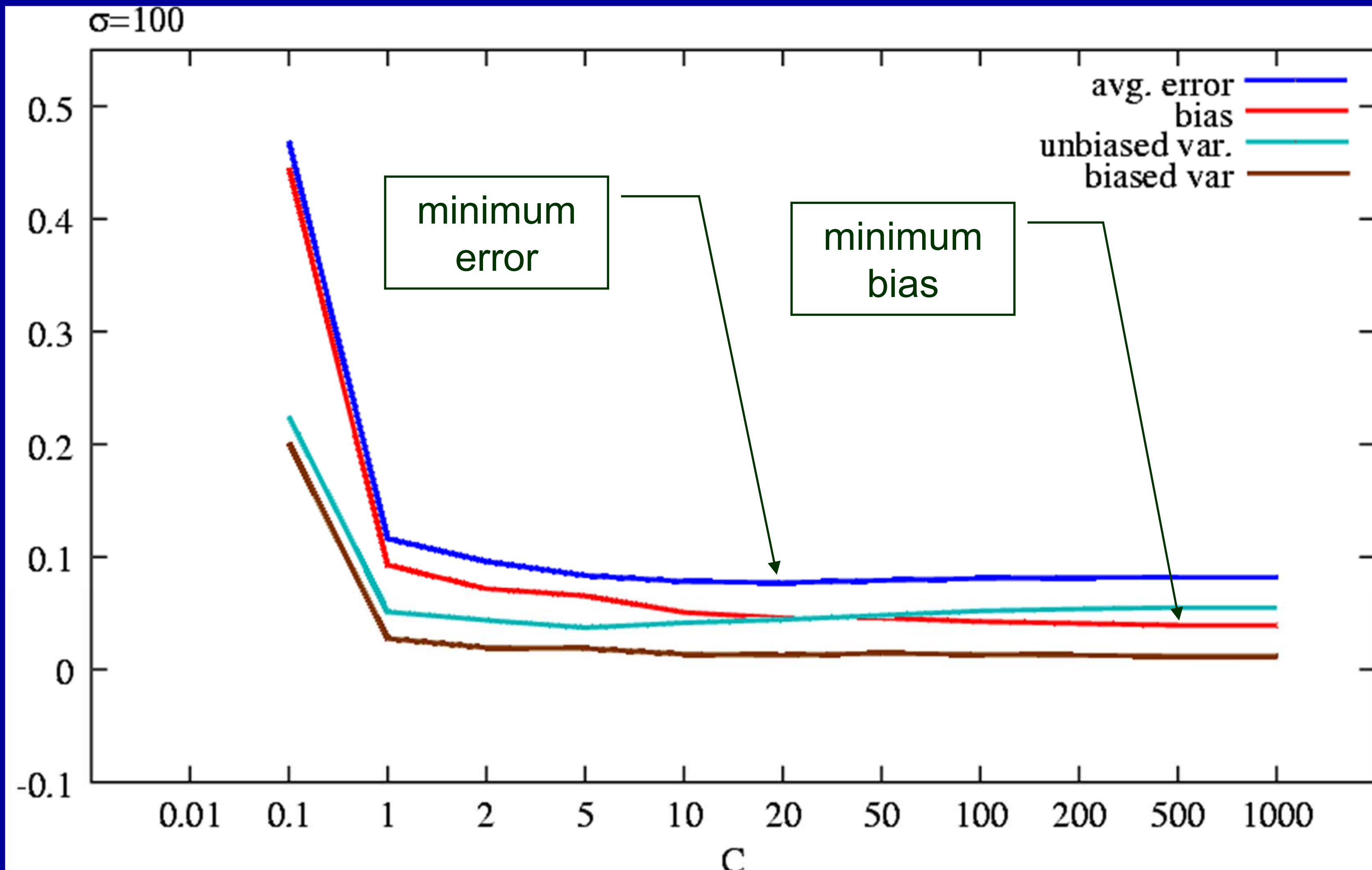
Lobag

- Given:
  - Training examples $\{(x_i,y_i)\}^N_{i=1}$
  - Learning algorithm with tuning parameters α
  - Parameter settings to try $\{\alpha_1,\alpha_2,\ldots\}$
- Do:
  - Apply internal bagging to compute out-of-bag estimates of the bias of each parameter setting. Let α^* be the setting that gives minimum bias
  - Perform bagging using α^*

