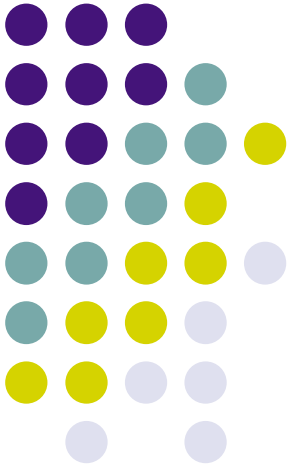
What is The Optimal Number of Features?

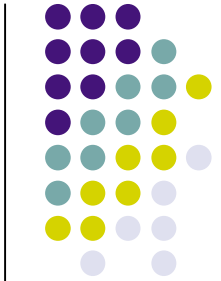
A learning theoretic Perspective

Amir Navot

Joint work with Ran Gilad-Bachrach, Yiftah Navot and Naftali Tishby



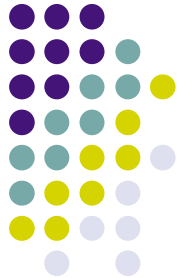
What is Feature Selection?



	Feature 1	Feature 2	Feature 3	Feature 4	...
Instance 1	1	5	3	5	...
Instance 2	1	5	8	98	...
Instance 3	10	77	7	4	...
Instance 4	0	45	59	3	...
⋮	⋮	⋮	⋮	⋮	⋱

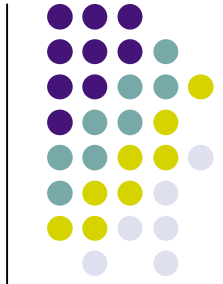
- **Feature selection: select a "good" small subset of features (out of the given set)**
- "Good" subset: enables to build good classifiers
- Feature selection is a special form of dimensionality reduction

Reasons to do Feature Selection



- Reduces computational complexity
- Saves the cost of measuring extra features
- The selected features can provide insights about the nature of the problem
- **Improves accuracy**

The Questions



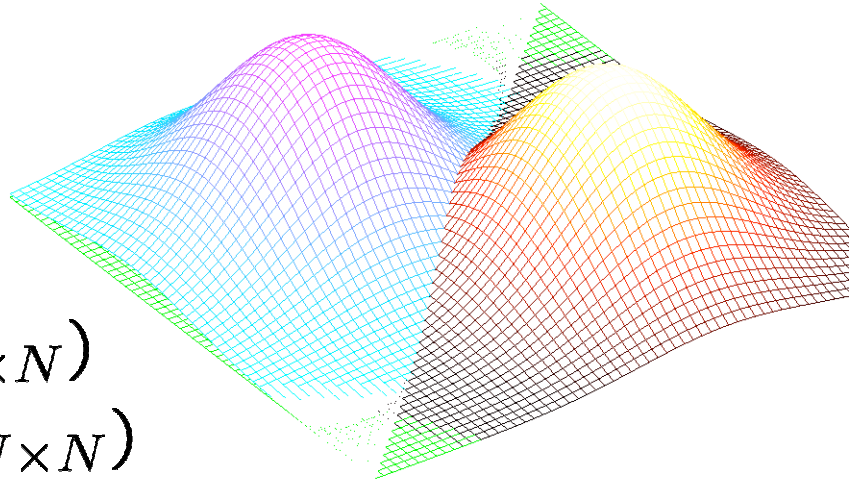
- Under which conditions can feature selection improve classification accuracy?
- What is the optimal number of features?
- How does this number depend on the training set size?

We discuss these questions by analyzing one simple setting

Two Gaussians - Problem Setting

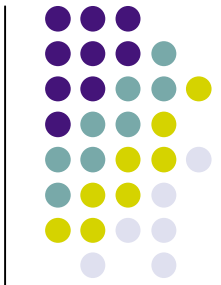
- Binary classification task

$$\mu \in R^N, y \in \{+1, -1\}$$
$$(x|y = +1) \sim \mathcal{N}(\mu, \Sigma = I_{N\times N})$$
$$(x|y = -1) \sim \mathcal{N}(-\mu, \Sigma = I_{N\times N})$$



- The coordinates are the **features**
- Optimal classifier: $h(x) = sign(\mu \cdot x)$
  - given features subset F : $h|_F(x) = sign(\mu|_F \cdot x|_F)$
- If μ is known: using all the features is optimal

Problem Setting – Cont.



