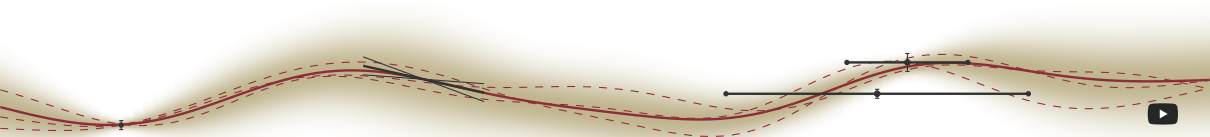
Probabilistic Machine Learning
Lecture 24
Customizing Probabilistic Models

Philipp Hennig

14 July 2020

Eberhard Karls
Universität
Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



#	date	content	Ex
1	20.04.	Introduction	1
2	21.04.	Reasoning under Uncertainty	
3	27.04.	Continuous Variables	2
4	28.04.	Monte Carlo	
5	04.05.	Markov Chain Monte Carlo	3
6	05.05.	Gaussian Distributions	
7	11.05.	Parametric Regression	4
8	12.05.	Learning Representations	
9	18.05.	Gaussian Processes	5
10	19.05.	Understanding Kernels	
11	26.05.	Gauss-Markov Models	
12	25.05.	An Example for GP Regression	6
13	08.06.	GP Classification	7

#	date	content	Ex
14	09.06.	Generalized Linear Models	
15	15.06.	Exponential Families	8
16	16.06.	Graphical Models	
17	22.06.	Factor Graphs	9
18	23.06.	The Sum-Product Algorithm	
19	29.06.	Example: Modelling Topics	10
20	30.06.	Mixture Models	
21	06.07.	EM	11
22	07.07.	Variational Inference	
23	13.07.	Tuning Inference Algorithms	12
24	14.07.	Kernel Topic Models	
25	20.07.	Outlook	
26	21.07.	Revision	

Designing a probabilistic machine learning method:

1. get the **data**
   - 1.1 try to collect as much meta-data as possible
2. build the **model**
   - 2.1 identify quantities and datastructures; assign names
   - 2.2 design a generative process (graphical model)
   - 2.3 assign (conditional) distributions to factors/arrows (use exponential families!)
3. design the **algorithm**
   - 3.1 consider conditional independence
   - 3.2 try standard methods for early experiments
   - 3.3 run unit-tests and sanity-checks
   - 3.4 identify bottlenecks, find customized approximations and refinements

The Toolbox

Framework:

$$\int p(x_1, x_2)\, dx_2 = p(x_1) \qquad p(x_1, x_2) = p(x_1 \mid x_2)p(x_2) \qquad p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Modelling:

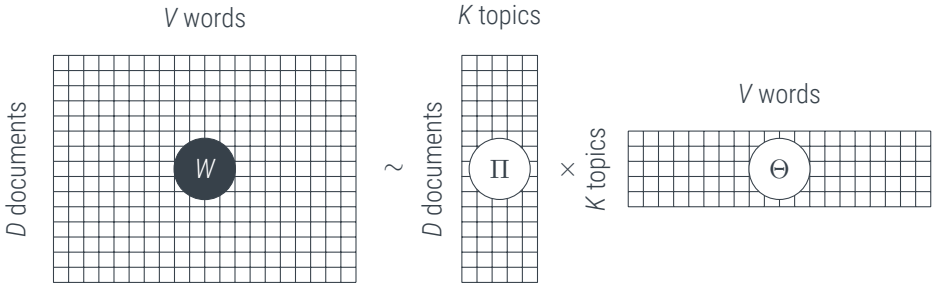
- graphical models
- Gaussian distributions
- (deep) learnt representations
- Kernels
- Markov Chains
- Exponential Families / Conjugate Priors
- Factor Graphs & Message Passing

Computation:

- Monte Carlo
- Linear algebra / Gaussian inference
- maximum likelihood / MAP
- Laplace approximations
- EM / variational approximations

Making Assumptions

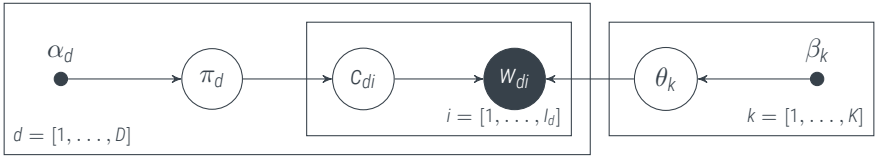
Our Data, our model



- a corpus of D documents
- each containing l_d words from a vocabulary of V words
- assumed to consist of K topics