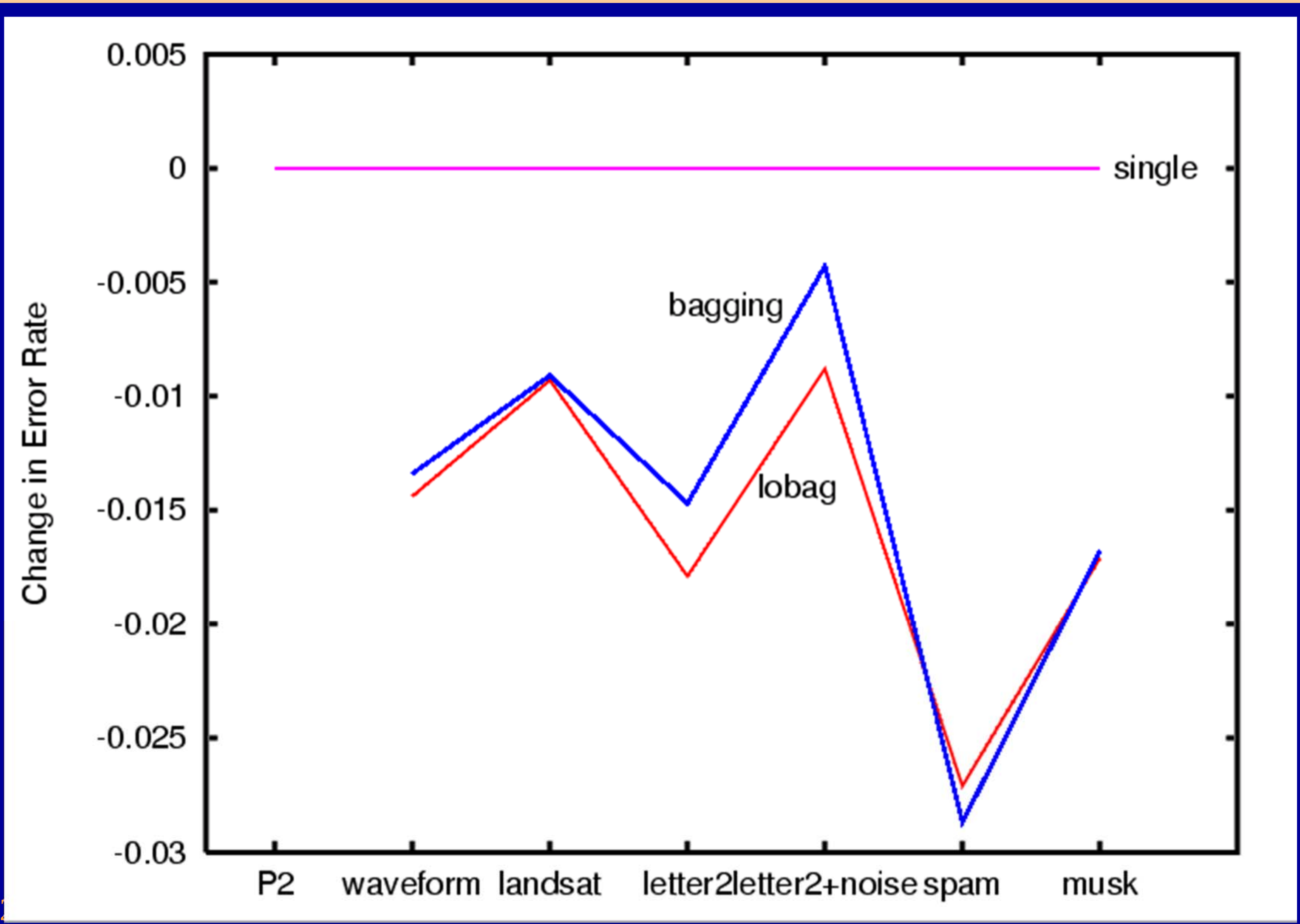
Experimental Study

- Seven data sets: P2, waveform, grey-landsat, spam, musk, letter2 (letter recognition 'B' vs 'R'), letter2+noise (20% added noise)
- Three kernels: dot product, RBF (σ = gaussian width), polynomial (σ = degree)
- Training set: 100 examples
- Bias and variance estimated on test set from 100 replicates