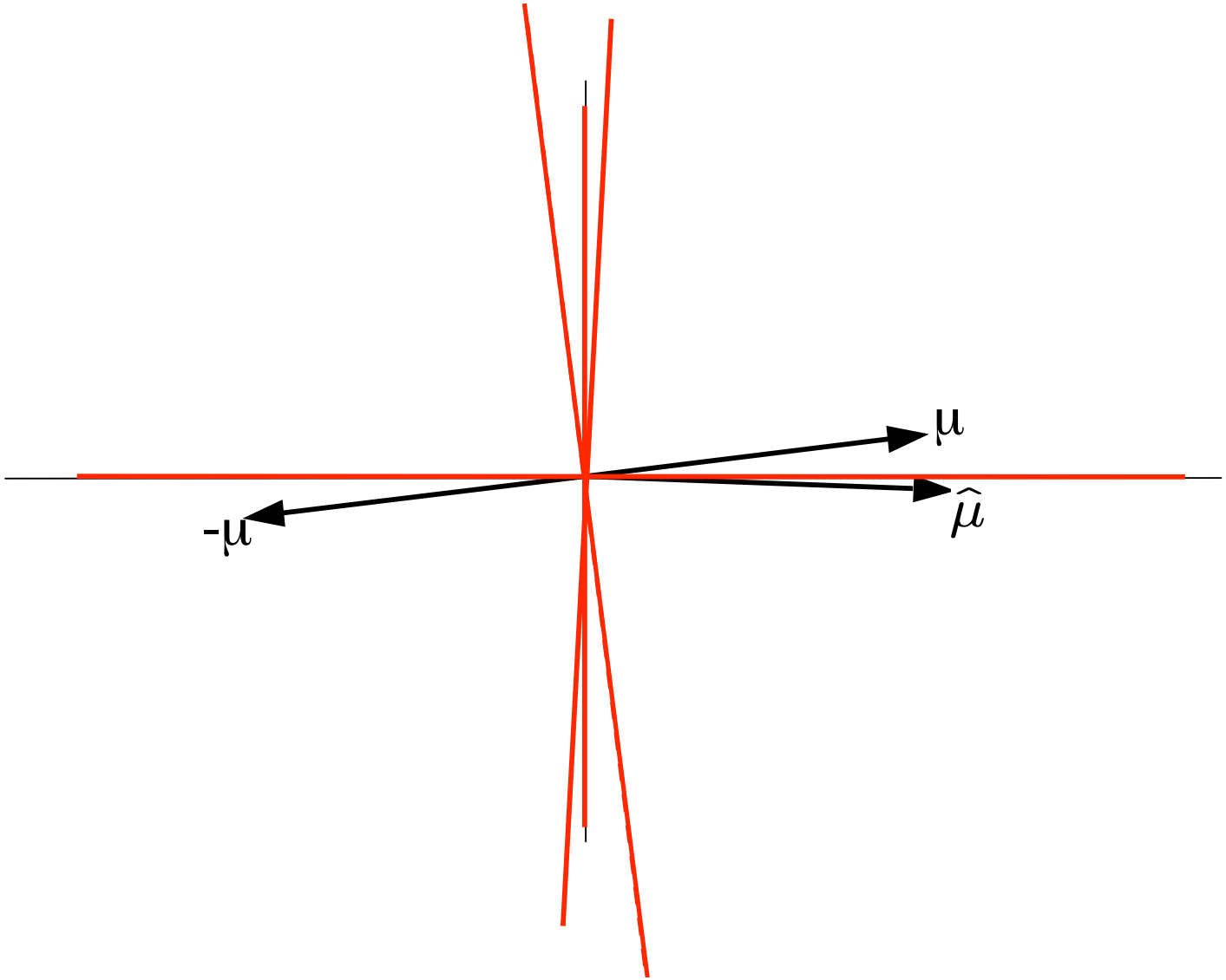
- Assume that μ is estimated from a sample of size m,

