Latent Dirichlet Allocation (Blei, Ng & Jordan)

LDA models each document as a mixture over latent topics. A topic is a distribution over words.

Generative Model:

For each document j and each word i independently

- Sample mixing coefficients $\boldsymbol{\theta}$ from $Dir(\alpha)$
- Sample word probabilities coefficients $\boldsymbol{\phi}$ from $Dir(\beta)$
- For each word i position in document j choose topic with $P[z_{ij} = k] = \theta_{jk}$
- For each word position i in document j and topic k choose $P[x_{ij} = w] = \phi_{kw}$.

β and α are hyperparameters that can be optimised (learned) from a training corpus of data.

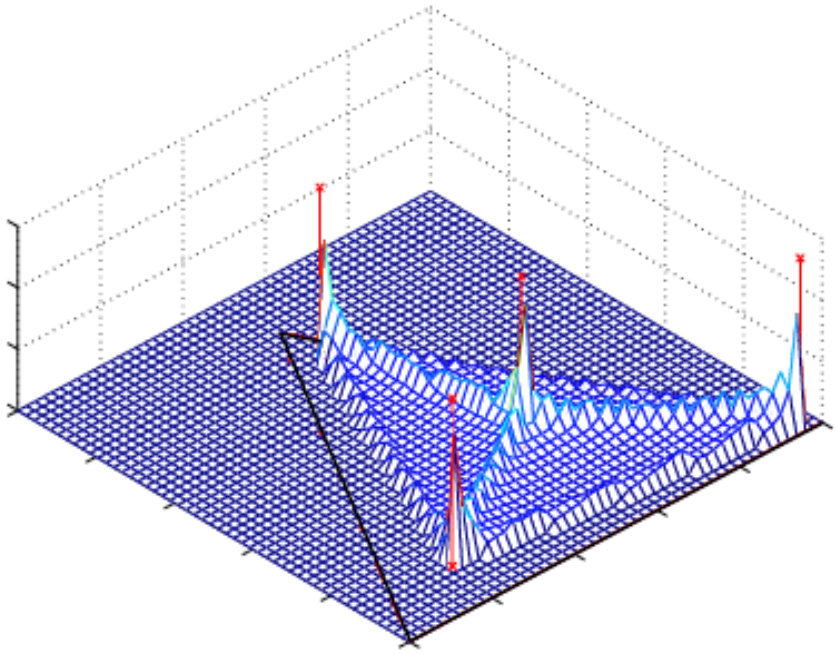
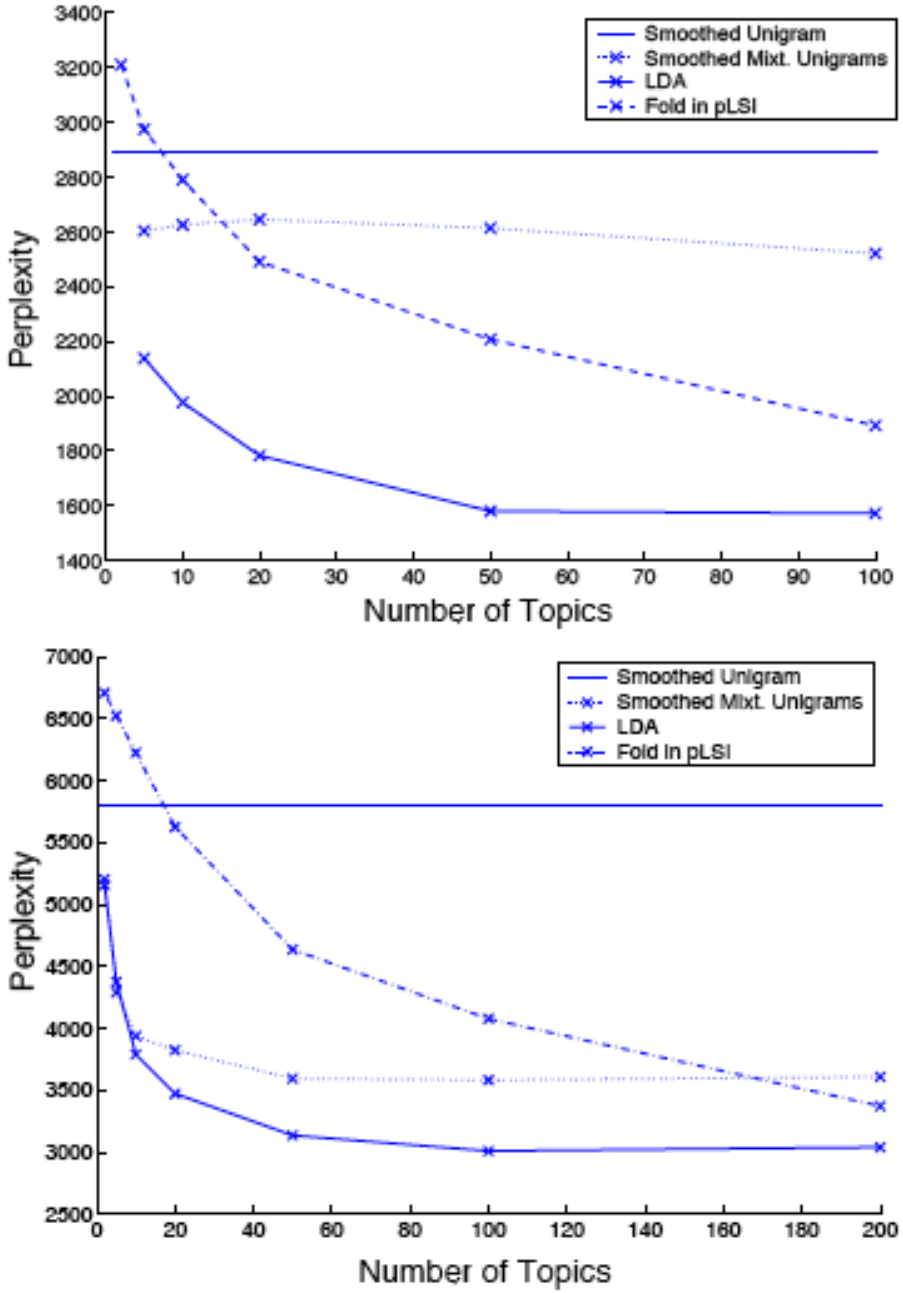


