Kernel Conditional Graphical Models

F. Pérez-Cruz, Z. Ghahramani and M. Pontil

13th October 2005

INDEX

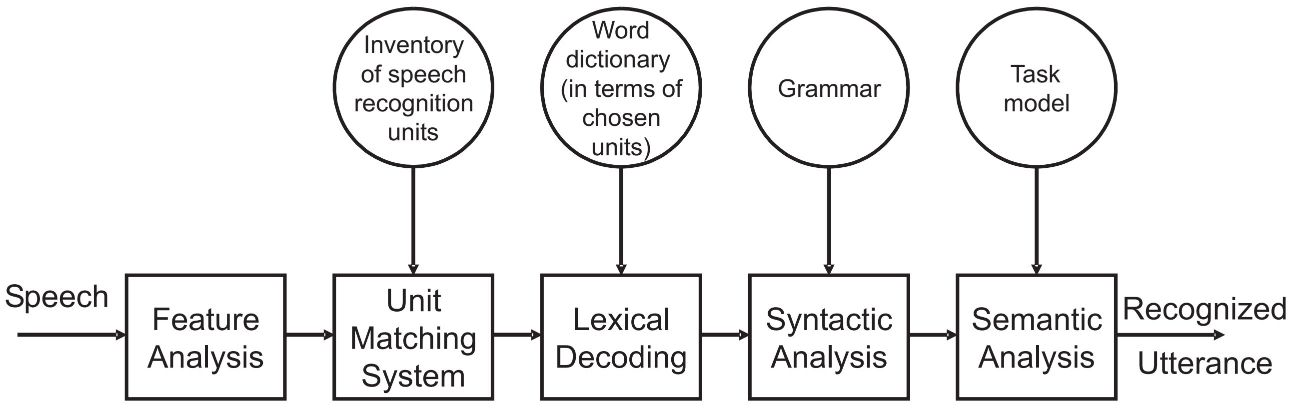
- Applications
- Common Notation
- Conditional Random Fields
- CRFs Extensions
- HM-SVMs and M^3 Networks.
- Kernel Conditional Graphical Model.
- Experiments with word recognition.

Applications

- Digital Communication.
- Automatic Speech Recognition.
- Speaker Identification.
- Part-of-Speech Tagging.
- Secondary structure of DNA.
- Mobile Fraud detection.
- Others.

Automatic Speech Recognition

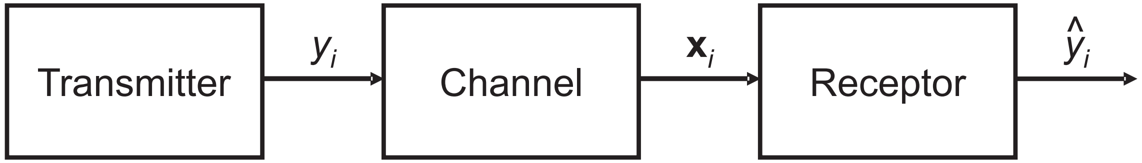
- The inputs to our ASR will be spoken words.
- The labels to each input will be their phonemes.
- ASR Classic model:



- Intra-phoneme temporal relation.
- Statistical relation between phonemes.

Channel Equalization

- The inputs to our receptor are the outputs of the communication channel.
- The Labels will be the transmitted symbols
- The receptor must recover the transmitted symbols.
- Simplified communication scheme:



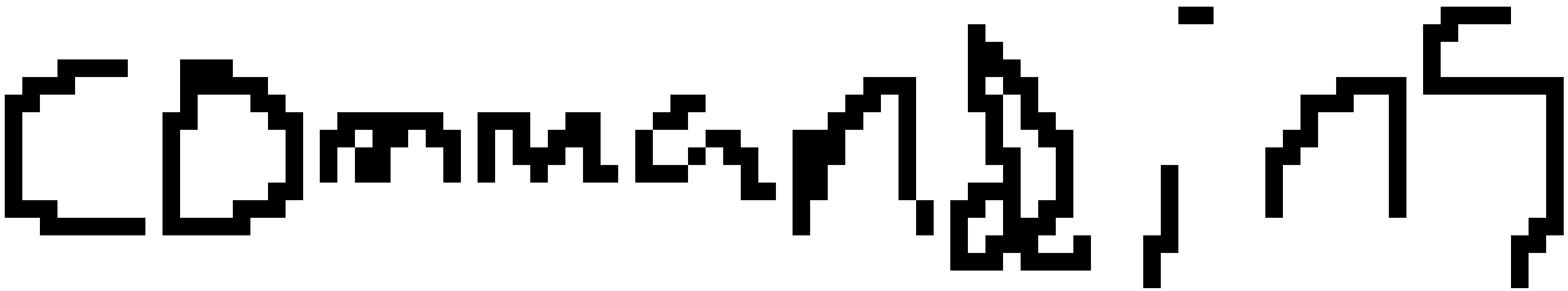
- The channel affects the transmitted signal with linear and nonlinear interferences and noise.

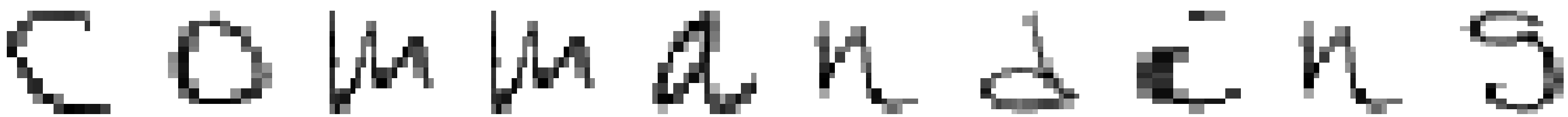
Channel Equalization II

- Known channel model: intersymbol interference and Gaussian noise.
  - Channel estimation (Markov chain)
  - Decoding using Viterbi's algorithm.
- Unknown channel model: + nonlinear interferences.
  - Symbol equalizer, using nonlinear techniques as neural networks or SVMs.
  - The temporal information is not exploit in the receptor ends.

Handwritten Word Recognition

- The inputs will be images of words.
- The labels will be the letters of each image.
- Examples of the word *commanding*:





AVAILABLE DATASETS

- MIT dataset:
  - 55 words of differen length. from 155 subjects.
  - Each letter is a 16×8 binary image.
  - Currecntly used in the literature.
- UCIIIM database:
  - 580 sheets of paper with 54 letters in each (Upper and lower cases Spanish alphabet). One sheet per person.
  - Each letter is a 28×28 8 bits image.
  - We can construct any desired text.

Common Notation

- Detection of multiple labels from samples.
- samples: $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$
- The inputs belong to $\mathbf{x}_n \in \mathcal{X}$.
- Las labels $\mathbf{y}_n$ present the following structure:

$$\mathbf{y}_n \in \mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_L \text{ and } \mathcal{Y}_l = \{1, 2, \ldots, q\}.$$

- The length of the sequence L can depend on each sample (L_n).
- The possible labellings grows exponentially with L (q^L).
- Our objective is estimate the label $\mathbf{y}_*$ given an input $\mathbf{x}_*$ and the data $\mathcal{D}$.

PREDICTIONS

- Linear model ($\mathbf{w}$) over a set of features ($f(\mathbf{x}_n, \mathbf{y}_n)$).
- Likelihood (Softmax function).

$$p(\mathbf{Y}|\mathbf{w}, \mathbf{X}) = \prod_{n=1}^{N} \frac{\exp(\mathbf{w}^\top f(\mathbf{x}_n, \mathbf{y}_n))}{\sum_{\mathbf{y}} \exp(\mathbf{w}^\top f(\mathbf{x}_n, \mathbf{y}))}$$

  The sum runs for exponentially increasing factors (q^L terms).