- Given an estimator $\hat{\mu} = \hat{\mu}(S^m)$ and a features subset F , we consider the classifier $h|_F(x) = sign(\hat{\mu}|_F \cdot x|_F)$
- We want to find a subset of features that minimizes the average generalization error:

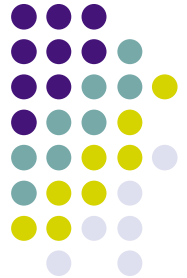
$$E_{S^m} \text{ error}(\hat{\mu}(S^m))$$

- We assume)
  we only need to find the optimal number of features
- We consider the optimal estimator: $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} y_i x_i$

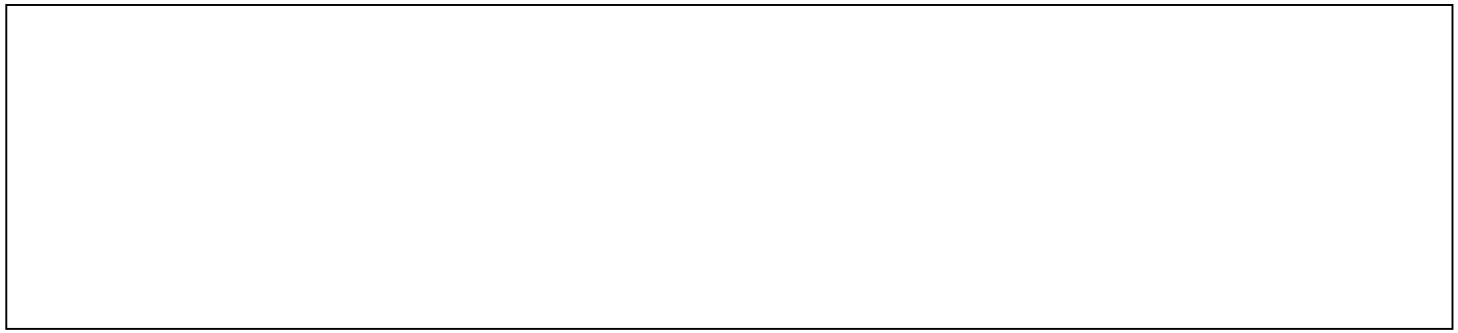
Illustration



Result

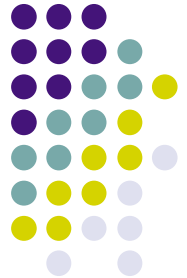


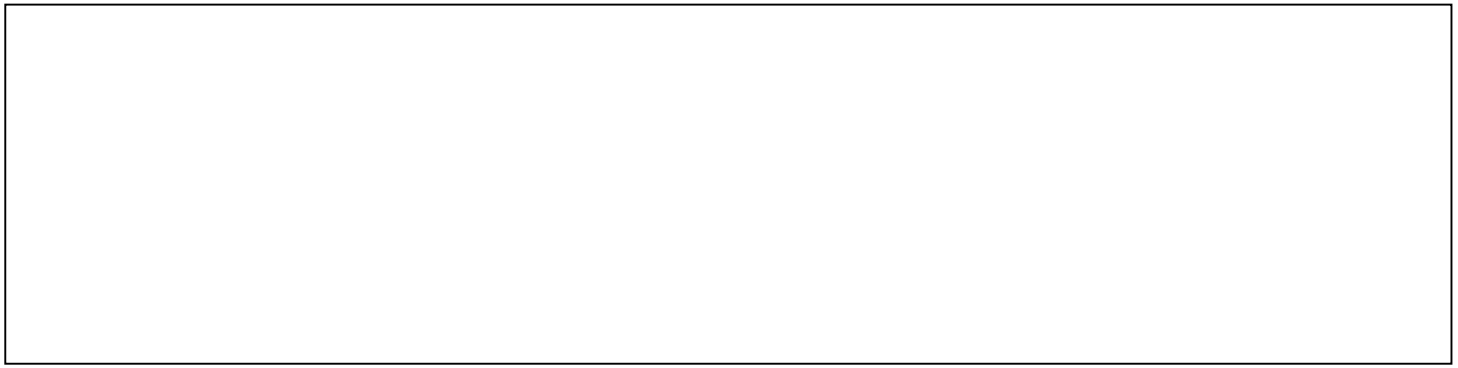
- **The number of features that minimizes the average error for a training set of size m is:**



- Observations:
  - If $\mu[j] \xrightarrow[j \to \infty]{} 0$, there is a non trivial optimal n
  - When $m \to \infty$ then $n_{opt} = N$ is optimal choice
  - Decision on adding depends on other features

Solving for Specific μ





Solving for Specific μ - Cont.