Figure 2: An example density on unigram distributions $p(w|\theta,\beta)$ under LDA for three words and four topics. The triangle embedded in the x-y plane is the 2-D simplex representing all possible multinomial distributions over three words. Each of the vertices of the triangle corresponds to a deterministic distribution that assigns probability one to one of the words; the midpoint of an edge gives probability 0.5 to two of the words; and the centroid of the triangle is the uniform distribution over all three words. The four points marked with an x are the locations of the multinomial distributions $p(w|z)$ for each of the four topics, and the surface shown on top of the simplex is an example of a density over the $(V-1)$-simplex (multinomial distributions of words) given by LDA.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



ure 9: Perplexity results on the nematode (Top) and AP (Bottom) corpora for LDA, the unigram model, mixture of unigrams, and pLSI.

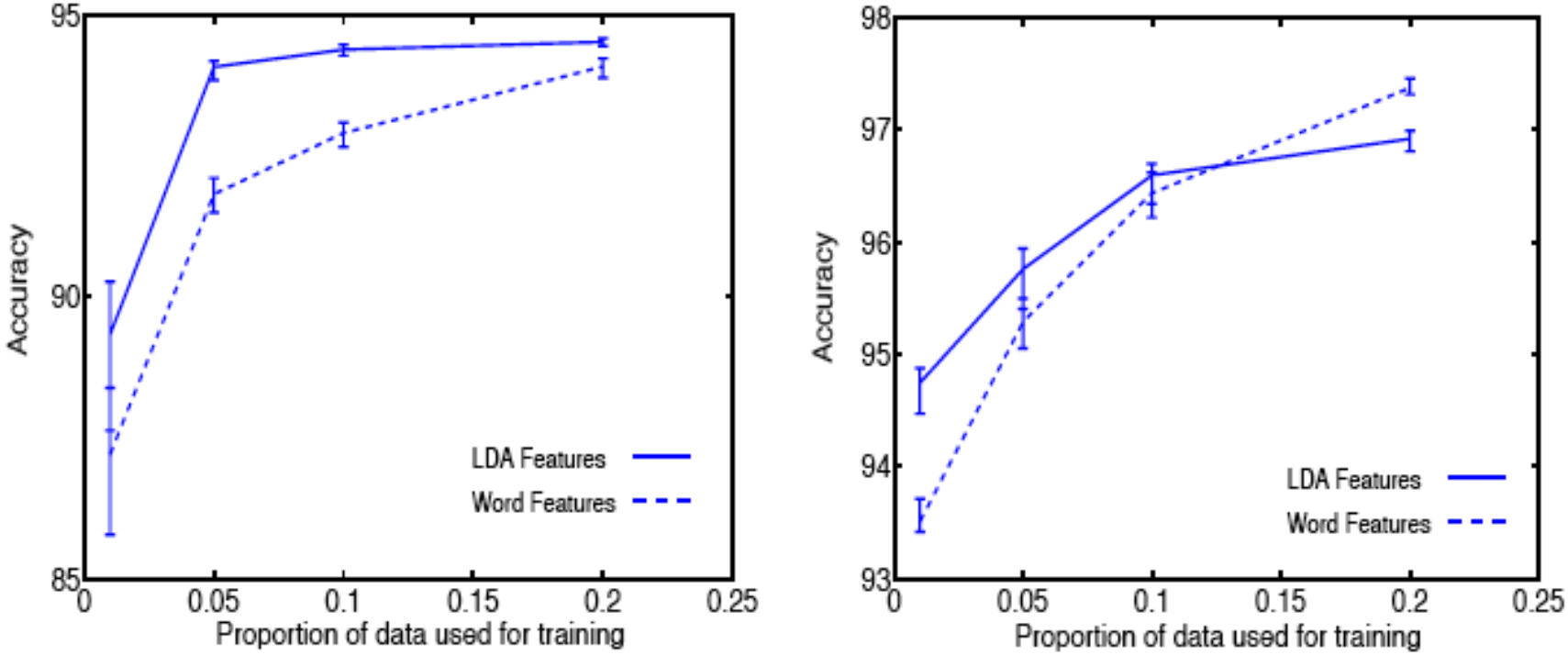


Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

Variational Approximation

Teh, Newman & Welling (NIPS06) show:

- $$q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{ij} q_{ij}(z_{ij}) \prod_{j} q_j(\boldsymbol{\theta}_j) \prod_{k} q_k(\boldsymbol{\phi}_k)$$

  does not work so well!

- Integrate out $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ and approximate

  $$q(\mathbf{z}) = \prod_{ij} q_{ij}(z_{ij})$$

  works better!