Latent Dirichlet Allocation

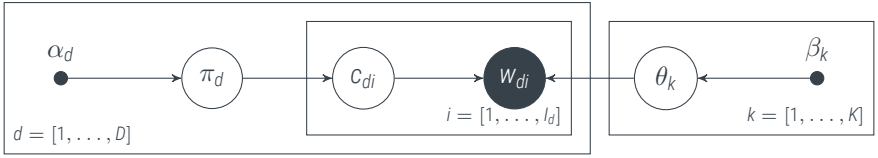
[Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) JMLR 3, 993–1022]



To draw I_d words $w_{di} \in [1, \ldots, V]$ of document $d \in [1, \ldots, D]$:

- Draw K topic distributions θ_k over V words from $p(\Theta \mid \boldsymbol{\beta}) = \prod_{k=1}^{K} \mathcal{D}(\theta_k; \beta_k)$
- Draw D document distributions over K topics from $p(\Pi \mid \boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{D}(\pi_d; \alpha_d)$
- Draw topic assignments c_{dik} of word w_{di} from $p(C \mid \Pi) = \prod_{i,d,k} \pi_{dk}^{c_{dik}}$
- Draw word w_{di} from $p(w_{di} = v \mid c_{di}, \Theta) = \prod_k \theta_{kv}^{c_{dik}}$

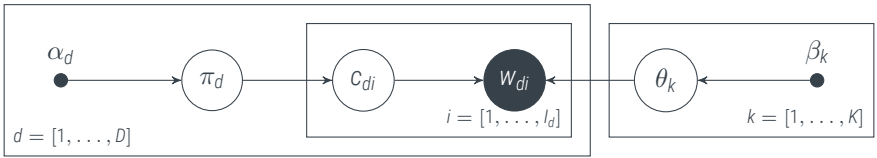
Useful notation: $n_{dkv} = \#\{i : w_{di} = v, c_{dik} = 1\}$. Write $n_{dk:} := [n_{dk1}, \ldots, n_{dkV}]$ and $n_{dk\cdot} = \sum_v n_{dkv}$, etc.



$$p(C, \Pi, \Theta, W) = \underbrace{\left(\prod_{d=1}^{D} \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d)\right)}_{p(\Pi|\boldsymbol{\alpha})} \cdot \underbrace{\left(\prod_{d=1}^{D}\prod_{i=1}^{l_d}\left(\prod_{k=1}^{K} \pi_{dk}^{c_{dik}}\right)\right)}_{p(C|\Pi)} \cdot \underbrace{\left(\prod_{d=1}^{D}\prod_{i=1}^{l_d}\left(\prod_{k=1}^{K} \theta_{kw_{di}}^{c_{dik}}\right)\right)}_{p(W|C,\Theta)} \cdot \underbrace{\left(\prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k)\right)}_{p(\Theta|\boldsymbol{\beta})}$$

$$= \underbrace{\left(\prod_{d=1}^{D} \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d)\right)}_{p(\Pi|\boldsymbol{\alpha})} \cdot \underbrace{\left(\prod_{d=1}^{D}\prod_{i=1}^{l_d}\left(\prod_{k=1}^{K} (\pi_{dk}\theta_{kw_{di}})^{c_{dik}}\right)\right)}_{p(W,C|\Theta,\Pi)} \cdot \underbrace{\left(\prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k)\right)}_{p(\Theta|\boldsymbol{\beta})}$$

$$= \left(\prod_{d=1}^{D} \frac{\Gamma(\sum_k \alpha_{dk})}{\prod_k \Gamma(\alpha_{dk})} \prod_{k=1}^{K} \pi_{dk}^{\alpha_{dk}-1+n_{dk\cdot}}\right) \cdot \left(\prod_{k=1}^{K} \frac{\Gamma(\sum_v \beta_{kv})}{\prod_v \Gamma(\beta_{kv})} \prod_{v=1}^{V} \theta_{kv}^{\beta_{kv}-1+n_{\cdot kv}}\right)$$



