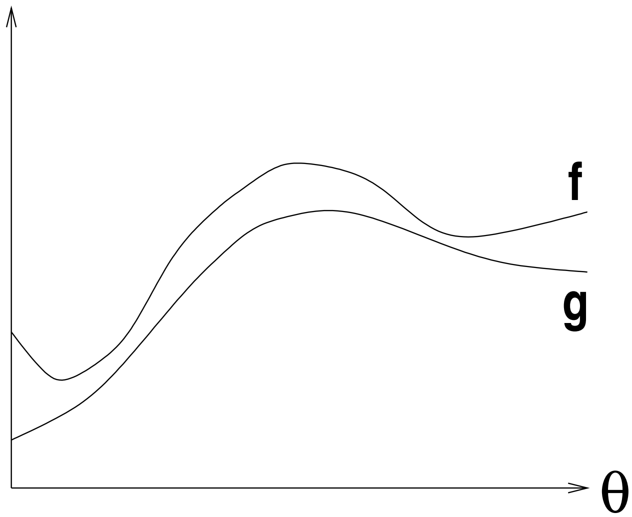
# Variational Bounds via Reversing EM

Thomas Minka

Just Research

# Variational bounds for integration




$$\int f(\theta)d\theta \geq \int g(\theta)d\theta$$

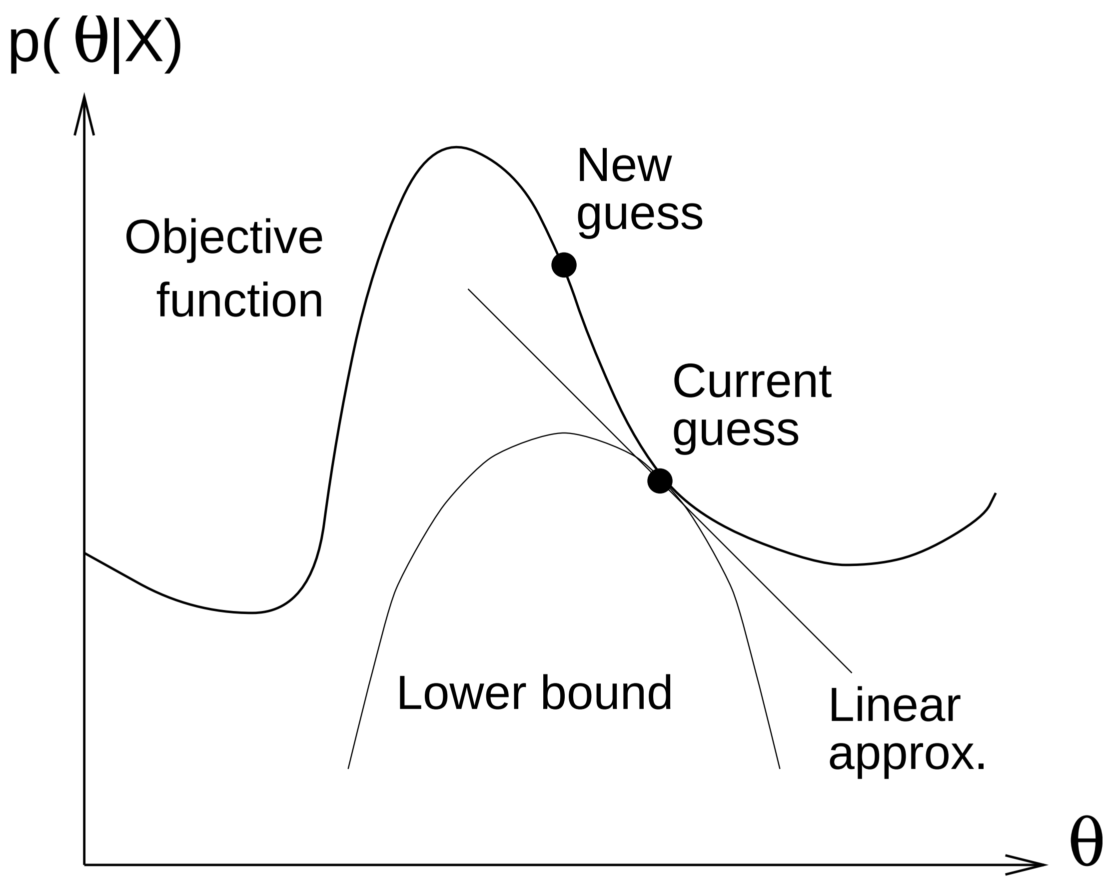
Integrate out parameters in order to:

- select hyperparameters
  (MacKay, 1995)
- select size of model
  (Attias, 1999)
- predict or classify new data
  (Waterhouse et al, 1996) (Jaakkola and Jordan, 1999)
- remove nuisance parameters
  (this talk)

# Outline

- EM as a variational bounding method
- Reversing EM to perform integration
- Faster EM -> Better bounds
- Language modeling example

# The EM Algorithm




E-step: Fit lower bound at current theta

M-step: Maximize the bound over theta
(Guaranteed to improve theta)

# EM for mixture weights

Maximize over w:

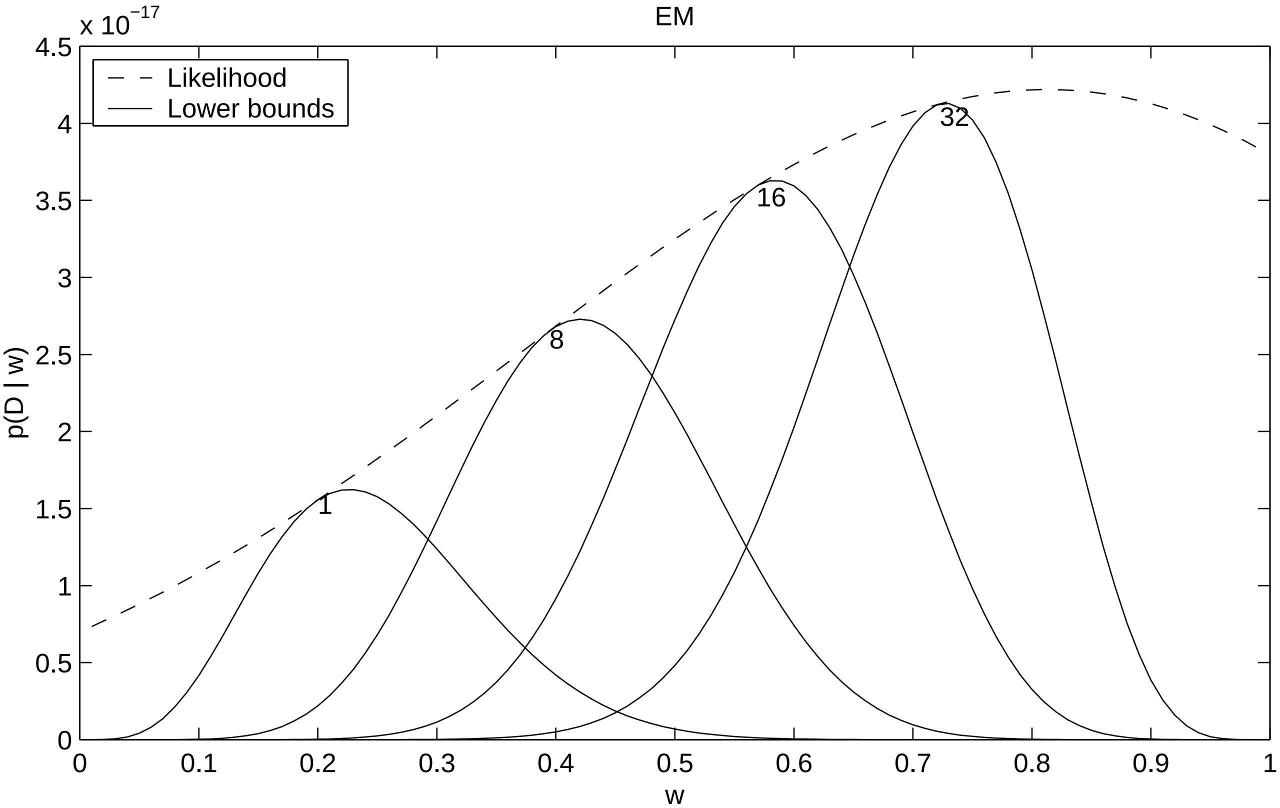
$$p(x_1..x_N|w) = \prod_i (p_1(x_i)w + p_2(x_i)(1-w))$$