$\mu[j] = \frac{1}{\sqrt{2^j}}$

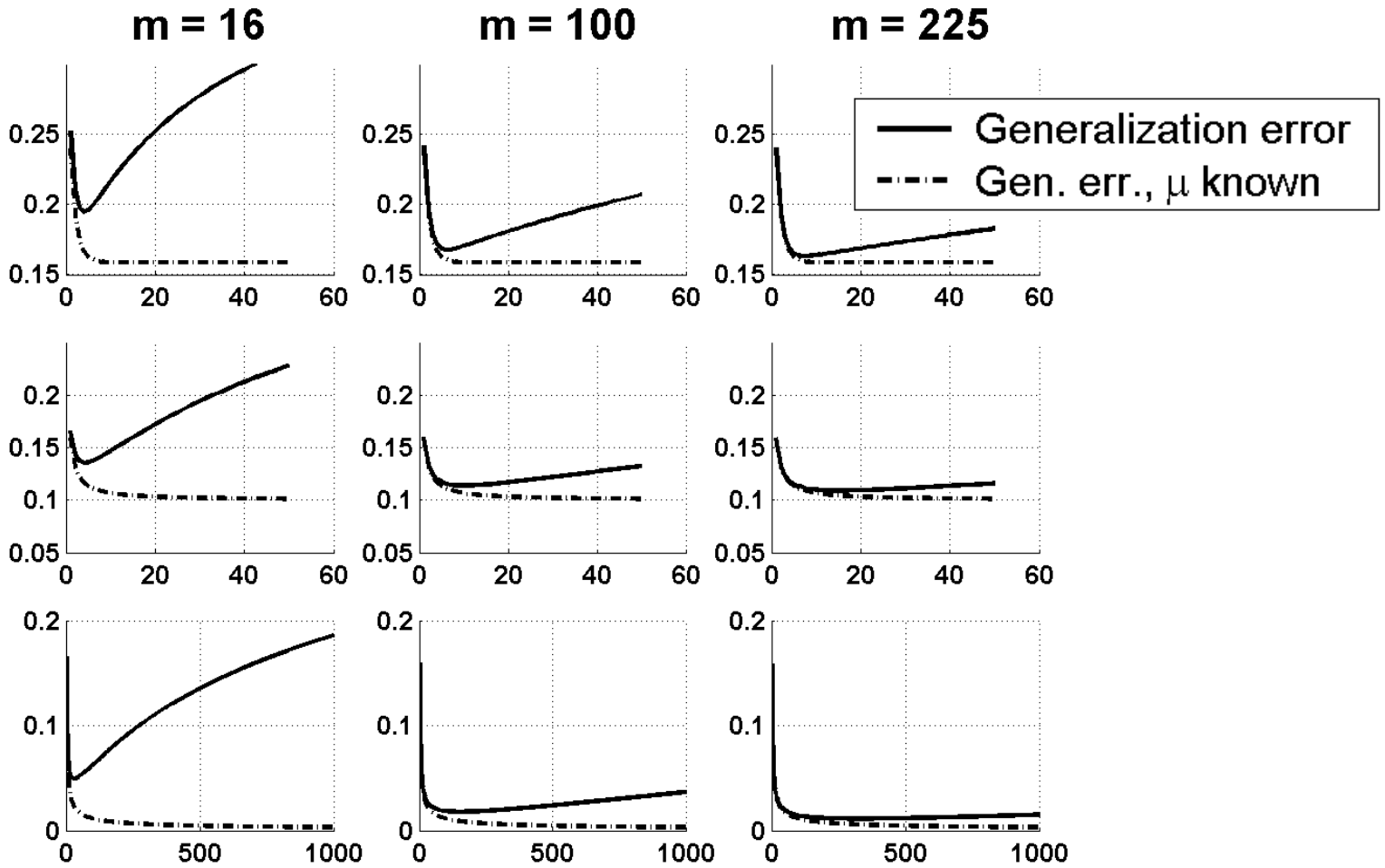
$(n_{opt} \cong \log m)$

$\mu[j] = \frac{1}{j}$

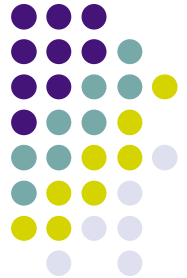
$(n_{opt} \cong \sqrt{m})$

$\mu[j] = \frac{1}{\sqrt{j}}$

$(n_{opt} \cong m)$



Proof



- For a given estimator $\widehat{\mu}$ for μ, the generalization error of the classifier $h(x) = sign(\widehat{\mu} \cdot x)$ is:

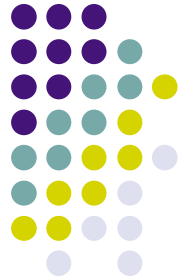
$$\text{error}(\widehat{\mu}) =$$

$$0)$$

$$= Pr(\widehat{\mu} \cdot x < 0)$$

(Φ is the CDF function of a standard Gaussian)

Proof



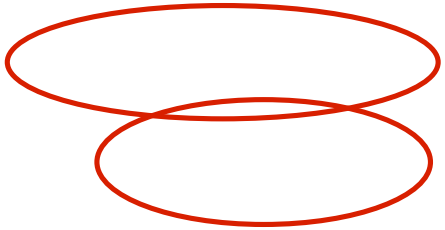
- For a given estimator $\widehat{\mu}$ for μ, the generalization error of the classifier $h(x) = sign(\widehat{\mu} \cdot x)$ is:

$$= Pr_x(\widehat{\mu} \cdot x < 0)$$

- We denote this error by $error(\widehat{\mu})$

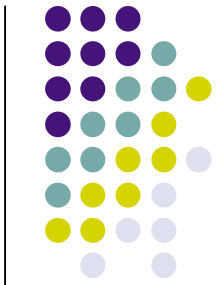
Proof – Cont.

- We want to find the number of features n that minimizes the average error:



- **Lemma:** $\Phi\left(-\frac{E_{S^m}\sum_{j=1}^{n}\hat{\mu}[j]\mu[j]}{\sqrt{E_{S^m}\sum_{j=1}^{n}\hat{\mu}[j]^2}}\right)$ is a good approximation for $E_{S^m}(\text{error}(\hat{\mu}))$ when m is large enough.

Proof – Cont.



- Therefore, minimizing $E_{S^m}(\text{error}(\hat{\mu}))$ is

  equivalent to maximizing $f(n) = \frac{E_{S^m} \sum_{j=1}^{n} \hat{\mu}[j]\mu[j]}{\sqrt{E_{S^m} \sum_{j=1}^{n} \hat{\mu}[j]^2}}$

- For the optimal estimator $\hat{\mu}(S^m) = \frac{1}{m} \sum_{i=1}^{m} y_i x_i$,