$$p(C, \Pi, \Theta, W) = \left(\prod_{d=1}^{D} \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d)\right) \cdot \left(\prod_{d=1}^{D}\prod_{i=1}^{I_d}\left(\prod_{k=1}^{K} \pi_{dk}^{c_{dik}}\right)\right) \cdot \left(\prod_{d=1}^{D}\prod_{i=1}^{I_d}\left(\prod_{k=1}^{K} \theta_{kw_{di}}^{c_{dik}}\right)\right) \cdot \left(\prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k)\right)$$

$$p(C \mid \Theta, \Pi, W) = \frac{p(W, C, \Theta, \Pi)}{\sum_C p(W, C, \Theta, \Pi)} = \prod_{d=1}^{D}\prod_{i=1}^{I_d} \frac{\prod_{k=1}^{K}(\pi_{dk}\theta_{kw_{di}})^{c_{dik}}}{\sum_{k'}(\pi_{dk'}\theta_{k'w_{di}})}$$

$$p(\Theta, \Pi \mid C, W) = \frac{p(C, W, \Pi, \Theta)}{\int p(\Theta, \Pi, C, W)\, d\Theta\, d\Pi} = \left(\prod_d \mathcal{D}(\pi_d; \alpha_{d:} + n_{d:.})\right)\left(\prod_k \mathcal{D}(\theta_k; \beta_{k:} + n_{\cdot k:})\right)$$

Our best Algorithm

collapsed variational inference

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Teh, Newman, Welling, NeurIPS 2007

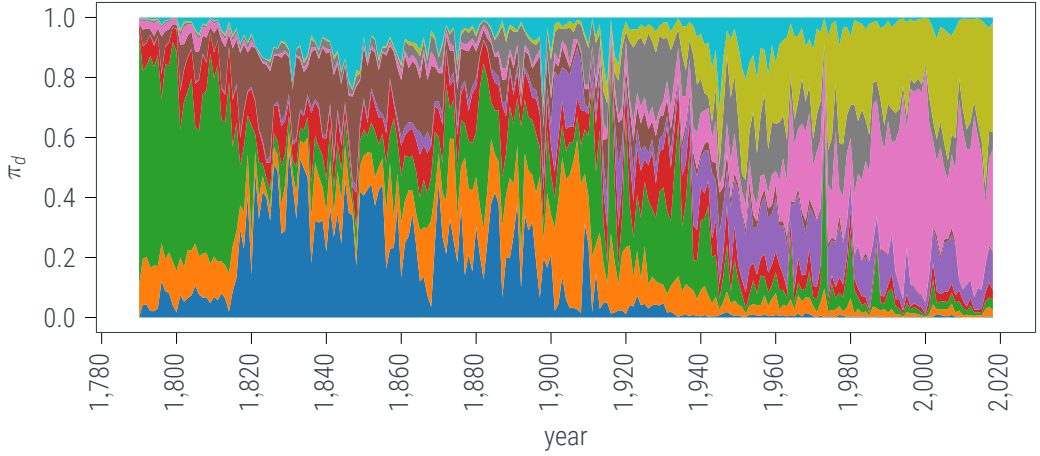
$$p(C) = \prod_d^D \prod_i^{I_d} \prod_k^K \gamma_{dik}$$

$$\gamma_{dik} \propto (\alpha_{dk} + \mathbb{E}[n_{dk.}^{\setminus di}])(\beta_{kw_{di}} + \mathbb{E}[n_{.kw_{di}}^{\setminus di}]) \left(\sum_v \beta_{kv} + \mathbb{E}[n_{.kv}^{\setminus di}] \right)^{-1}$$

$$\cdot \exp\left(- \frac{\mathrm{var}_q[n_{dk.}^{\setminus di}]}{2(\alpha_{dk} + \mathbb{E}_q[n_{dk.}^{\setminus di}])^2} - \frac{\mathrm{var}_q[n_{.kw_{di}}^{\setminus di}]}{2(\beta_{kw_{di}} + \mathbb{E}_q[n_{.kw_{di}}^{\setminus di}])^2} + \frac{\mathrm{var}_q[n_{.k.}^{\setminus di}]}{2(\sum_v \beta_{kv} + \mathbb{E}_q[n_{.kv}^{\setminus di}])^2} \right)$$

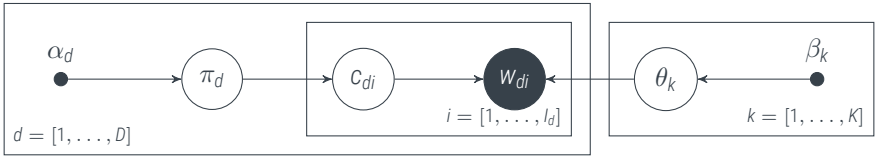
Some Output

can we be happy with this?