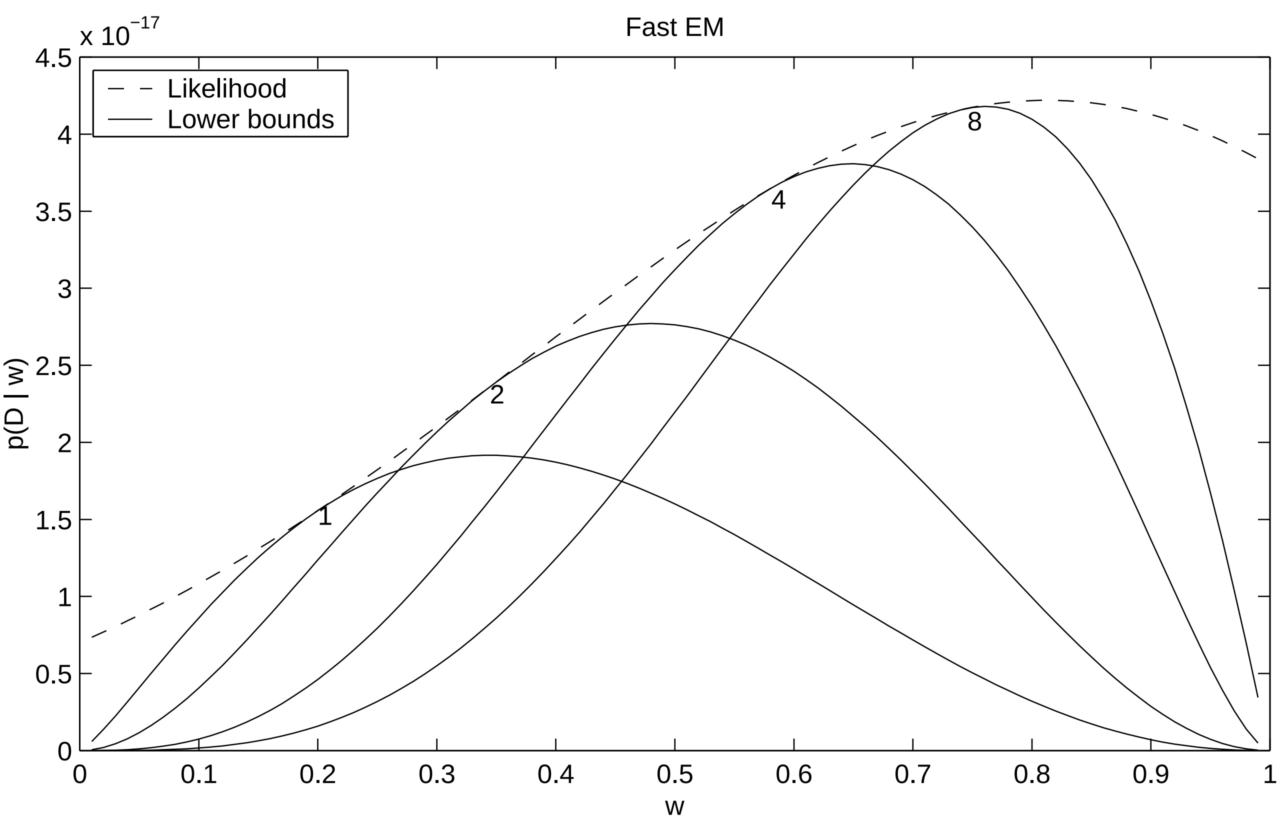
The bound:

$$\begin{aligned} p(x|w) &= p_1(x)w + p_2(x)(1-w) \\ &\geq \left(\frac{p_1(x)w}{q}\right)^q \left(\frac{p_2(x)(1-w)}{1-q}\right)^{1-q} = g(x|w;q) \end{aligned}$$

E-step: $$q_i = \frac{p_1(x_i)w}{p_1(x_i)w + p_2(x_i)(1-w)}$$

M-step: $$w = \frac{\sum_i q_i}{\sum_i q_i + N - \sum_i q_i} = \frac{1}{N}\sum_i q_i$$


EM
x 10^-17
Likelihood
Lower bounds
1
8
16
32
p(D | w)
w
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
1.5
2
2.5
3
3.5
4
4.5


Fast EM
x 10^-17
Likelihood
Lower bounds
1
2
4
8
p(D | w)
w
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
1.5
2
2.5
3
3.5
4
4.5

# Reversing EM to maximize bound area

Want to compute

$$
\begin{aligned}
p(x_1..x_N) &= \int \prod_i p(x_i|w)dw \\
&\geq \int \prod_i g(x_i|w;q_i)dw = g(x_1..x_N;q)
\end{aligned}
$$

Maximize the bound area with EM

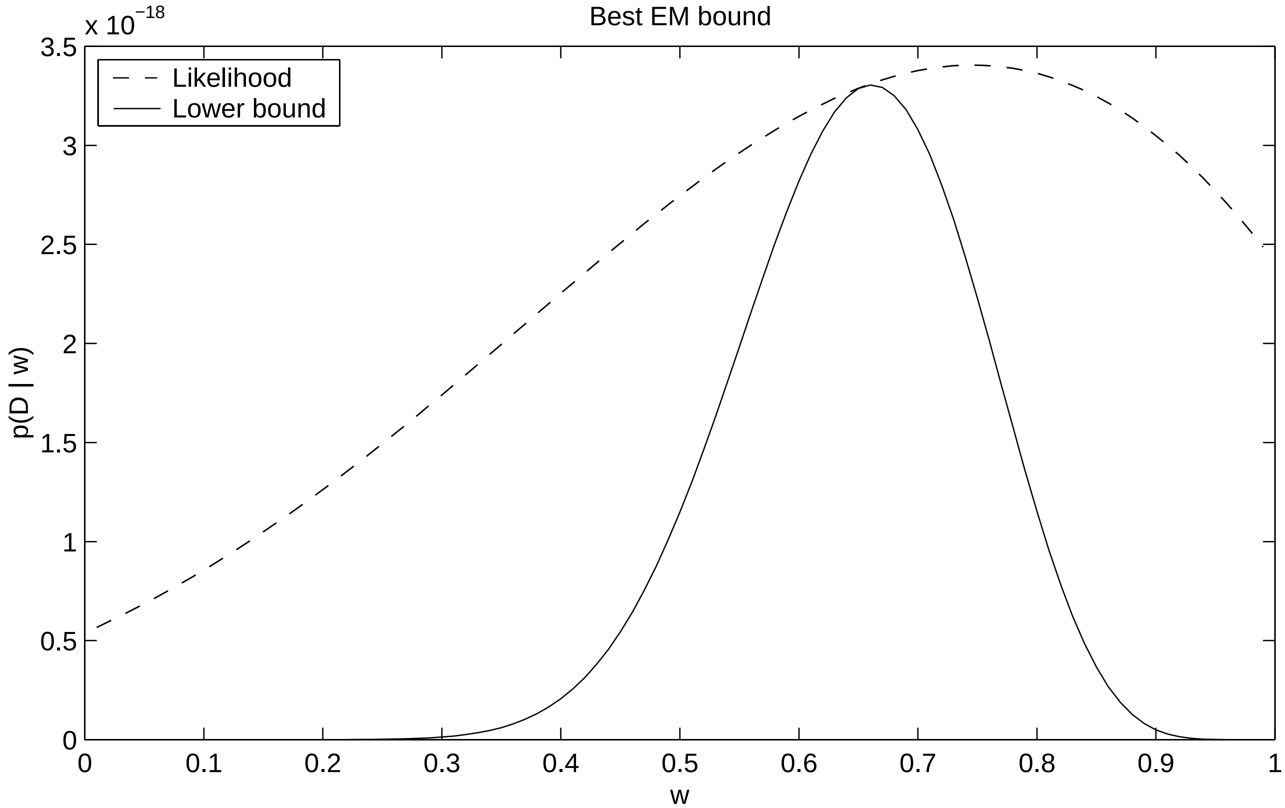
(maximize over q with w hidden)