- Gaussian Prior:

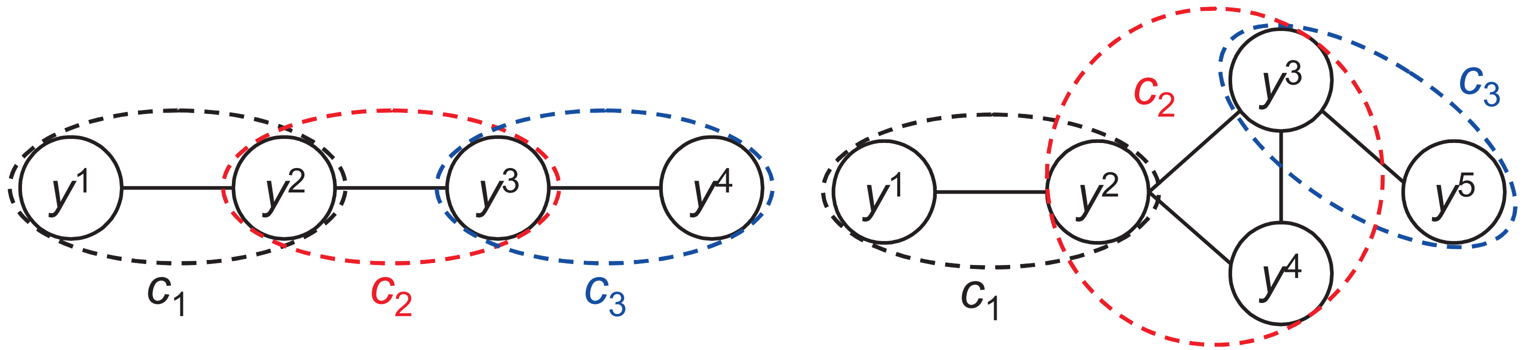
$$p(\mathbf{w}) = \frac{1}{\sqrt{(2\pi C)^{|\mathcal{W}|}}} \exp\left(-\frac{||\mathbf{w}||^2}{2C}\right)$$

- Predictions (integrate over $\mathbf{w}$):

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) = \int p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}, \mathbf{X}) d\mathbf{w}$$

Conditional Random Fields

- Option 1: Direct Multi-class problem. $\longrightarrow$ Complexity q^L.
- Option 2: Assume each label is independent. $\longrightarrow$ Lose information about relationship between labels.
- Option 3 (CRF): Assume a graphical model ($G = (V, E)$ with T cliques $c_1, c_2, \ldots, c_T$.) over the labels that encode the dependencies (independencies) of the labels given the input vector.



Likelihood model

- We get the following model for the likelihood:

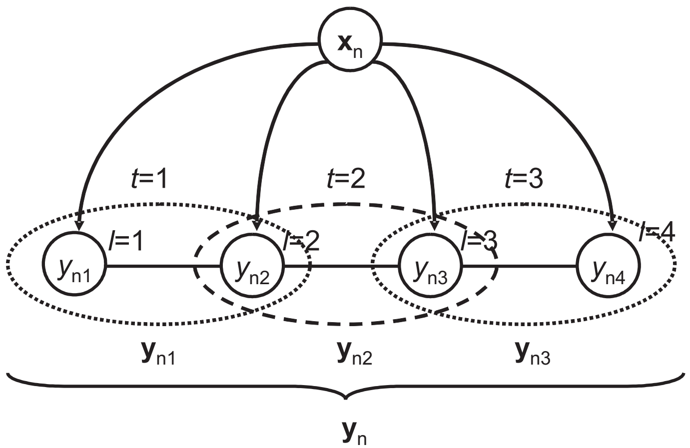
$$p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{w}) \propto \exp\left(\sum_{t=1}^{T} \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt})\right)$$

Fundamental theorem of random fields (Hammersley & Clifford, 1971).