Can we be happy with this model?

Have we described all we know about the data?



$$p(C, \Pi, \Theta, W) = \underbrace{\left(\prod_{d=1}^{D} \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d)\right)}_{p(\Pi|\boldsymbol{\alpha})} \cdot \underbrace{\left(\prod_{d=1}^{D}\prod_{i=1}^{l_d}\left(\prod_{k=1}^{K}(\pi_{dk}\theta_{kw_{di}})^{c_{dik}}\right)\right)}_{p(W,C|\Theta,\Pi)} \cdot \underbrace{\left(\prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k)\right)}_{p(\Theta|\boldsymbol{\beta})}$$

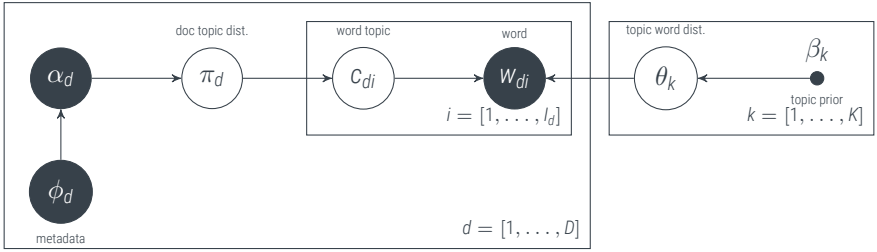
Meta-Data

It's right there!

Adams_1797.txt
Adams_1798.txt
Adams_1799.txt
Adams_1800.txt
Adams_1825.txt
Adams_1826.txt
Adams_1827.txt
Adams_1828.txt
Arthur_1881.txt
Arthur_1882.txt
Arthur_1883.txt
Arthur_1884.txt
Buchanan_1857.txt
Buchanan_1858.txt
Buchanan_1859.txt
Buchanan_1860.txt
Buren_1837.txt
Buren_1838.txt
Buren_1839.txt
Buren_1840.txt
Bush_1989.txt
Bush_1990.txt
Bush_1991.txt
Bush_1992.txt
Bush_2001.txt
Bush_2002.txt
Bush_2003.txt
Cleveland_1887.txt
Cleveland_1888.txt
Cleveland_1893.txt
Cleveland_1894.txt
Cleveland_1895.txt
Cleveland_1896.txt
Clinton_1993.txt
Clinton_1994.txt
Clinton_1995.txt
Clinton_1996.txt
Clinton_1997.txt
Clinton_1998.txt
Clinton_1999.txt
Clinton_2000.txt
Coolidge_1923.txt
Coolidge_1924.txt
Coolidge_1925.txt
Coolidge_1926.txt
Coolidge_1927.txt
Coolidge_1928.txt
Eisenhower_1954.txt
Eisenhower_1955.txt
Eisenhower_1956.txt
Eisenhower_1957.txt
Eisenhower_1958.txt
Eisenhower_1959.txt
Eisenhower_1960.txt
Grant_1873.txt
Grant_1874.txt
Grant_1875.txt
Grant_1876.txt
Harding_1921.txt
Harding_1922.txt
Harrison_1889.txt
Harrison_1890.txt
Harrison_1891.txt
Harrison_1892.txt
Hayes_1877.txt
Hayes_1878.txt
Hayes_1879.txt
Hayes_1880.txt
Hoover_1929.txt
Hoover_1930.txt
Hoover_1931.txt
Hoover_1932.txt
Jackson_1829.txt
Jackson_1830.txt
Jackson_1831.txt
Jackson_1832.txt
Jackson_1833.txt
Jackson_1834.txt
Jackson_1835.txt
Jackson_1836.txt
Jefferson_1801.txt
Johnson_1964.txt
Johnson_1965.txt
Johnson_1966.txt
Johnson_1967.txt
Johnson_1968.txt
Johnson_1969.txt
Kennedy_1962.txt
Kennedy_1963.txt
Lincoln_1861.txt
Lincoln_1862.txt
Lincoln_1863.txt
Lincoln_1864.txt
Madison_1809.txt
Madison_1810.txt
Madison_1811.txt
Madison_1812.txt
Madison_1813.txt
Madison_1814.txt
Madison_1815.txt
Madison_1816.txt
McKinley_1897.txt
McKinley_1898.txt
McKinley_1899.txt
McKinley_1900.txt
Monroe_1817.txt
Monroe_1818.txt
Monroe_1819.txt
Obama_2010.txt
Obama_2011.txt
Obama_2012.txt
Obama_2013.txt
Obama_2014.txt
Obama_2015.txt
Obama_2016.txt
Pierce_1853.txt
Pierce_1854.txt
Pierce_1855.txt
Pierce_1856.txt
Polk_1845.txt
Polk_1846.txt
Polk_1847.txt
Polk_1848.txt
Reagan_1982.txt
Reagan_1983.txt
Reagan_1984.txt
Reagan_1985.txt
Reagan_1986.txt
Reagan_1987.txt
Reagan_1988.txt
Roosevelt_1901.txt
Roosevelt_1902.txt
Roosevelt_1903.txt
Roosevelt_1904.txt
Roosevelt_1905.txt
Roosevelt_1942.txt
Roosevelt_1943.txt
Roosevelt_1944.txt
Roosevelt_1945.txt
Taft_1909.txt
Taft_1910.txt
Taft_1911.txt
Taft_1912.txt
Taylor_1849.txt
Truman_1946.txt
Truman_1947.txt
Truman_1948.txt
Truman_1949.txt
Truman_1950.txt
Truman_1951.txt
Truman_1952.txt
Truman_1953.txt
Trump_2017.txt
Trump_2018.txt
Tyler_1841.txt
Tyler_1842.txt
Tyler_1843.txt
Tyler_1844.txt
Washington_1790.txt
Washington_1791.txt
Washington_1792.txt
Washington_1793.txt