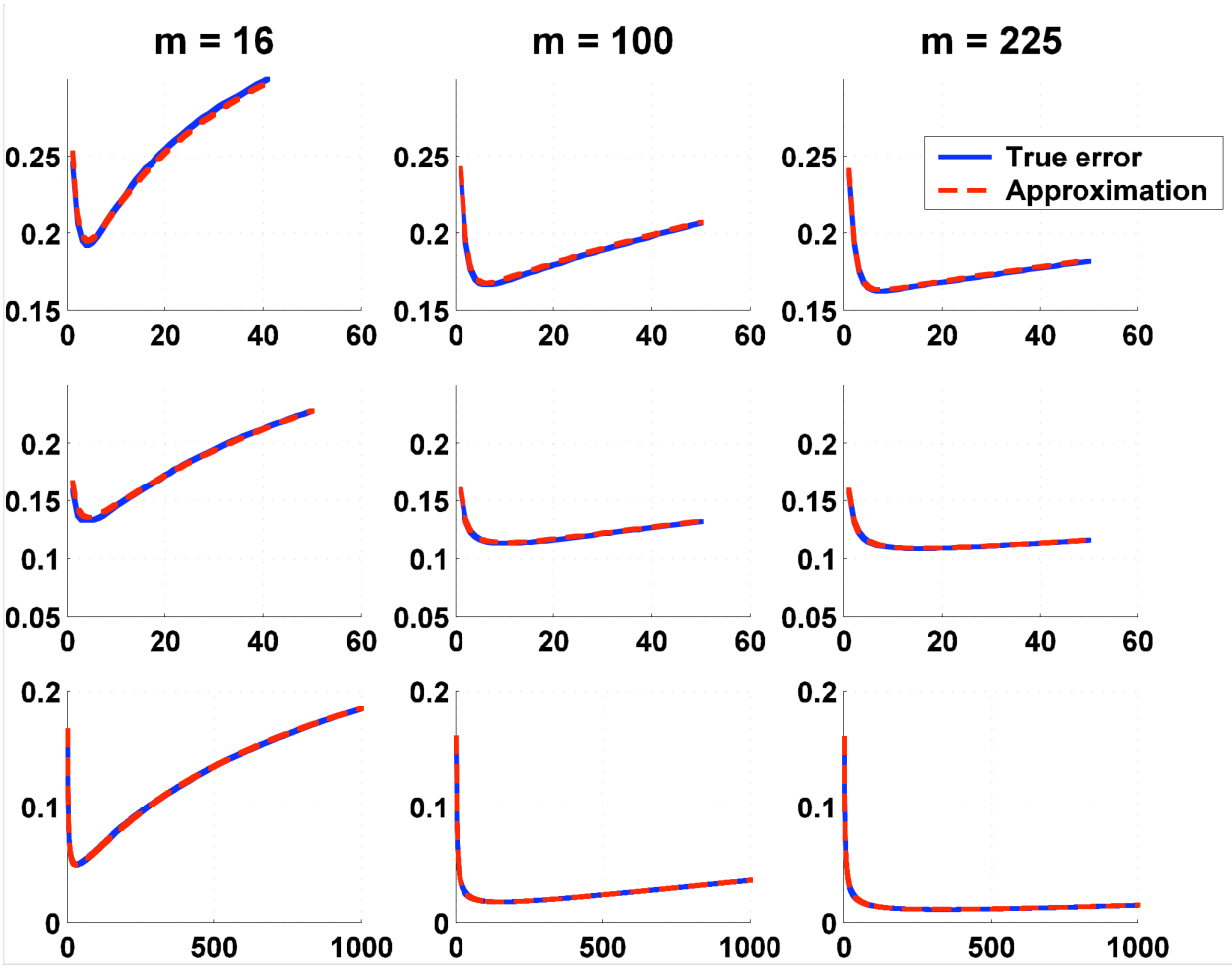
  $E_{S^m}(\hat{\mu}[j]) = \mu[j]$ and $E_{S^m}(\hat{\mu}[j]^2) = \mu[j]^2 + \frac{1}{m}$

  Therefore,

$$f(n) = \frac{\sum_{j=1}^{n} \mu[j]^2}{\sqrt{\frac{n}{m} + \sum_{j=1}^{n} \mu[j]^2}}$$