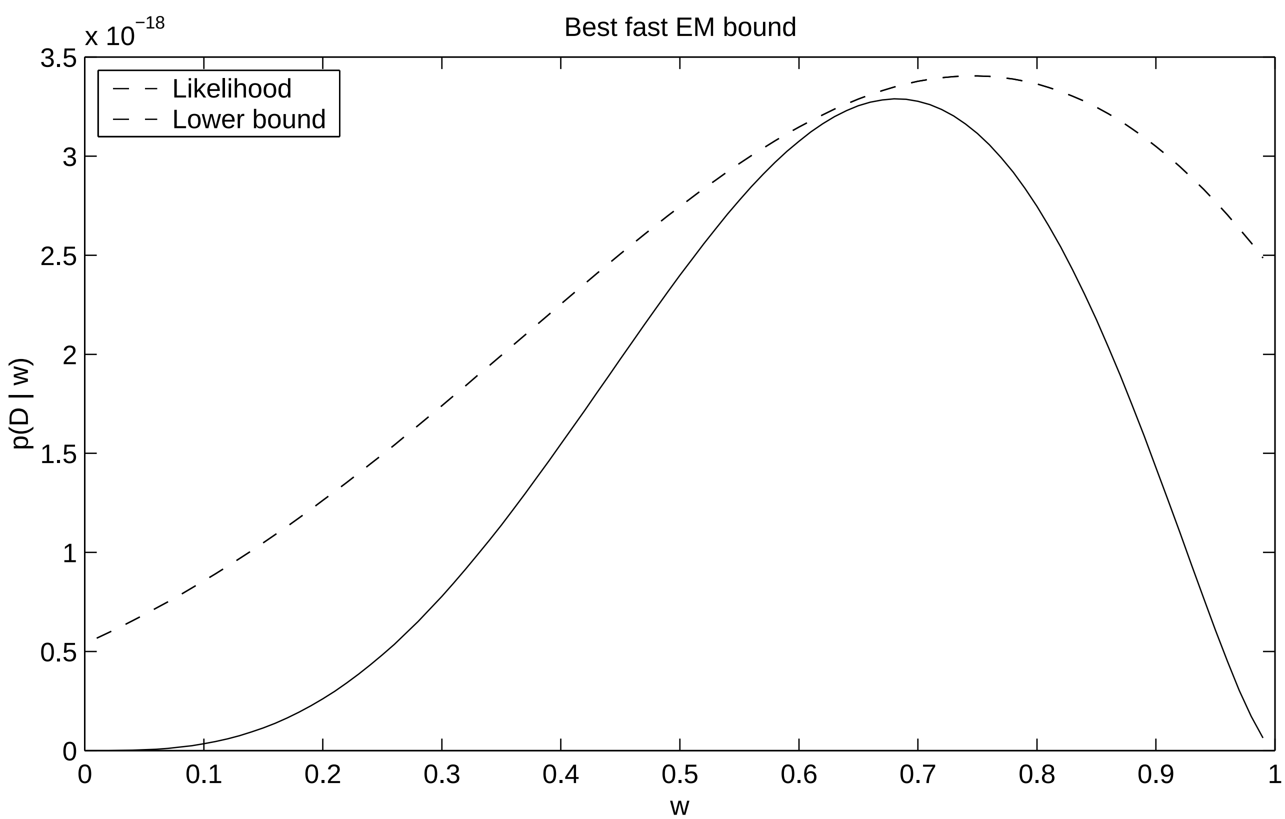
E-step:

$$
\begin{aligned}
p(w|q,D) &\sim \mathcal{B}(\alpha_1,\alpha_2) \\
\alpha_1 &= 1 + \sum_i q_i \\
\alpha_2 &= 1 + N - \sum_i q_i
\end{aligned}
$$

M-step:

$$
q_i = \frac{p_1(x_i)\exp(\Psi(\alpha_1))}{p_1(x_i)\exp(\Psi(\alpha_1)) + p_2(x_i)\exp(\Psi(\alpha_2))}
$$


Best EM bound
x 10^-18
3.5
3
2.5
2
1.5
1
0.5
0
p(D | w)
Likelihood
Lower bound
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
w


Best fast EM bound
x 10^-18
3.5
3
2.5
2
1.5
1
0.5
0
p(D | w)
Likelihood
Lower bound
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
w

## Faster EM -> Better integral approximation

Faster EM -> less curved bound (Dempster et al, 1977)

-> greater bound area

Fast EM algs now known for:

- Mixture weights (Pilla & Lindsay, 1996)
- Student T estimation (Meng & van Dyk, 1997)
- Poisson imaging model "
- Random effects regression (Liu, Rubin, Wu, 1998)
- Factor analysis "
- Probit regression "

# Pilla & Lindsay's algorithm

The bound:

$$
\begin{aligned}
p(x|w) &= (p_1(x) - p_2(x))w + p_2(x) \\
&\geq \left(\frac{(p_1(x) - p_2(x))w}{q_1}\right)^{q_1} \left(\frac{p_2(x)}{1-q_1}\right)^{1-q_1} \\
\text{or} &\quad \left(\frac{p_1(x)}{1-q_2}\right)^{1-q_2} \left(\frac{(p_2(x) - p_1(x))(1-w)}{q_2}\right)^{q_2}
\end{aligned}
$$