Adding more Information

a model for document metadata

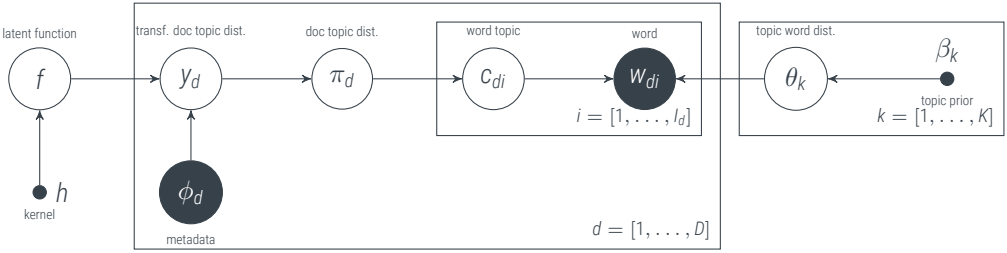


The Price of Packaged Solutions

Toolboxes speed up development, but also make it brittle

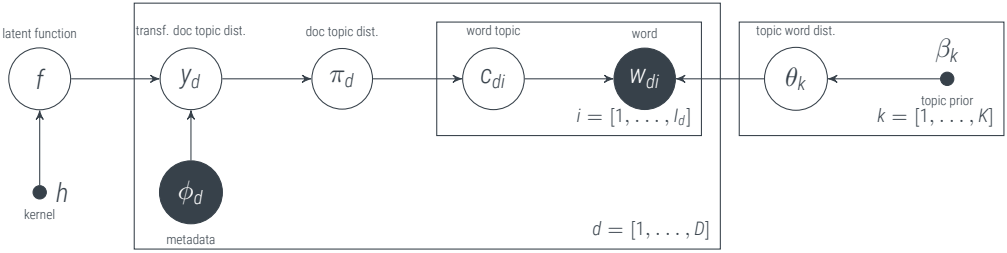
https://github.com/scikit-learn/scikit-learn/blob/fd237278e/sklearn/decomposition/_lda.py#L134

- toolboxes are extremely valuable for quick early development. Use them to your advantage!
- but their interface often enforces and restricts *model* design decisions
- to really *solve* a probabilistic modelling task, build your own *craftware*