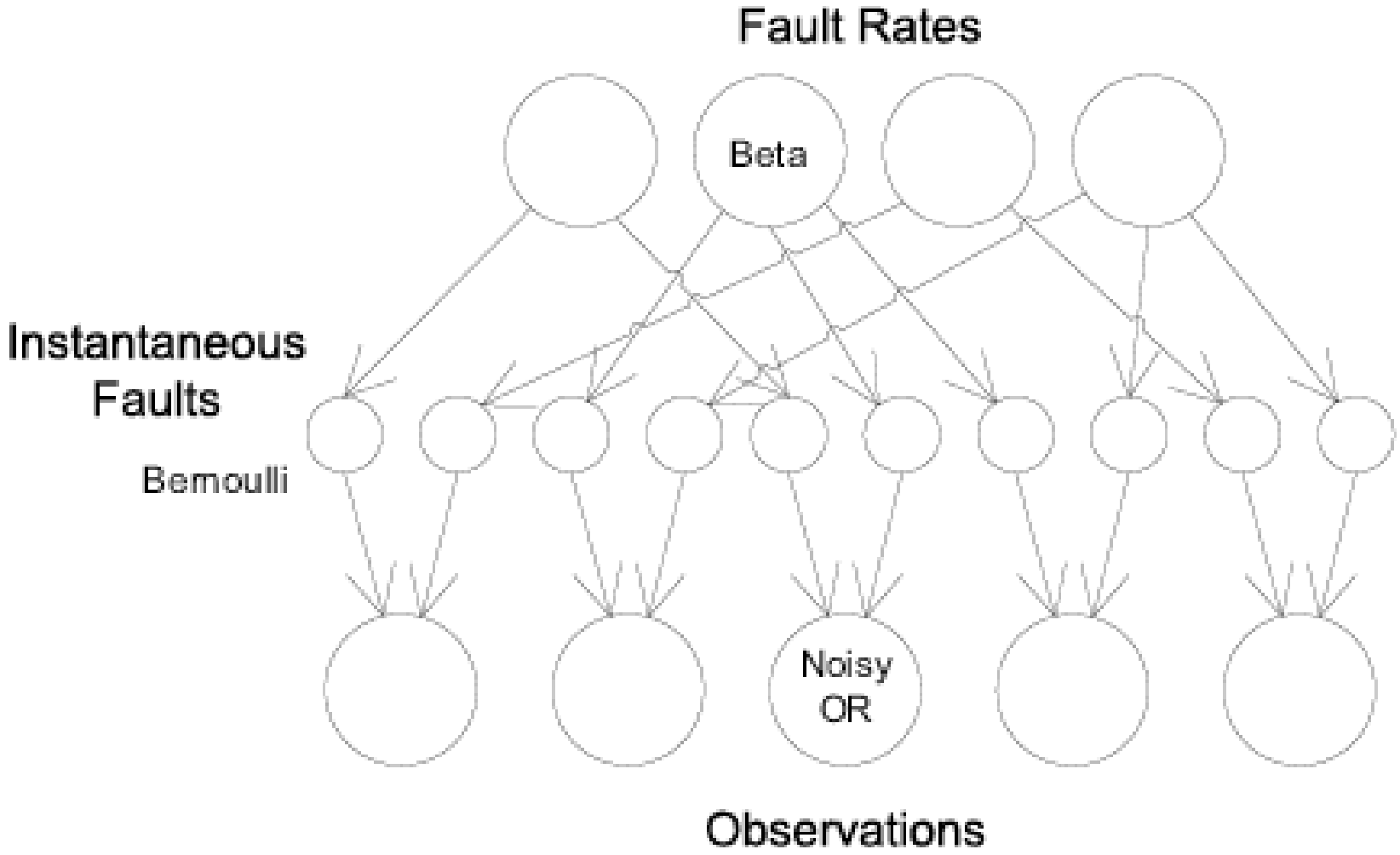


Figure 1: The full graphical model for the diagnosis of Internet faults

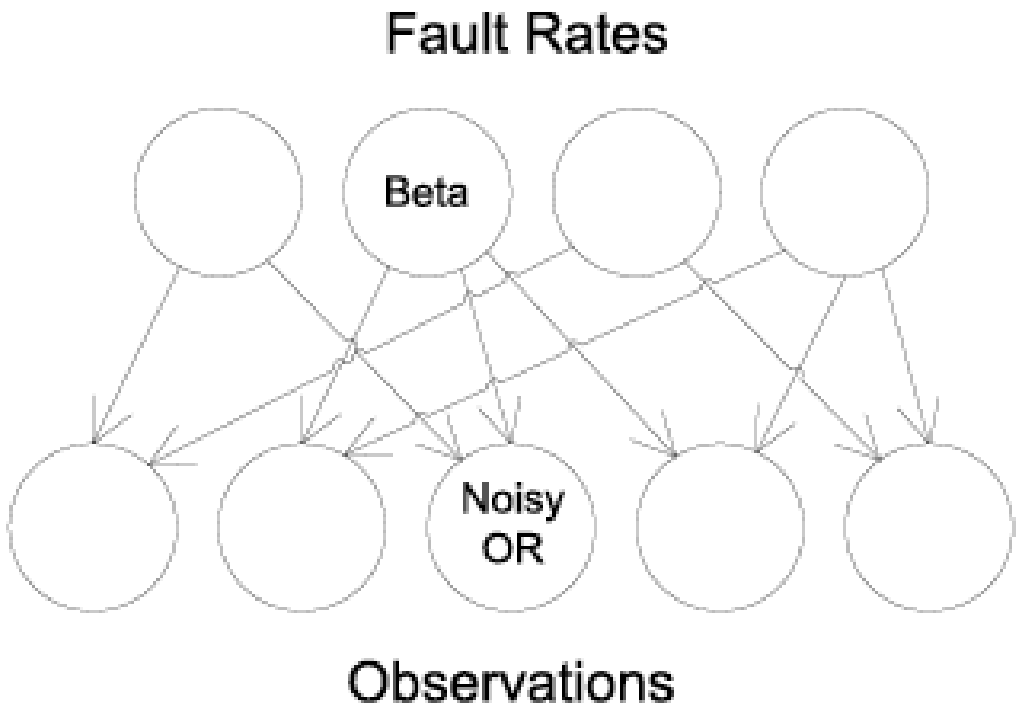


Figure 2: Graphical model after integrating out instantaneous faults: a bipartite noisy-OR