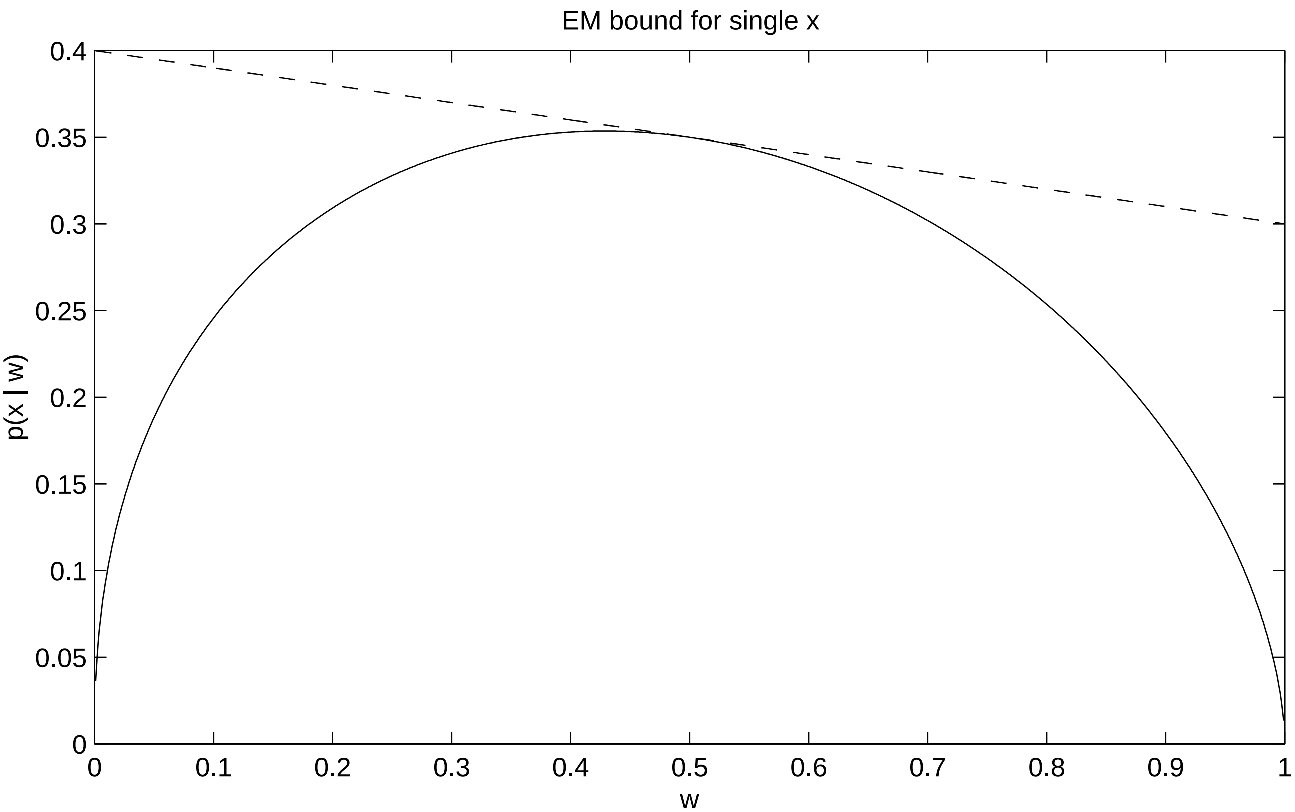
E-step:

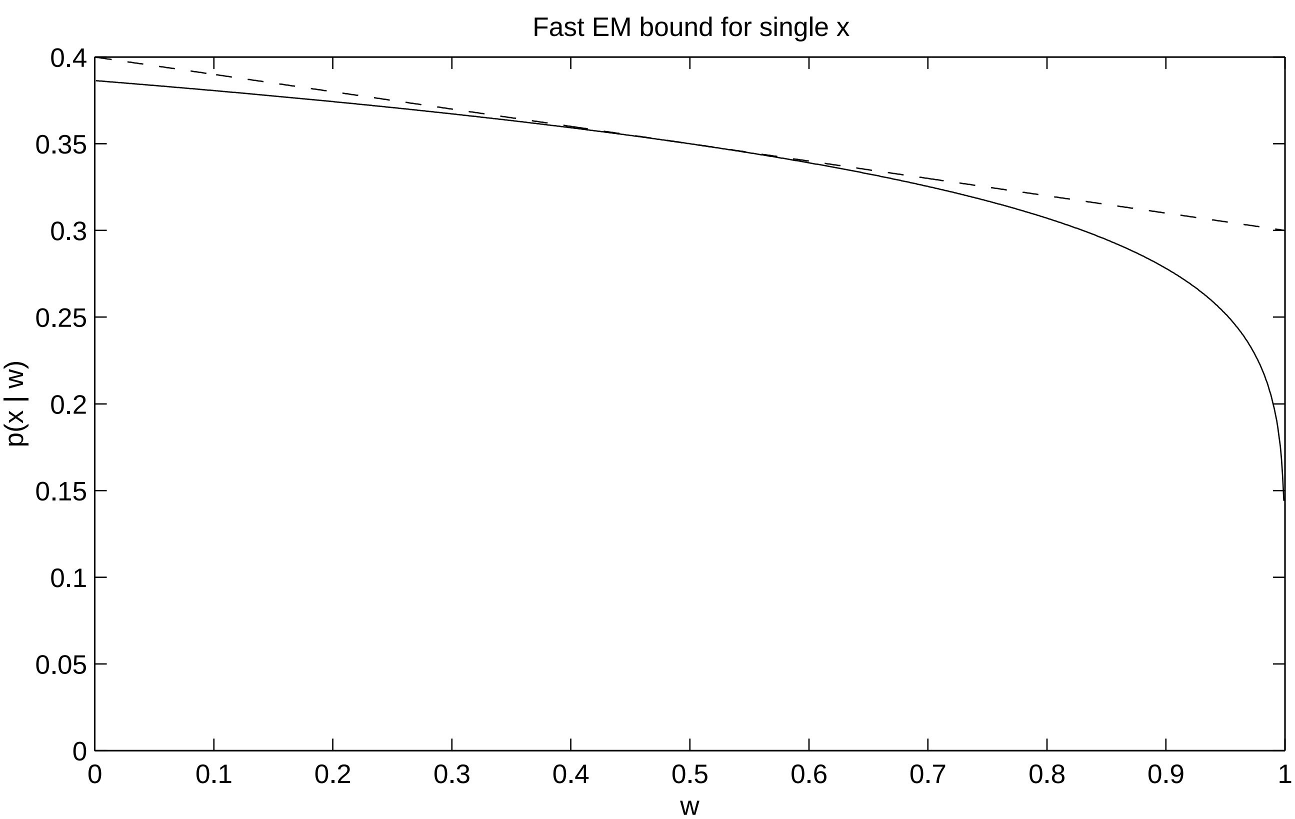
$$
\begin{aligned}
q_{i1} &= \frac{(p_1(x_i) - p_2(x_i))w}{p_1(x_i)w + p_2(x_i)(1-w)} \\
q_{i2} &= 0
\end{aligned}
$$

or

$$
\begin{aligned}
q_{i1} &= 0 \\
q_{i2} &= \frac{(p_2(x_i) - p_1(x_i))(1-w)}{p_1(x_i)w + p_2(x_i)(1-w)}
\end{aligned}
$$

M-step: $$w = \frac{\sum_i q_{i1}}{\sum_i q_{i1} + \sum_i q_{i2}}$$


EM bound for single x
p(x | w)
w
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


Fast EM bound for single x
p(x | w)
w
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

# Reverse of Pilla & Lindsay's algorithm

E-step:

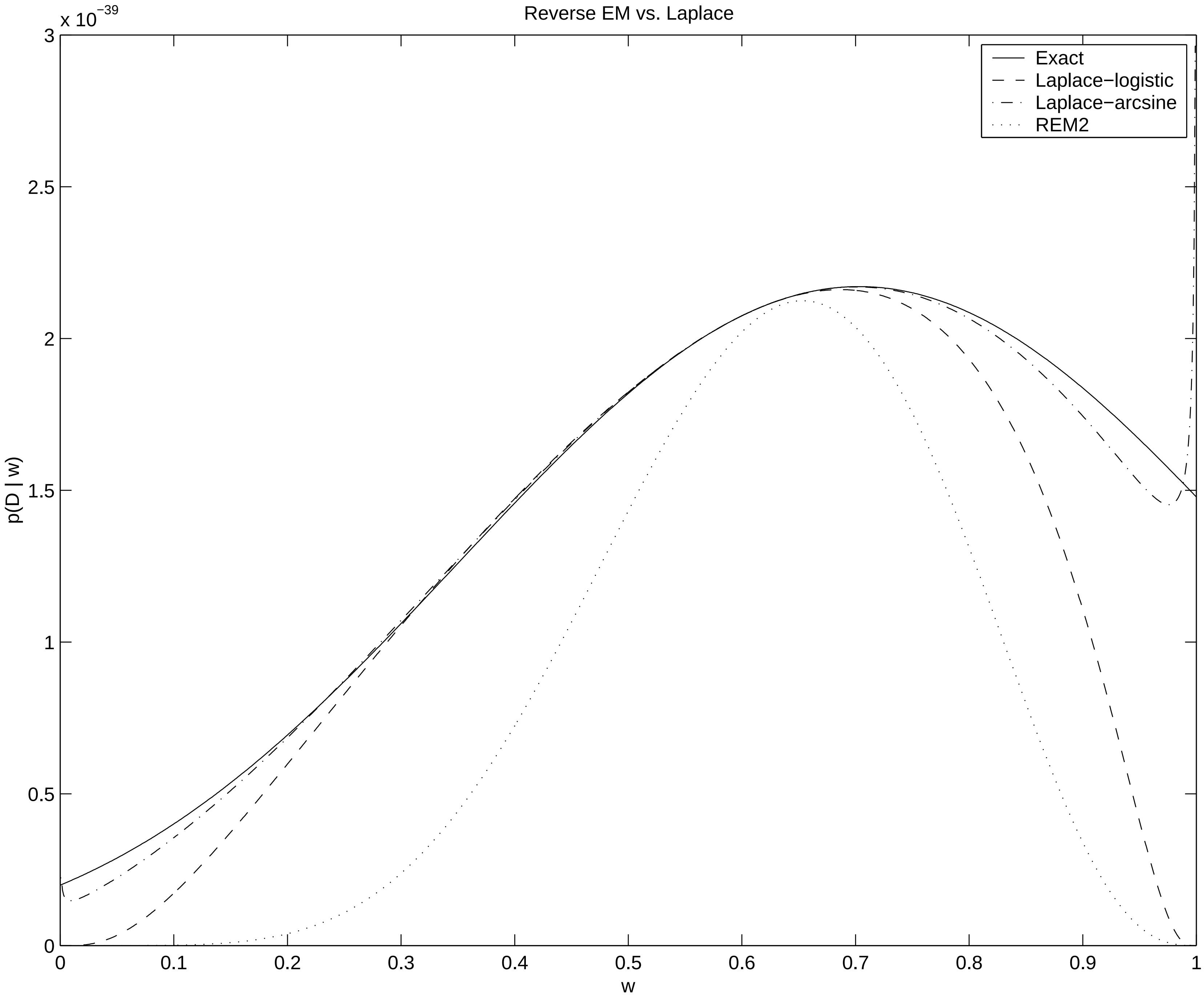
$$\begin{aligned} p(w|q,D) &\sim \mathcal{B}(\alpha_1,\alpha_2) \\ \alpha_1 &= 1+\sum_i q_{i1} \\ \alpha_2 &= 1+\sum_i q_{i2} \end{aligned}$$

M-step:

$$\begin{aligned} q_{i1} &= \frac{(p_1(x_i)-p_2(x_i))\bar{w}_i}{p_1(x_i)\bar{w}_i+p_2(x_i)(1-\bar{w}_i)} \\ q_{i2} &= 0 \\ \bar{w}_i &= \exp(\Psi(\alpha_1)-\Psi(\alpha_1+\alpha_2)) \end{aligned}$$

or

$$\begin{aligned} q_{i1} &= 0 \\ q_{i2} &= \frac{(p_2(x_i)-p_1(x_i))(1-\bar{w}_i)}{p_1(x_i)\bar{w}_i+p_2(x_i)(1-\bar{w}_i)} \\ \bar{w}_i &= 1-\exp(\Psi(\alpha_2)-\Psi(\alpha_1+\alpha_2)) \end{aligned}$$



Reverse EM vs. Laplace
x 10^−39
3
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
w
p(D | w)
Exact
Laplace−logistic
Laplace−arcsine
REM2

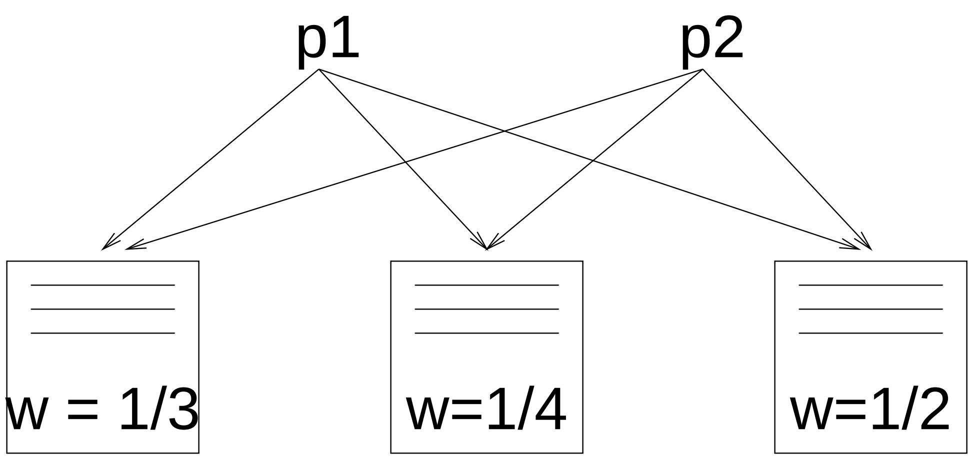
## Language modeling example

(aspect model)

Every document has its own word distribution:

a mixture of fixed multinomials

(Hofmann, 1998)




Want to estimate the mixture components

$\longrightarrow$ mixing weights are nuisance parameters

# Simplified model

Two-word case:

$$\begin{aligned} p(x|w, p_1, p_2) &= p_1^x (1-p_1)^{1-x} w + p_2^x (1-p_2)^{1-x} (1-w) \\ p(d = \{x_1..x_N\}|p_1, p_2) &= \int_0^1 \prod_i p(x_i|w, p_1, p_2) dw \end{aligned}$$

Maximize over p1,p2:

$$p(d_1..d_M|p_1, p_2) = \prod_j p(d_j|p_1, p_2)$$

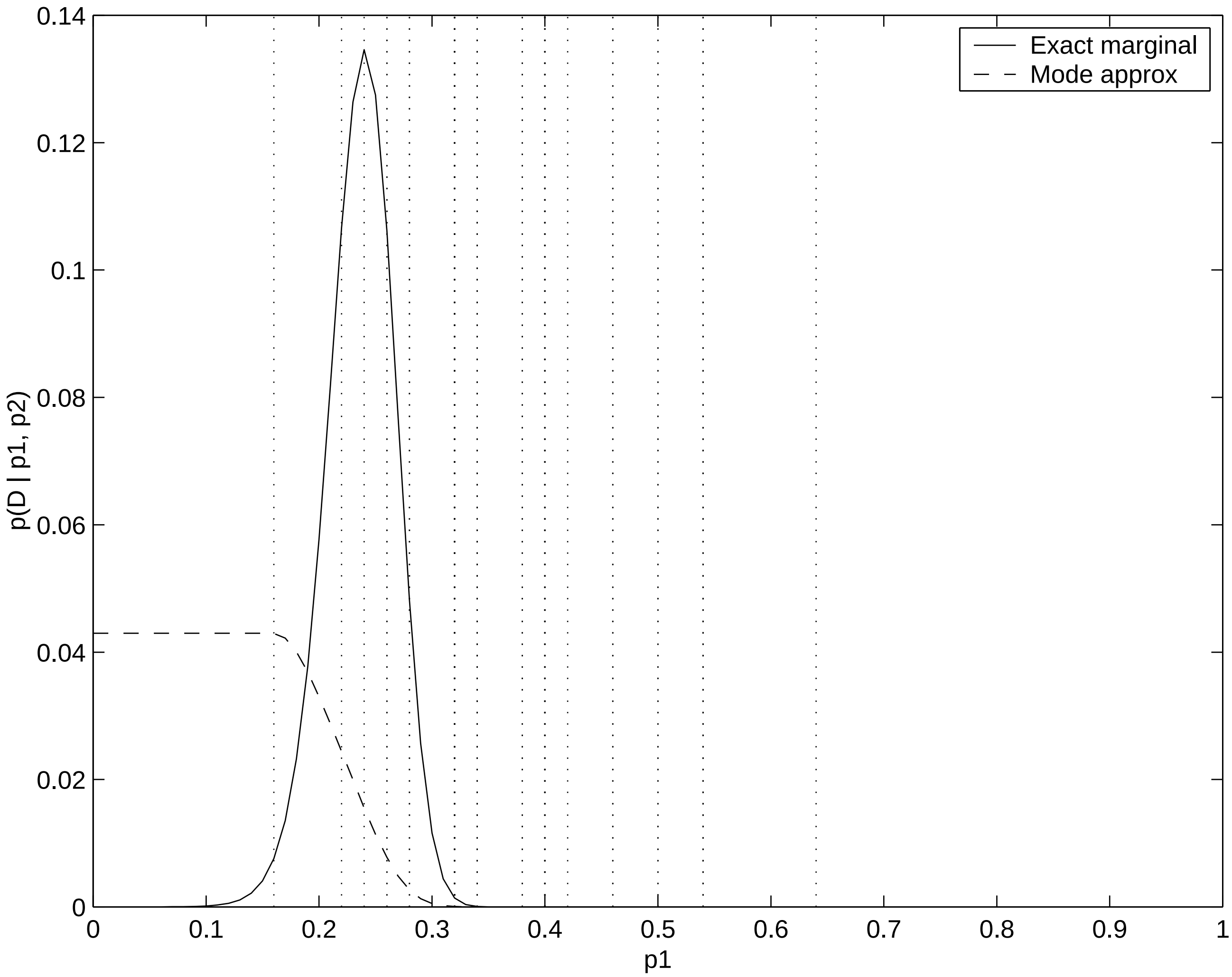
NOT joint maximum of (p1, p2, w1, ..., wM)