- In which the graphs enforces a given structure for the features:

$$f(\mathbf{x}_n, \mathbf{y}_n) = \begin{bmatrix} f_1(\mathbf{x}_n, \mathbf{y}_{n1}) \\ \vdots \\ f_T(\mathbf{x}_n, \mathbf{y}_{nT}) \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_T \end{bmatrix}$$



SOLUTION

- Maximum a Posteriori (Conditional Random Fields, CRF):

$$\min_{\mathbf{w}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{n=1}^{N} \left[\log \left(\sum_{\mathbf{y}} \exp\left(o(\mathbf{x}_n, \mathbf{y})\right) \right) - o(\mathbf{x}_n, \mathbf{y}_n) \right] \right\}$$

where

$$o(\mathbf{x}_n, \mathbf{y}) = \sum_{t=1}^{T} \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t)$$

- CRF initially proposed with Maximum Likelihood.
- The logarithm of the partition function still has exponentially many terms. But it can be computed efficiently using a forward-backward algorithm.

CRFs Extensions

- Bayesian CRF (Qi et al., 2005):
  - Posterior approximation using Power EP.
  - Predictions using EP, independent distribution per label.
- Kernel CRF (Lafferty et al., 2004;Altun et al., 2004):
  - Representer theorem: $\mathbf{w}_t = \sum_{n'=1}^{N} \sum_{\mathbf{y}'_t} \beta^t_{n',\mathbf{y}'_t} \mathbf{f}_t(\mathbf{x}_{n'}, \mathbf{y}'_t)$.
  - Tractability: clique selection or sparse greedy approximation.
- Gaussian Process CRF (Altun et al., 2004):
  - Posterior approximation using delta over the mode.
  - Representer theorem.

Hidden Markov SVMs (Altun et al., 2003)

- Change loss-function to hinge loss plus margin:

$$\min_{\mathbf{w}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{n=1}^{N} \left[\max_{\mathbf{y}} \Big(M_{\mathbf{y}_n,\mathbf{y}} + o(\mathbf{x}_n, \mathbf{y}) - o(\mathbf{x}_n, \mathbf{y}_n) \Big) \right]_+ \right\}$$

  Margin $M_{\mathbf{y}_n,\mathbf{y}}$ depends on the true label and the possible labelling.

- Constraint optimization:

$$\min_{\mathbf{w},\xi_n} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{n=1}^{N} \xi_n \right\}$$

  subject to: $o(\mathbf{x}_n, \mathbf{y}_n) - o(\mathbf{x}_n, \mathbf{y}) \geq M_{\mathbf{y}_n,\mathbf{y}} - \xi_n \qquad \forall n, \forall \mathbf{y} \quad M_{\mathbf{y}_n,\mathbf{y}_n} = 0$

- Exponentially many (possible) Support Vectors.

- Optimization exploits the fact that many SVs will be zero.

Max-Margin Markov Nets (Taskar et al., 2004)

- Lagrange Multipliers

$$L(\mathbf{w}, \xi_n, \alpha_{n,\mathbf{y}}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$
$$-\sum_{n=1}^{N}\sum_{\mathbf{y}}\alpha_{n,\mathbf{y}}\left(\sum_{t=1}^{T}\mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt}) - \sum_{t=1}^{T}\mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t) - M_{\mathbf{y}_n,\mathbf{y}} + \xi_n\right)$$

- Margin: Hamming distance between the labels, $\sum_{l=1}^{L}[1-\delta(y_{nl}, y_l)]$.