network with Beta distributions as hidden variables

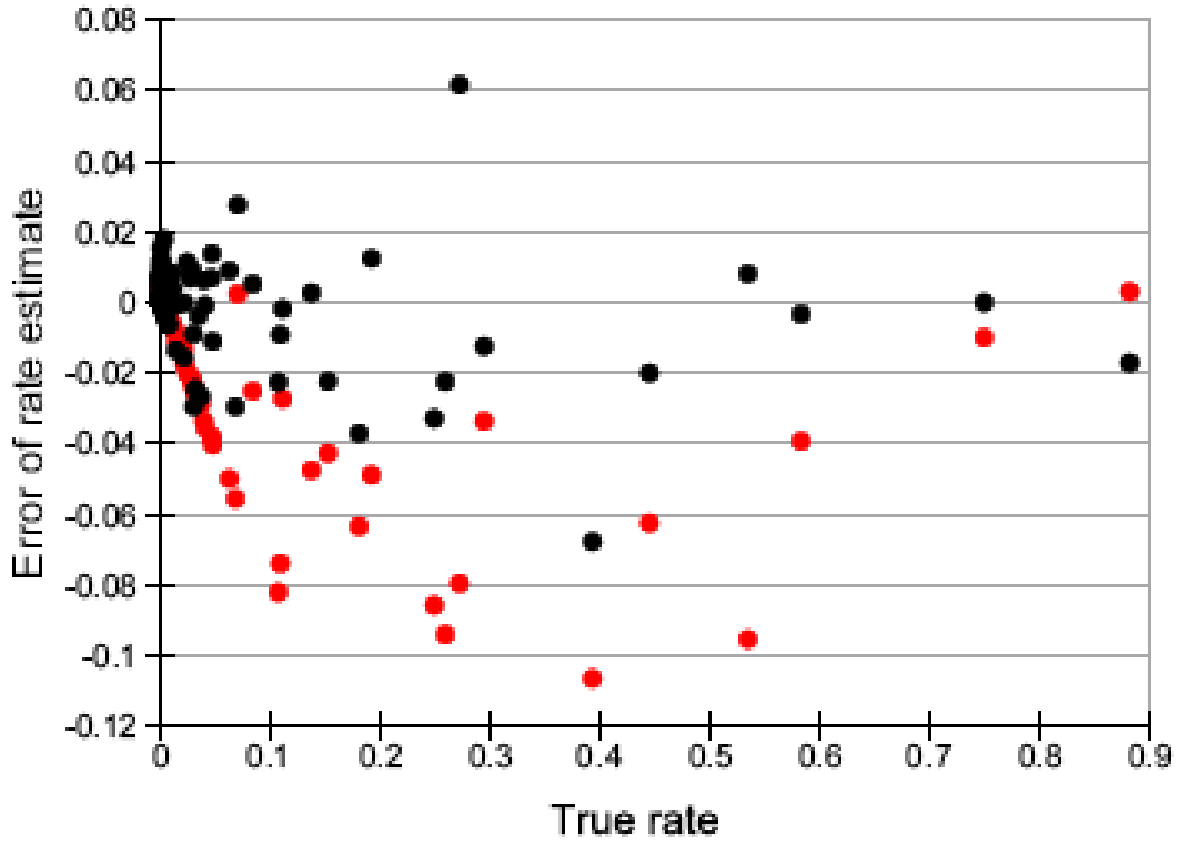


Figure 3: The error in estimate of rate versus true underlying rate. Black dots are L-BFGS, Red dots are Stochastic Gradient Descent with 20 epochs.