(Hofmann)

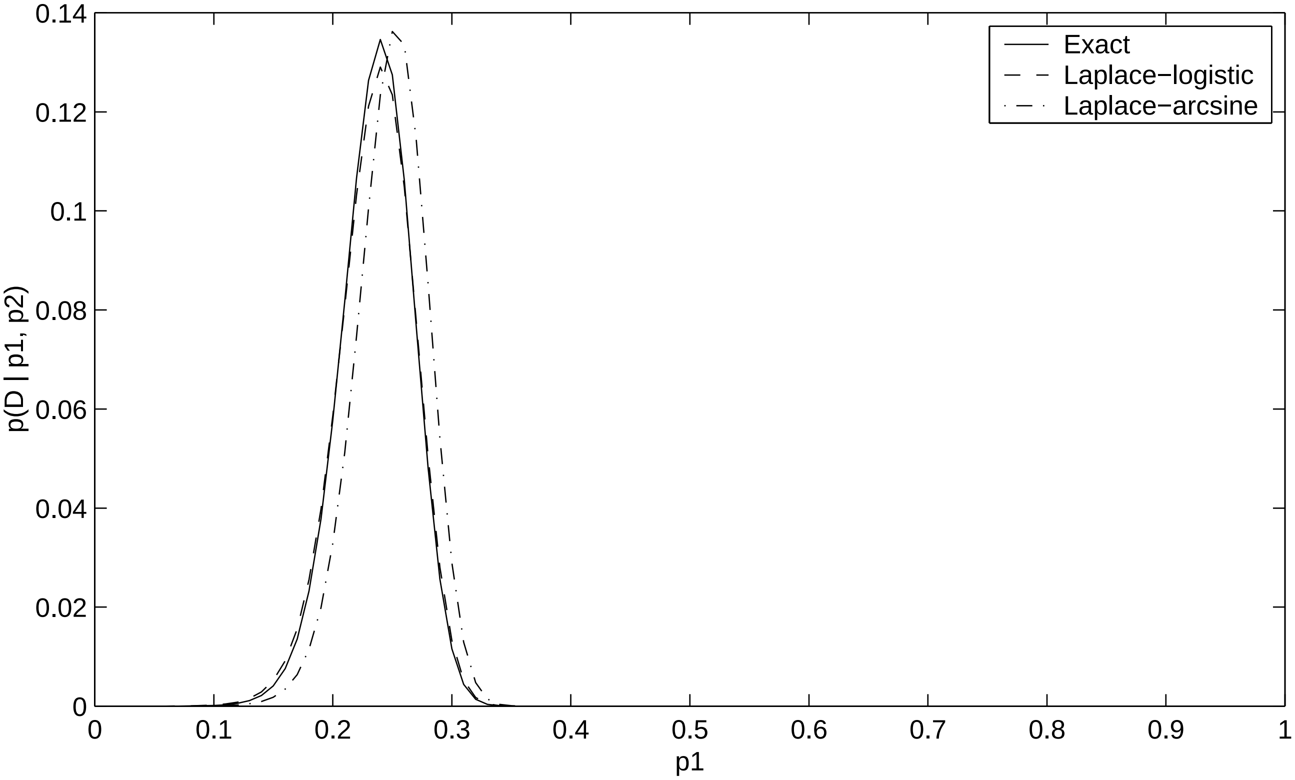
# Case 1

M = 30 docs, N = 50 words/doc

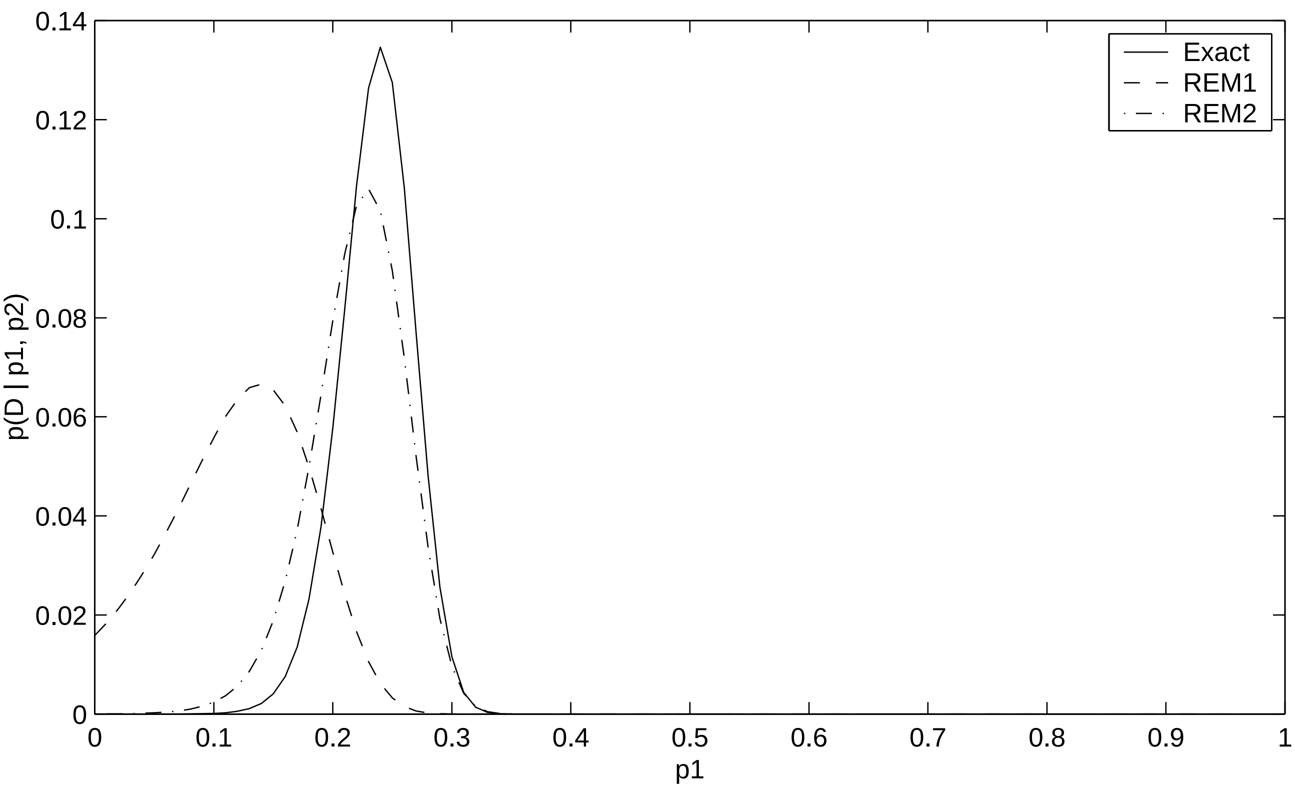
True p1 = 1/4 p2 fixed at 1/2



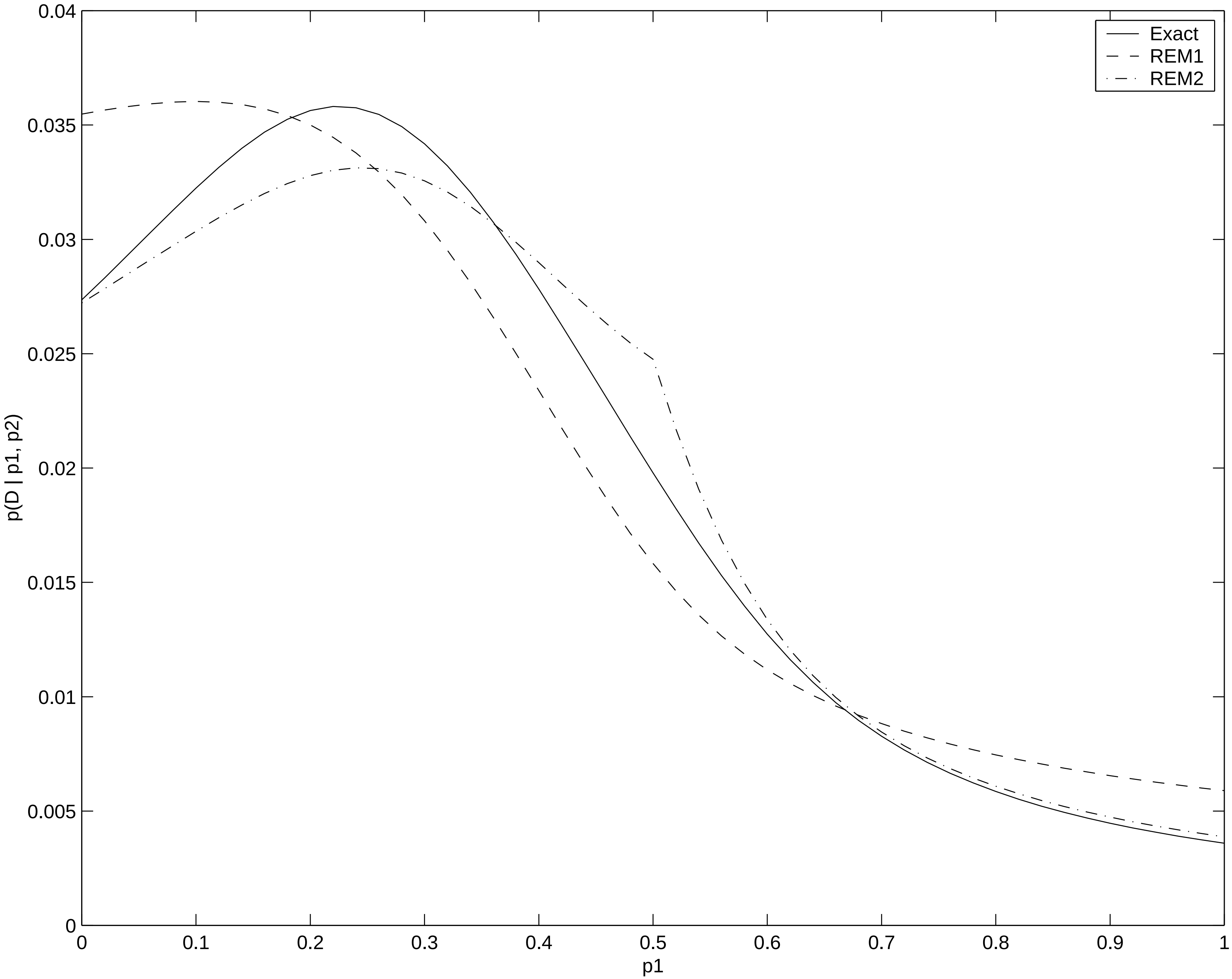

Dotted are empirical fractions: $\frac{1}{N}\sum_i x_i$