- Reduce the exponentially many $\alpha_{n,\mathbf{y}}$ to:

$$\beta^t_{n,\mathbf{y}_t} = \sum_{\mathbf{y}\sim\mathbf{y}_t}\alpha_{n,\mathbf{y}} \quad \forall n, \forall t, \forall \mathbf{y}_t$$
$$\beta^l_{n,y_l} = \sum_{\mathbf{y}\sim y_l}\alpha_{n,\mathbf{y}} \quad \forall n, \forall l, \forall y_l \qquad\qquad \beta^l_{n,y_l} = \sum_{\mathbf{y}_t\sim y_l}\beta^t_{n,\mathbf{y}_t}$$

OUR PROPOSAL

- Define the margin per clique: $M_{\mathbf{y}_n,\mathbf{y}} = \sum_{t=1}^{T} M'_{\mathbf{y}_{nt},\mathbf{y}_t}$.
- Redefine the slack variables as: $\xi_n = \sum_{t=1}^{T} \xi_{nt}$.
- We can now rewrite the Lagrange functional only in terms of $\beta^t_{n,\mathbf{y}_t}$:

$$L(\mathbf{w}, \xi_n, \beta^t_{n,\mathbf{y}_t}) = \sum_{t=1}^{T} \left(\frac{1}{2}\|\mathbf{w}_t\|^2 + C \sum_{n=1}^{N} \xi_{nt} - \sum_{n=1}^{N} \eta_{nt}\xi_{nt} \right. \\ \left. - \sum_{n=1}^{N} \sum_{\mathbf{y}_t} \beta^t_{n,\mathbf{y}_t} \left(\mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt}) - \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t) - M'_{\mathbf{y}_{nt},\mathbf{y}_t} + \xi_n \right) \right)$$

  which can be solved independently per clique, if there were not any constraints liking the $\beta^t_{n,\mathbf{y}_t}$ of each clique.
- Let's drop this constraint ...

Kernel Conditional Graphical Model

- We propose the following constrained problem:

$$\min_{\mathbf{w},\xi_{nt}} \sum_{t=1}^{T} \left\{ \frac{1}{2}\|\mathbf{w}_t\|^2 + C \sum_{n=1}^{N} \xi_{nt} \right\}$$

subject to:

$$\mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt}) - \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t) \geq M'_{\mathbf{y}_{nt},\mathbf{y}_t} - \xi_{nt}$$

- Previous constraint:

$$\sum_{t=1}^{T} \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt}) - \sum_{t=1}^{T} \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t) \geq \sum_{t=1}^{T} M'_{\mathbf{y}_{nt},\mathbf{y}_t} - \sum_{t=1}^{T} \xi_{nt}$$

- The new constraints are more restrictive.

- The new constraints reduces the number of constraints.

Example

- Given a sample $\mathbf{x}_n$ and $\mathbf{y}_n$ with a graph with 10 cliques.
- the outputs for the true labelling an other possible labelling are:

t	1	2	3	4	5	6	7	8	9	10
$\mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt})$	1.5	0.8	2.0	1.1	2.5	2.1	1.7	1.9	2.7	1.2
$\mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t)$	0.2	1.7	2.0	1.1	2.5	0.3	1.7	0.7	1.0	1.2
$M'_{\mathbf{y}_{nt},\mathbf{y}_t}$	1	1	0	0	0	1	0	1	1	0

$\sum \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt}) = 17.5$, $\sum \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t) = 12.4$ and $\sum M'_{\mathbf{y}_{nt},\mathbf{y}_t} = 5$.

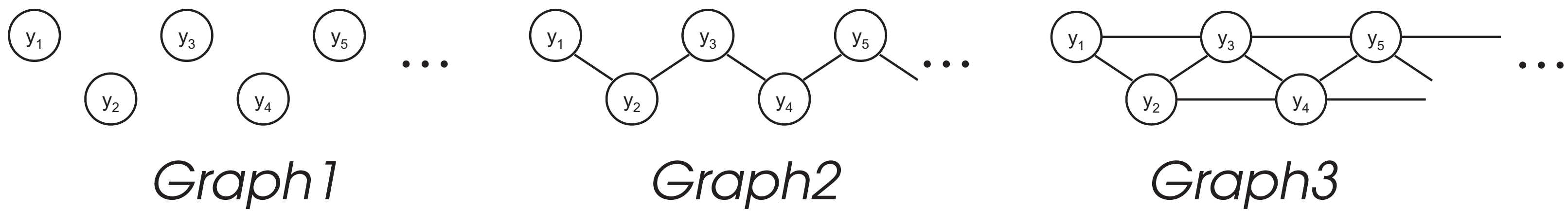
- the outputs for the true labelling an other possible labelling are:

t	1	2	3	4	5	6	7	8	9	10
$\mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt})$	1.5	0.8	2.0	1.1	2.5	2.1	1.7	1.9	2.7	1.2
$\mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t)$	0.2	1.9	2.0	1.1	2.5	0.3	1.7	0.7	1.0	1.2
$M'_{\mathbf{y}_{nt},\mathbf{y}_t}$	1	1	0	0	0	1	0	1	1	0

$\sum \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt}) = 17.5$, $\sum \mathbf{w}_t^\top \mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_t) = 12.6$ and $\sum M'_{\mathbf{y}_{nt},\mathbf{y}_t} = 5$.

EXPERIMENTS

- Written word recognition (Thanks to B. Taskar for proving the dataset).
- We have 55 different-length words written by $\sim$150 subjects.
- First experiment: Obtain a machine for recognising these words. We use the following 3 models:



- Second experiment: Obtain a machine for recognising any word. Using the letters we have created a training data set with all the words with more than 4 letter in: *"The picture of Dorian Grey"*.

First Expeimrent

- We divided de data set in 10 sets. We have used one for training and 9 for testing (Comparison with Taskar's results).

- We have used a multiclass feature:

$$f_t(\mathbf{x}_n, \mathbf{y}_{nt}) = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \phi_t(\mathbf{x}_{nt}) \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}_{q^{|\mathbf{y}_t|} \times 1} \leftarrow \text{position} \quad \mathbf{y}_{nt}$$

  $\phi(\mathbf{x}_{nt})$ is a nonlinear transformation characterised by an RBF kernel.

- Computed the result using a stationary and nonstationary models.

Results

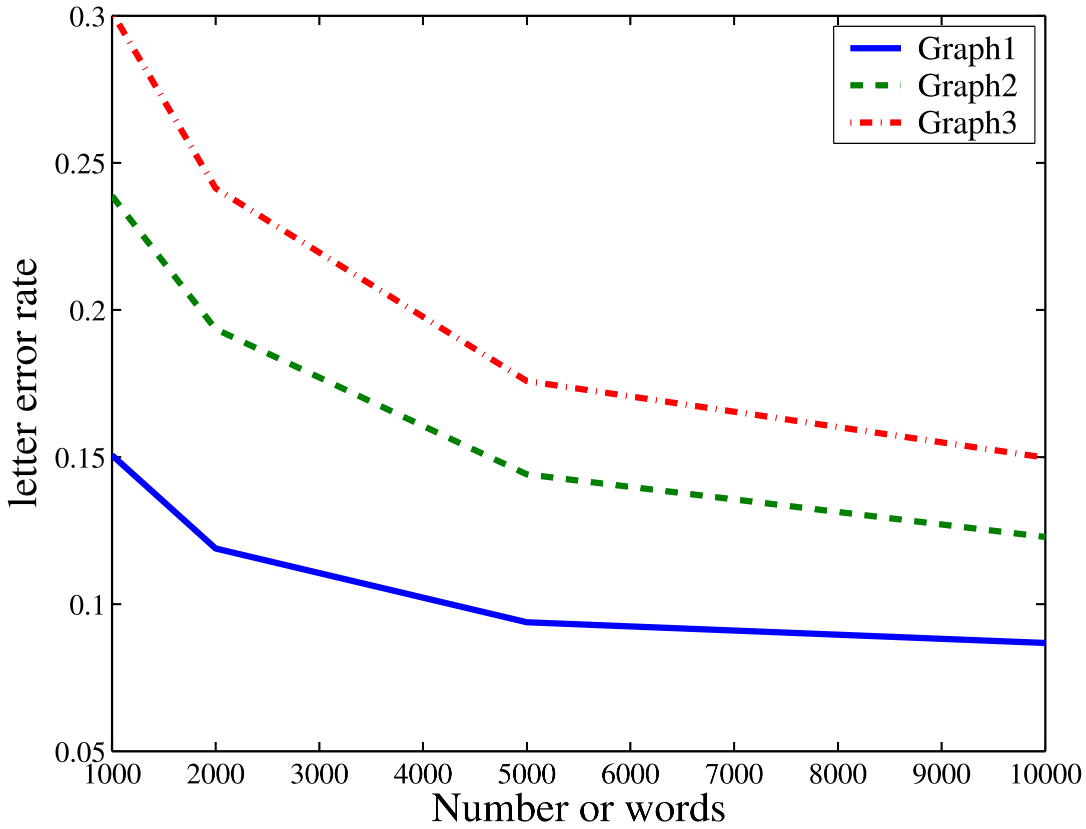
- Results for 10-fold cross validation:

	non-stationary	Stationary
graph1	27.16%	17.12%
graph2	12.48%	20.61%
graph3	5.79%	10.23%

- We used a Gaussian kernel with $\sigma = \sqrt{d/16}$ and $C = 5$.
- Results for the Maximum Margin Network for letter recognition $\sim$13%.

Second Experiment

- We have trained the KCGM using an increasing number of words.



Second Feature

- For *Graph 2*:

$$\mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt}) = [\mathbf{0}, \ldots, \phi(\mathbf{x}_{nt_1}), \ldots \mid \mathbf{0}, \ldots, \phi(\mathbf{x}_{nt_2}), \ldots]^\top$$

- For *Graph 3*:

$$\mathbf{f}_t(\mathbf{x}_n, \mathbf{y}_{nt}) = [\mathbf{0}, \ldots, \phi(\mathbf{x}_{nt_1}), \ldots \mid \mathbf{0}, \ldots, \phi(\mathbf{x}_{nt_2}), \ldots \mid \mathbf{0}, \ldots, \phi(\mathbf{x}_{nt_3}), \ldots]^\top$$

- This feature combines the individual decision per letters with the error correcting given by the graph.
- The weight vectors that classify the same letter are linked together. i.e. *Graph 2*:

$$\mathbf{w}_t = [\mathbf{w}_{t_1}, \mathbf{w}_{t_2}, \ldots, \mathbf{w}_{t_q}, \mathbf{w}_{t_1}, \mathbf{w}_{t_2}, \ldots, \mathbf{w}_{t_q}, \ldots, \mathbf{w}_{t_1}, \mathbf{w}_{t_2}, \ldots, \mathbf{w}_{t_q} \mid \\ \mathbf{w}_{t_1}, \mathbf{w}_{t_1}, \ldots, \mathbf{w}_{t_1}, \mathbf{w}_{t_2}, \mathbf{w}_{t_2}, \ldots, \mathbf{w}_{t_2}, \ldots \mathbf{w}_{t_q}, \mathbf{w}_{t_q}, \ldots, \mathbf{w}_{t_q}]$$

RESULTS

- Results for 10-fold cross validation:

	non-stationary	Stationary	Mixture
graph1	27.16%	17.12%	—
graph2	9.67%	12.43%	3.03%
graph3	2.15%	2.24%	0.71%
graph total	0.61%	0.21%	—

- We used a Gaussian kernel with $\sigma = \sqrt{d/16}$ and $C = 5$.
- Results for the Maximum Margin Network for letter recognition $\sim$13%.
- Results for a SVM with 55 different classes: 2.22%

Second Experiment

- We have trained the KCGM using an increasing number of words.