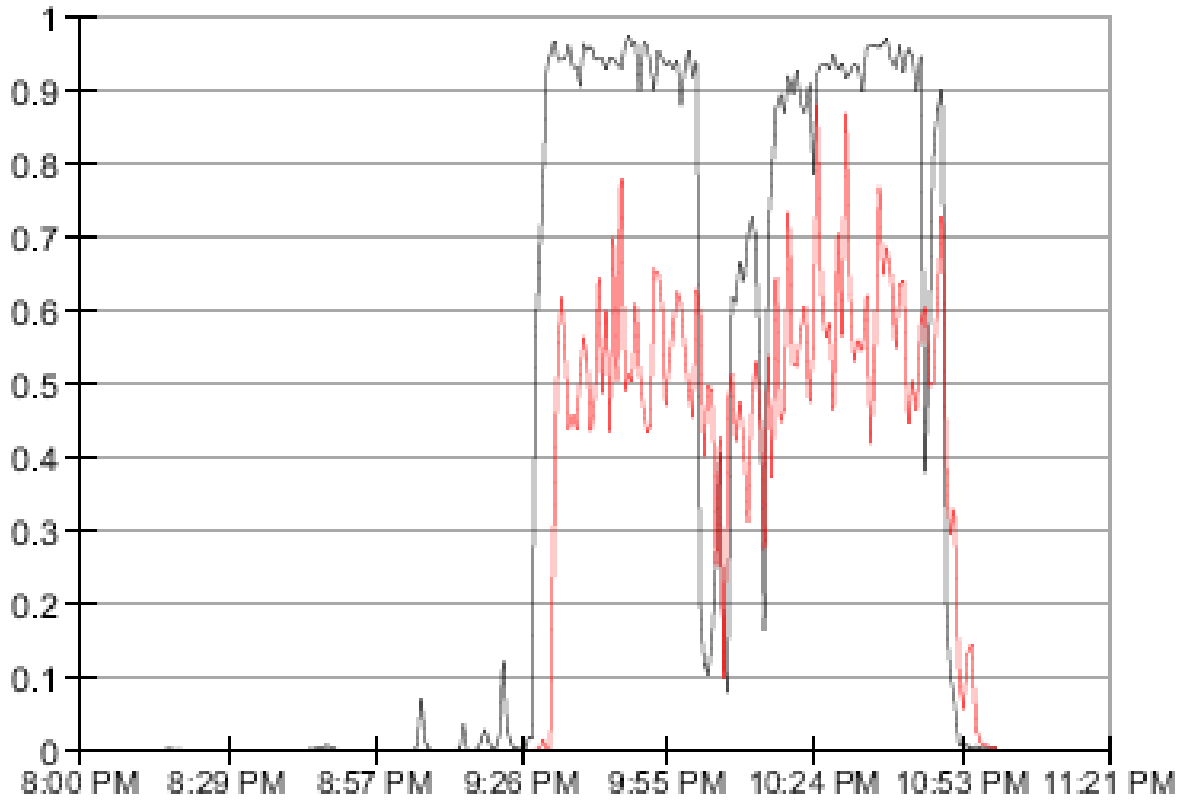


Figure 4: The inferred fault rate for two Autonomous Systems, as a function of time. These are the only two faults with high rate.

Minimising the other KL

If we could, we would rather minimize the other KL

$$KL(p,q) = \int d\mathbf{x}\; p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{q(\mathbf{x})} = \text{const} - \int d\mathbf{x}\; p(\mathbf{x}) \ln q(\mathbf{x})$$

If $q(\mathbf{x}) = \prod_i q_i(x_i)$, we have to minimize

$$-\sum_i \int dx\; p_i(x) \ln q_i(x)$$

which is minimized by the true marginal $q_i = p_i$.

On the other hand for exponential families

$$q(x|\boldsymbol{\theta}) = f(x)\; \exp[\boldsymbol{\psi}(\boldsymbol{\theta}) \cdot \boldsymbol{\phi}(x) + g(\boldsymbol{\theta})]\;.$$

we see that the optimal $\boldsymbol{\psi}$ is such that general moments match $\langle \boldsymbol{\phi}(\mathbf{x}) \rangle_q = \langle \boldsymbol{\phi}(\mathbf{x}) \rangle_p$.