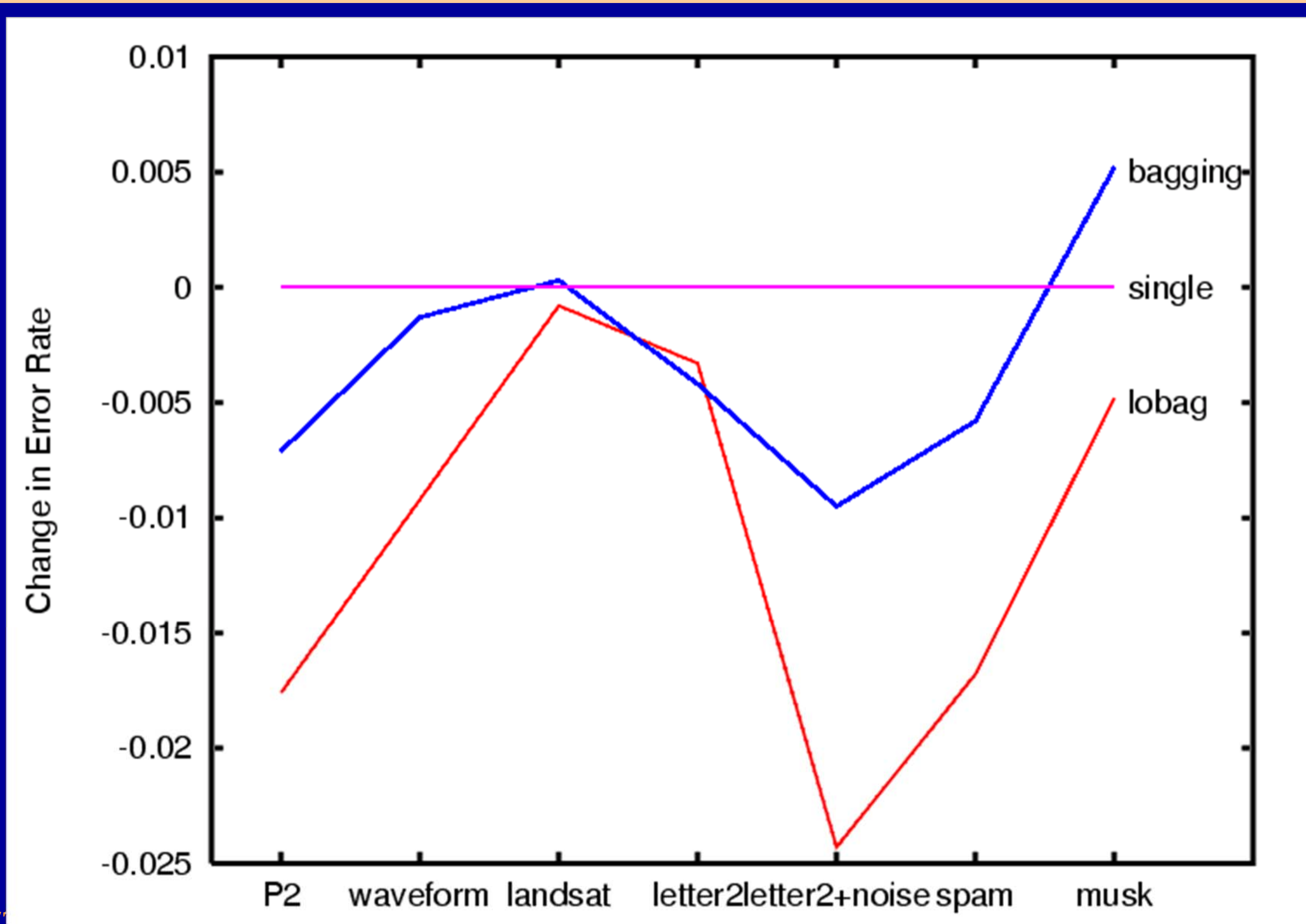
Example: Letter2, RBF kernel, σ = 100



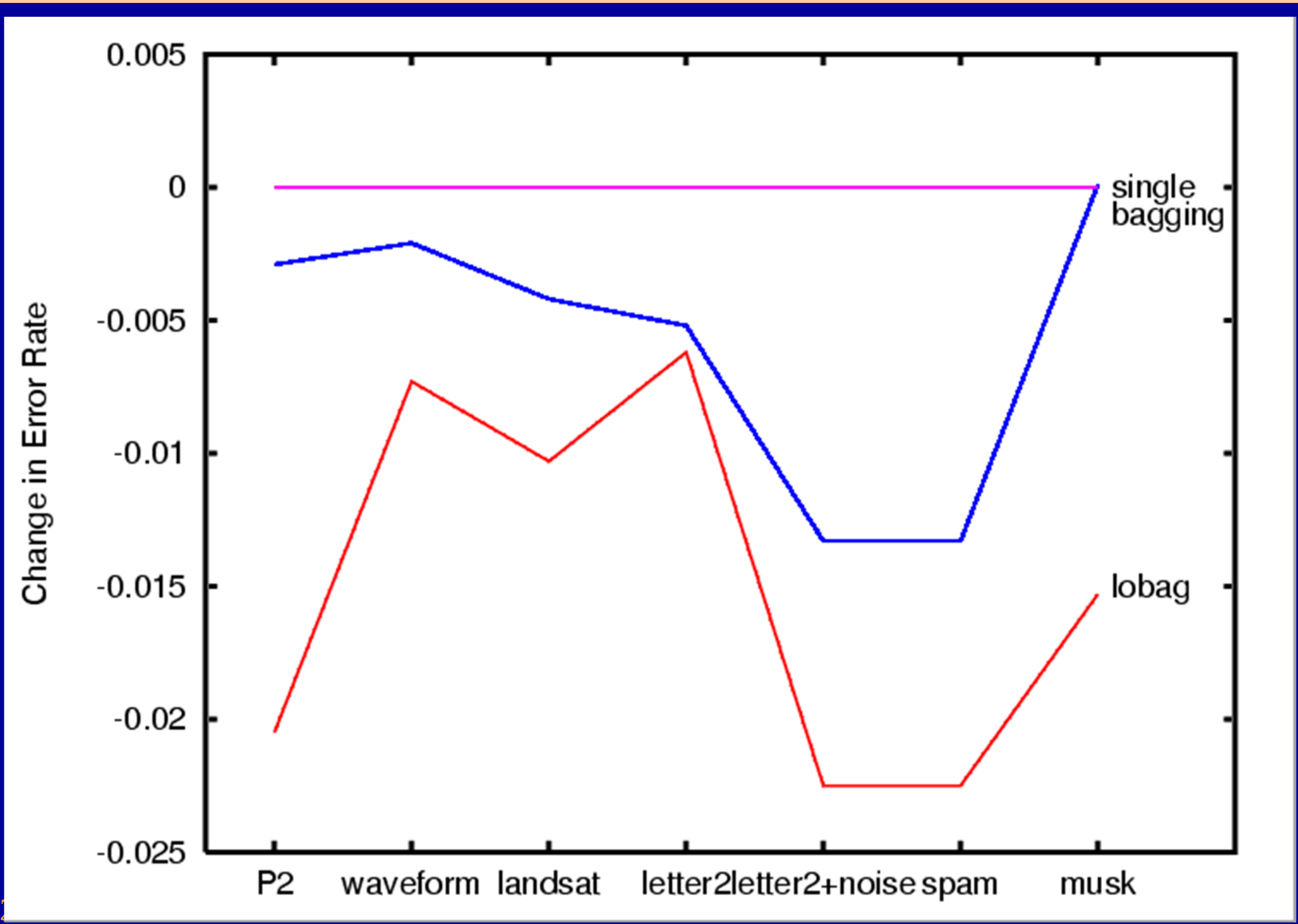
Results: Dot Product Kernel



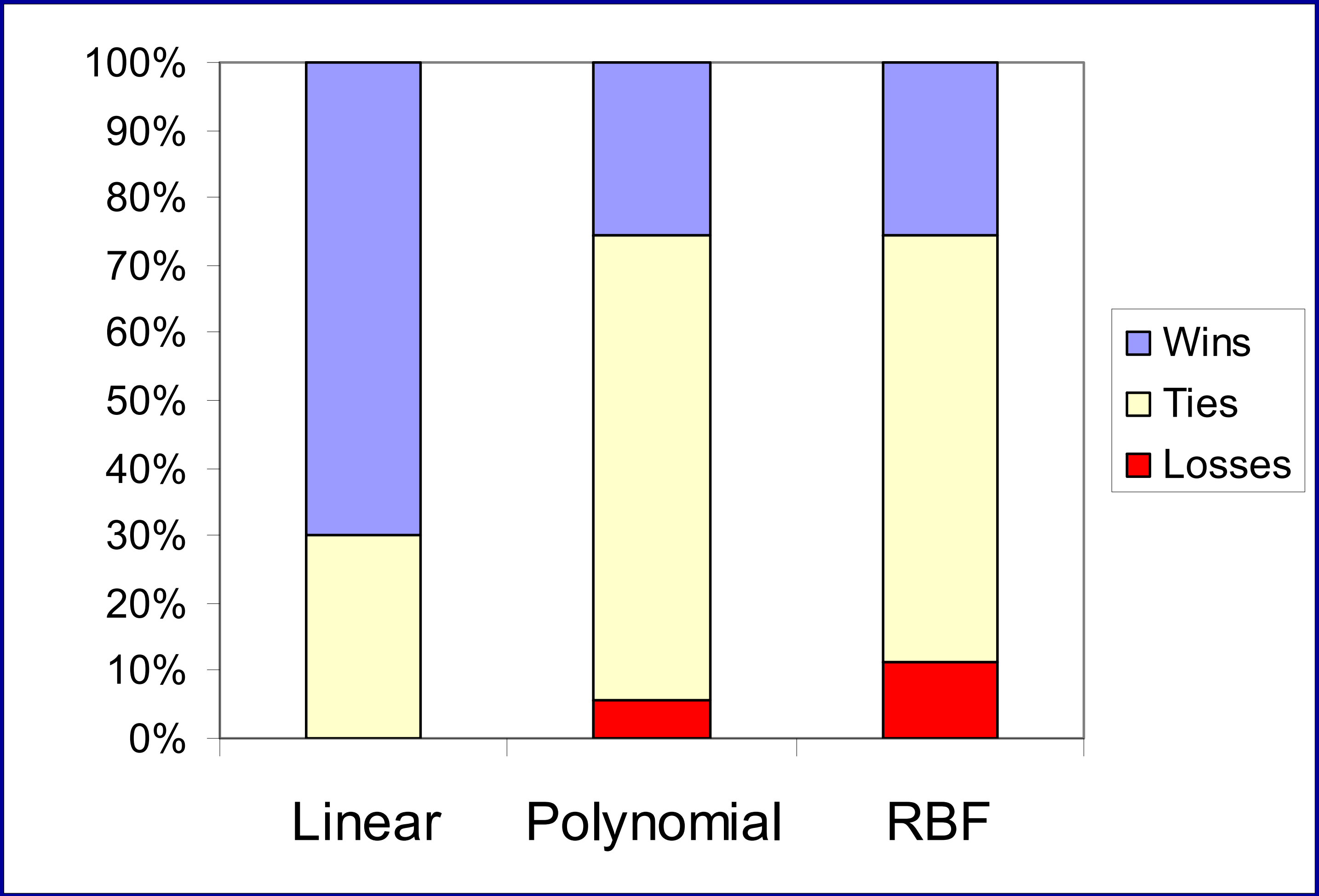
Results (2): Gaussian Kernel



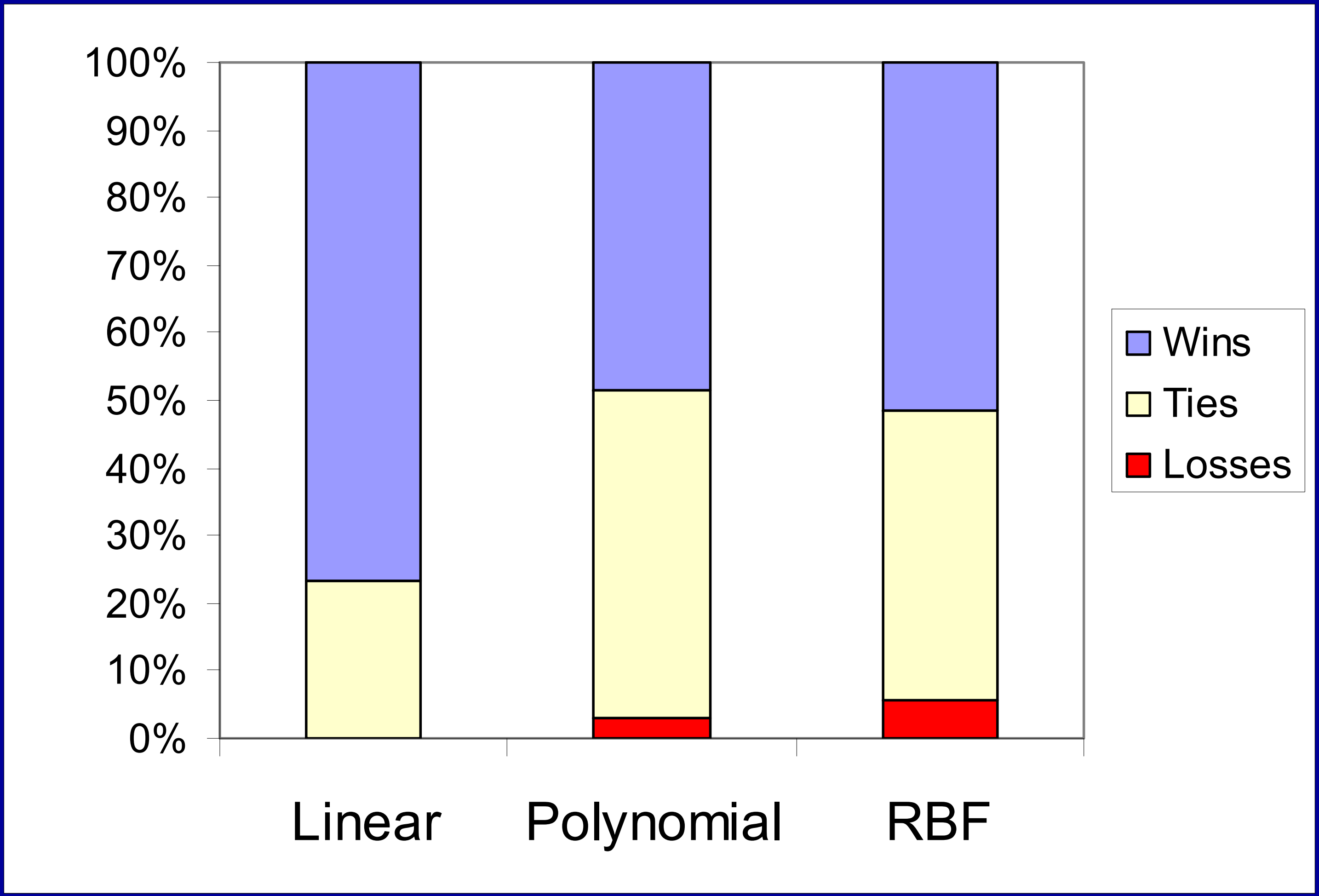
Results (3): Polynomial Kernel