□

“Empirical Proof” of the Lemma



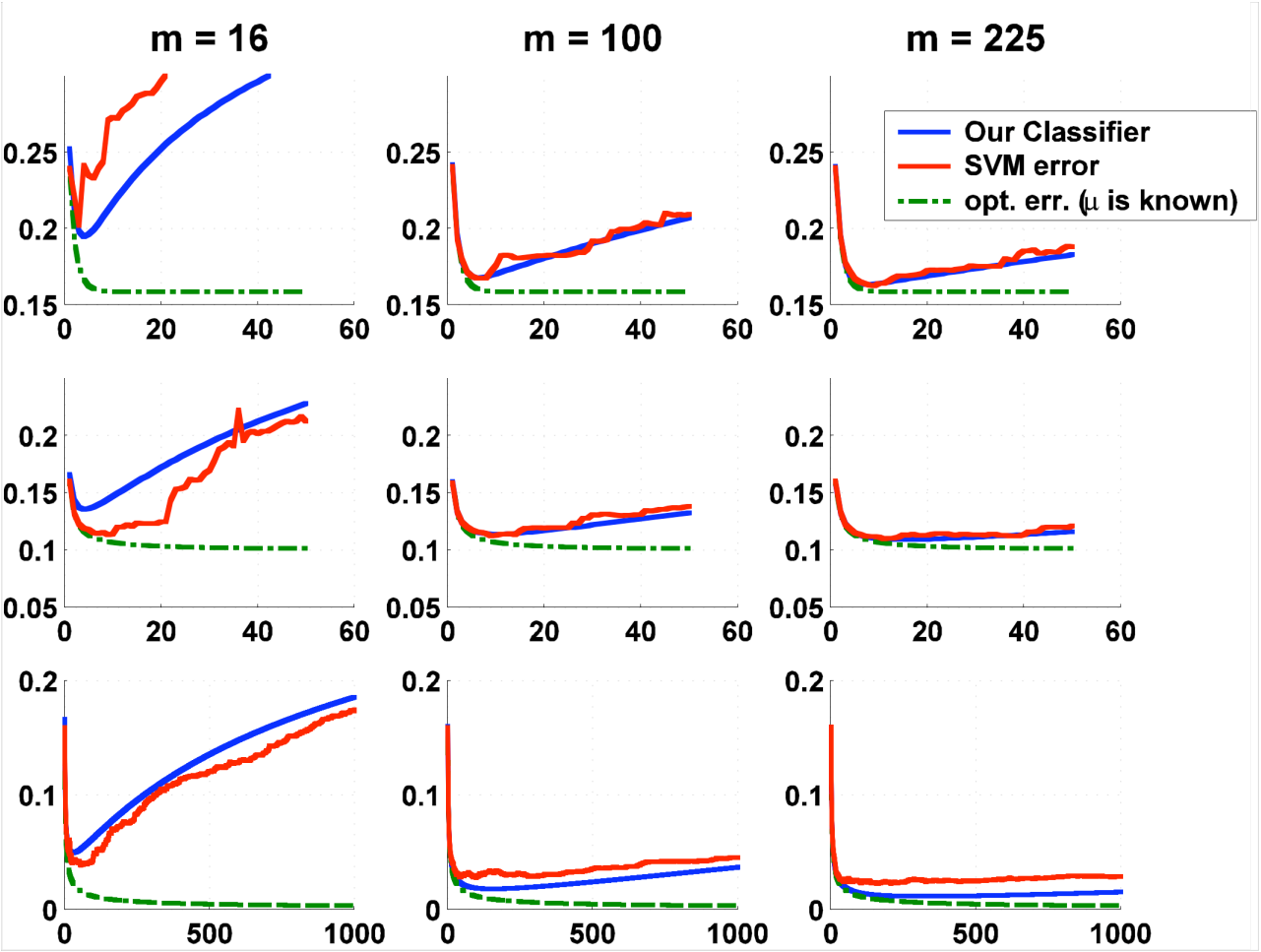
Linear SVM Error

(averaged on 200 repeats, c=0.01, using Gavin Cawley's tool box)

$$\mu[j] = \frac{1}{\sqrt{2j}}$$

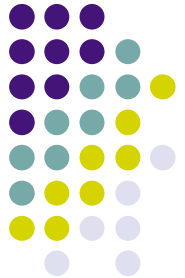
$$\mu[j] = \frac{1}{j}$$

$$\mu[j] = \frac{1}{\sqrt{j}}$$



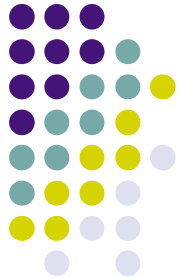
x-axis: number of features

Conclusions



- Even when all the features carry information and are independent,
  - Using all the features may be suboptimal
  - The decision to add a feature depends on the others
- The optimal number of features depends critically on the sample size (= the quality of model estimation)

Further Research Directions



- More general setting
- What can we learn from the dependency between the # of samples and the optimal # of features on the problem in hand?
- Adapt to general dimensionality reduction

What is The Optimal Number of Features?

A learning theoretic Perspective

Amir Navot

Joint work with Ran Gilad-Bachrach, Yiftah Navot and Naftali Tishby