We next try to do this procedure approximately in an on–line algorithm.

Bayes Online (Assumed Density Filtering)

Exact update of the posterior, when new data y_{t+1} arrives

$$p(\mathbf{x}|D_{t+1}) = \frac{p(y_{t+1}|\mathbf{x})p(\mathbf{x}|D_t)}{\int d\mathbf{x} p(y_{t+1}|\mathbf{x})p(\mathbf{x}|D_t)}.$$

Replace $p(\mathbf{x}|D_t)$ by parametric approximation $q(\mathbf{x}|par(t))$ using the following steps:

- Update:

$$q(\mathbf{x}|y_{t+1}, par(t)) = \frac{p(y_{t+1}|\mathbf{x})p(\mathbf{x}|par(t))}{\int d\mathbf{x} p(y_{t+1}|\mathbf{x})q(\mathbf{x}|par(t))}.$$

- Project: Minimize

$$KL\left(q(\cdot|y_{t+1}, par(t))||q(\cdot|par)\right)$$

  For exponential families $p(\mathbf{x}|par) \propto \exp[\boldsymbol{\psi} \cdot \boldsymbol{\phi}(\mathbf{x})]$, we have $par = \boldsymbol{\psi}$. The projection leads to moment matching $\langle \boldsymbol{\phi}(\mathbf{x}) \rangle$ for the distributions $q(\mathbf{x}|par)$ and $q(\mathbf{x}|y_{t+1}, par(t))$.

Gaussian Approximation

for $q(\mathbf{x}|par)$. Set $par = (mean, covariance) = (\hat{\mathbf{x}}, \mathbf{C})$. Matching of moments results in the explicit update:

$$\begin{aligned} \hat{\mathbf{x}}(t+1) &= \hat{\mathbf{x}}(t) + \sum_j C_{ij}(t) \times \\ &\quad \times \partial_j \ln E_u[p(y_{t+1}|\hat{\mathbf{x}}(t)+u)] \end{aligned}$$

and

$$\begin{aligned} C_{ij}(t+1) &= C_{ij}(t) + \sum_{kl} C_{ik}(t) C_{lj}(t) \times \\ &\quad \times \partial_k \partial_l \ln E_u[p(y_{t+1}|\hat{\mathbf{x}}(t)+u)]. \end{aligned}$$

with $\partial_j \doteq \frac{\partial}{\partial \hat{x}_j}$.
$\int d\mathbf{x} p(y_{t+1}|\mathbf{x}) q(\mathbf{x}|par(t))$ was written as $E_u[p(y_{t+1}|\hat{\mathbf{x}}(t)+u)]$ where u is a zero mean Gaussian random vector with covariance $\mathbf{C}(t)$.

Asymptotic Error

Assume data are generated from **true density** $p^*(y)$

$$E_D[\epsilon_i(t)\epsilon_j(t)] = \frac{1}{t}\ (\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})_{ij}, \qquad t \to \infty.$$