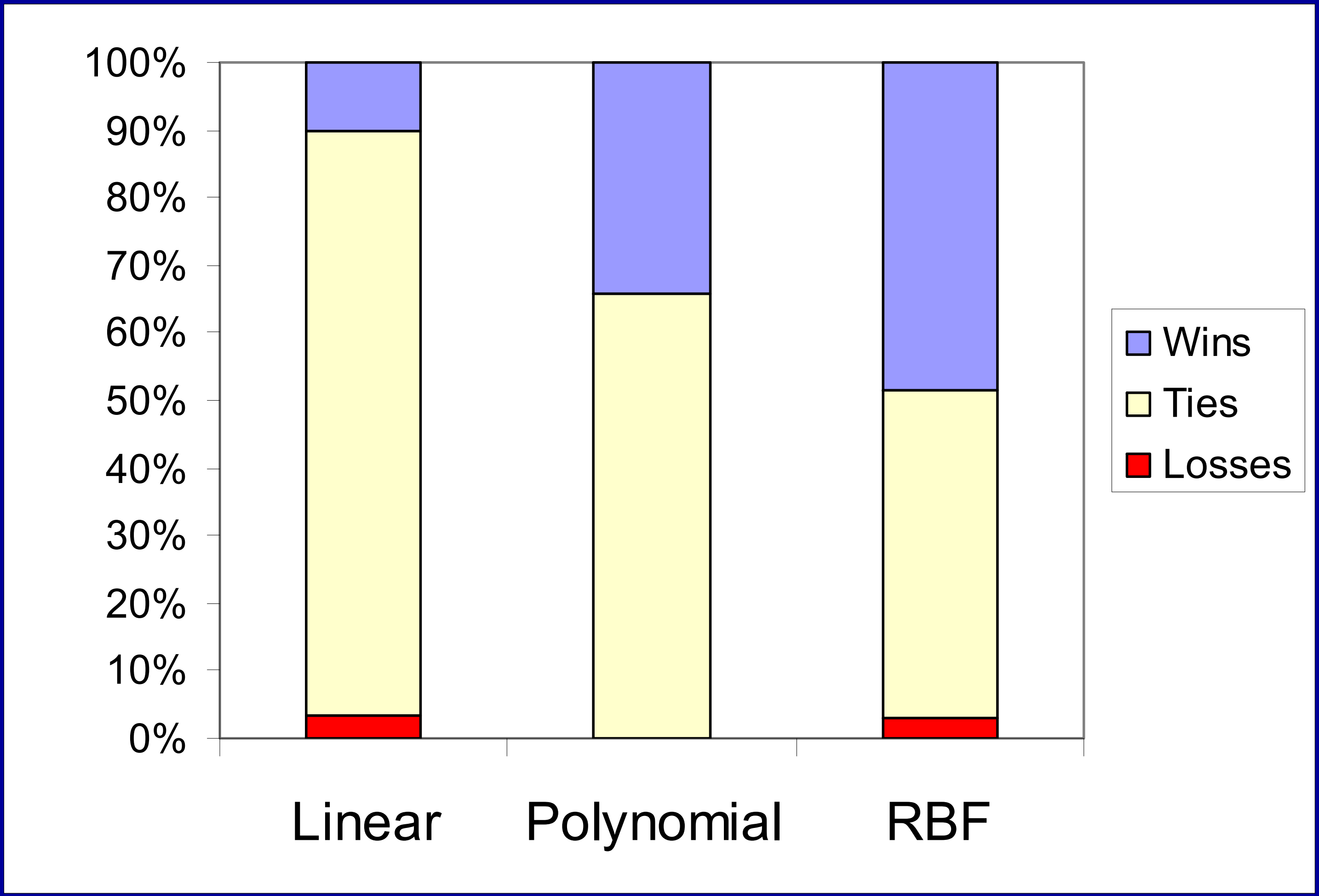
McNemar's Tests: Bagging versus Single SVM



McNemar's Test: Lobag versus Single SVM



McNemar's Test: Lobag versus Bagging



Results: McNemar's Test (wins – ties – losses)

Kernel	Lobag vs Bagging	Lobag vs Single	Bagging vs Single
Linear	3 – 26 – 1	23 – 7 – 0	21 – 9 – 0
Polynomial	12 – 23 – 0	17 – 17 – 1	9 – 24 – 2
Gaussian	17 – 17 – 1	18 – 15 – 2	9 – 22 – 4
Total	32 – 66 – 2	58 – 39 – 3	39 – 55 – 6

Discussion

- For small training sets
  - Bagging can improve SVM error rates, especially for linear kernels
  - Lobag is at least as good as bagging and often better
- Consistent with previous experience
  - Bagging works better with unpruned trees
  - Bagging works better with neural networks that are trained longer or with less weight decay

Conclusions

- Lobag is recommended for SVM problems with high variance (small training sets, high noise, many features)
- Small added cost:
  - SVMs require internal validation to set C and σ
  - Lobag requires internal bagging to estimate bias for each setting of C and σ
- Future research:
  - Smart search for low-bias settings of C and
  - Experiments with larger training sets