Exact
Laplace−logistic
Laplace−arcsine
p(D | p1, p2)
p1
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

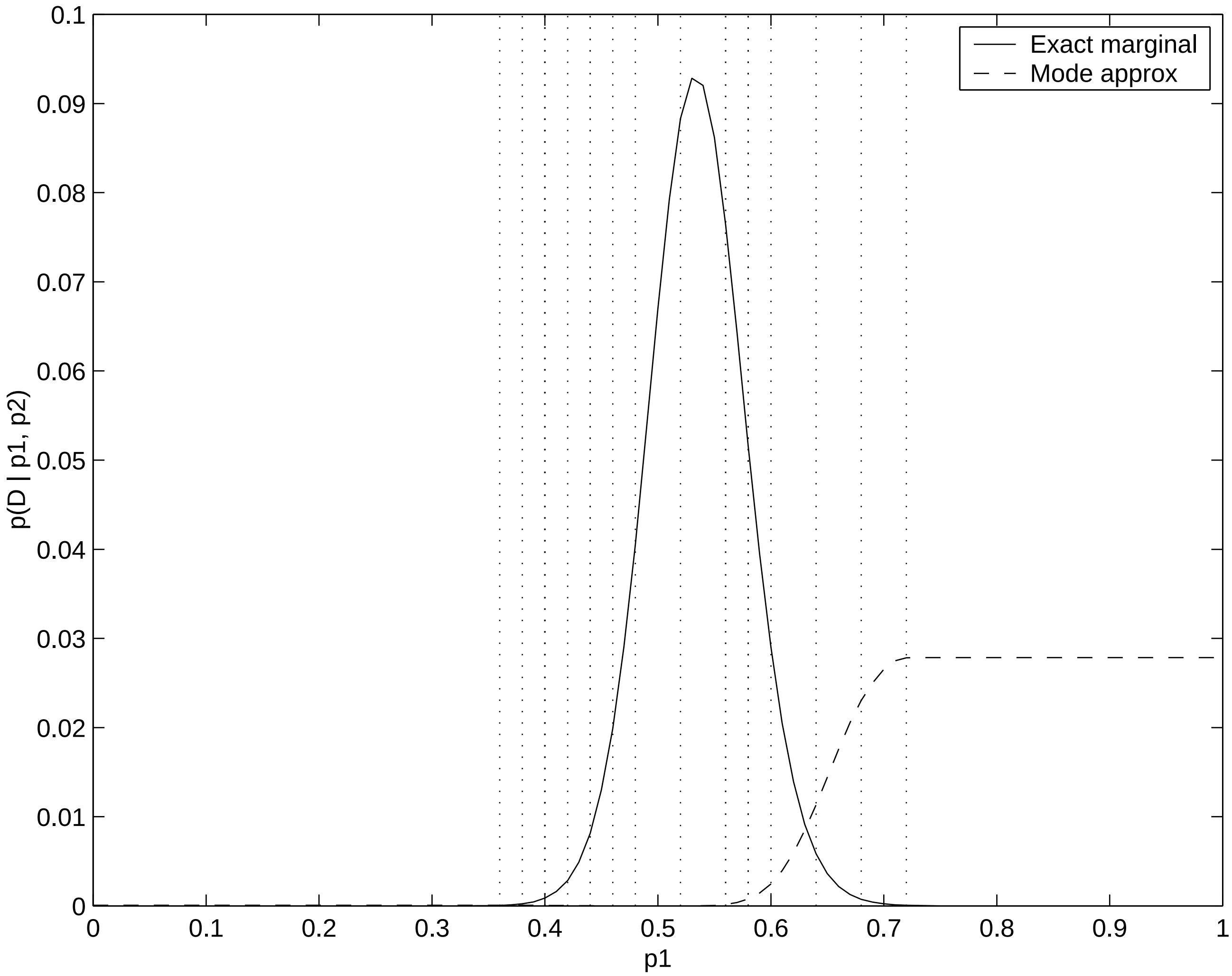
Exact
REM1
REM2
p(D | p1, p2)
p1
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