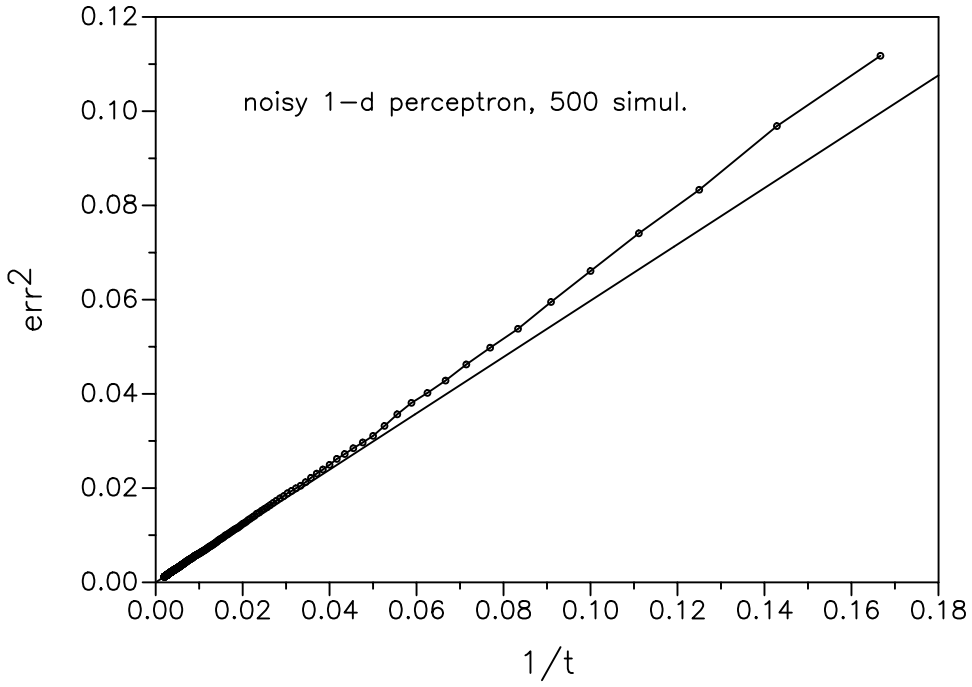
with

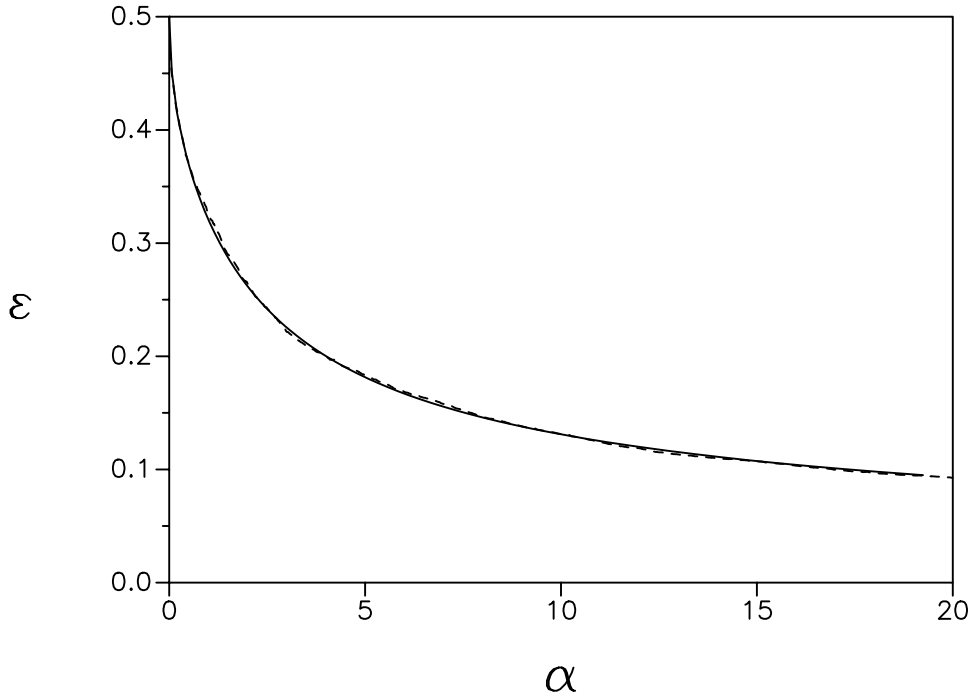
$$\begin{aligned} B_{ij} &= \int dy\ p^*(y)\partial_i \ln p(y|\mathbf{x}^*)\partial_j \ln p(y|\mathbf{x}^*) \\ A_{ij} &= -\int dy\ p^*(y)\partial_i\partial_j \ln p(y|\mathbf{x}^*). \end{aligned}$$

The same rate as for batch algorithms (Max. Likelihood or Bayes) : Asymptotic Efficiency!

Toy applications: Perceptrons

Results for $d = 1$ and $d = 50$ (probit model, spherical Gaussian inputs, realizable target, $\alpha \doteq \frac{\#\text{data}}{d}$). Right–dashed line: Bayes optimal (batch).





Method Kernelizes (Csato & Opper) also basis of *Informative Vector Machine* (Lawrence et al).