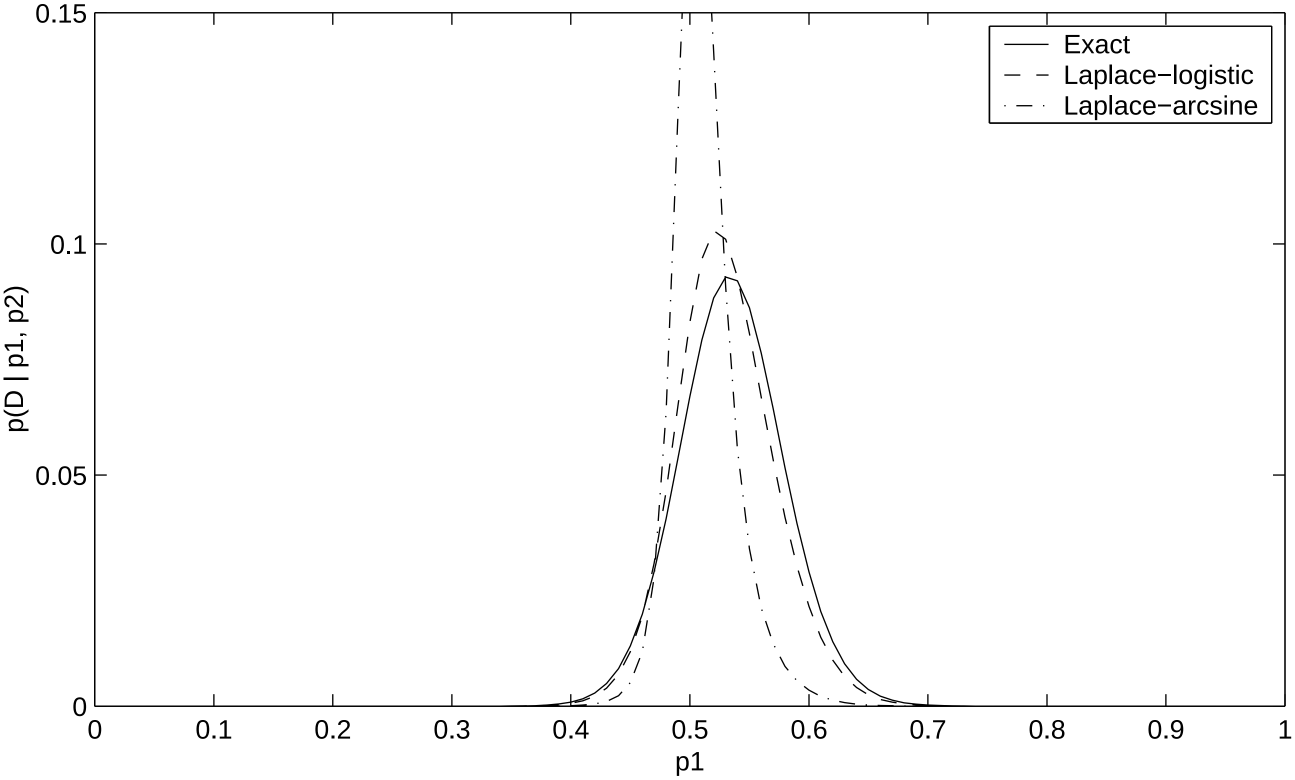

Exact
REM1
REM2
p(D | p1, p2)
p1
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

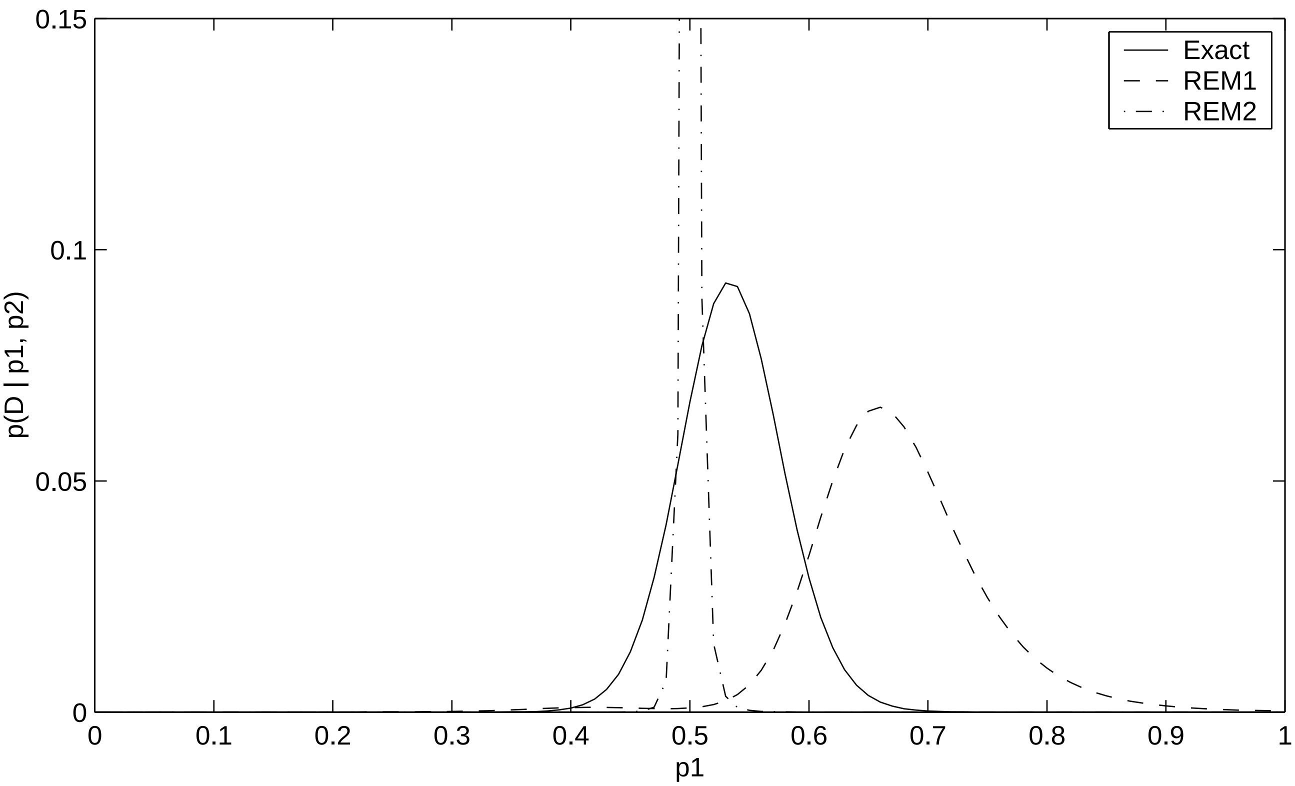
## Case 2

M = 20 docs, N = 50 words/doc

True p1 = 1/2 p2 fixed at 1/2




Exact
Laplace−logistic
Laplace−arcsine
p(D | p1, p2)
p1
0.15
0.1
0.05
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Exact
REM1
REM2
p(D | p1, p2)
p1
0.15
0.1
0.05
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1