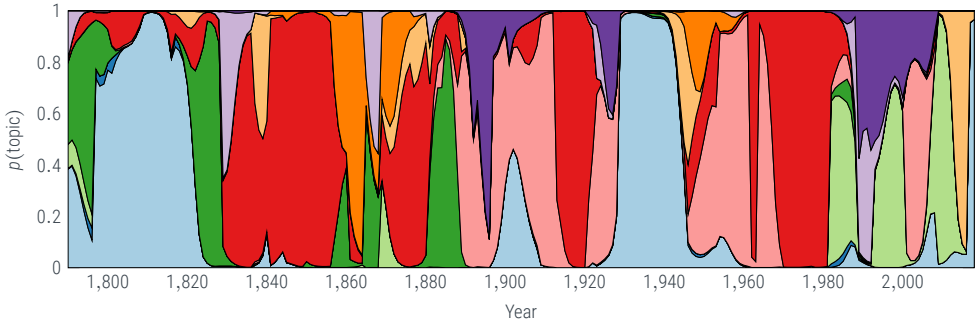
To generate the words W of documents $d = 1, \ldots, D$ with features $\phi_d \in \mathbb{F}$:

- draw function $f : \mathbb{F} \to \mathbb{R}^K$ from $p(f \mid h) = \mathcal{GP}(f; 0, h)$
- draw document topic distribution y_d from $p(y_d \mid f, \phi_d) = \mathcal{N}(y_d; f(\phi_d), \eta^2)$
- map through **softmax** to get $\pi_d = \sigma(y_d)$, with $[\sigma(x)]_k = \frac{e^{x_k}}{\sum_j e^{x_j}}$



To generate the words W of documents $d = 1, \ldots, D$ with features $\phi_d \in \mathbb{F}$:

- draw topic-word distributions $p(\Theta \mid \beta) = \prod_{k=1}^{K} \mathcal{D}(\theta_k, \beta_k)$
- draw each word's topic $p(C_{d::} \mid \Pi) = \prod_{d=1}^{D} \prod_{i=1}^{l_d} \prod_k \pi_{dk}^{c_{dik}}$
- draw the word w_{di} with probability $\theta_{kw_{di}}^{c_{dik}}$.

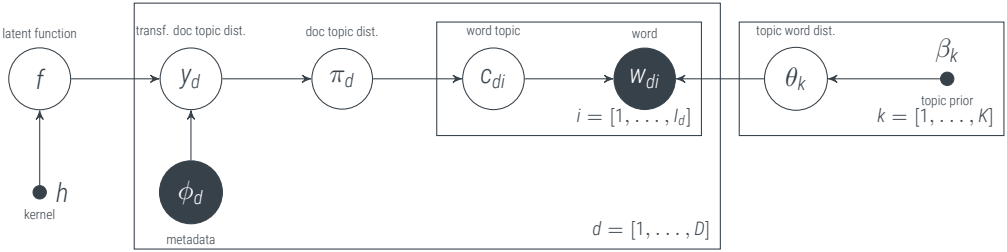


$$k(x_a, x_b) = \theta^2 \left(1 + \frac{(x_a - x_b)^2}{2\alpha\ell^2}\right)^{-\alpha} \cdot \begin{cases} 1.00 & \text{if } \texttt{president}(x_a) = \texttt{president}(x_b) \\ \gamma & \text{otherwise} \end{cases}$$

$$\theta = 5 \qquad \ell = 10\text{years} \qquad \alpha = 0.5 \qquad \gamma = 0.9$$

Algorithmic Considerations

efficient but feasible

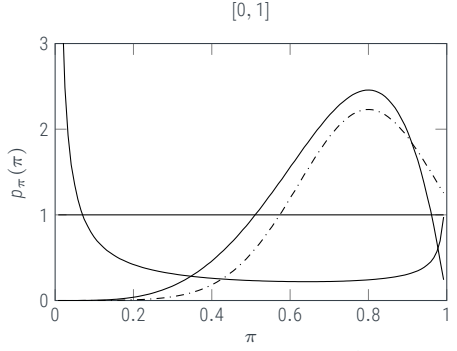


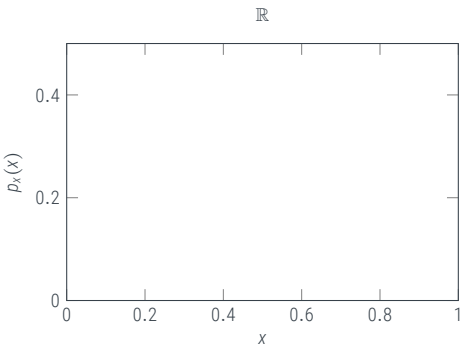
- We could construct a variational bound on (Y, Π, Θ, C), but it would need to factorize into many terms, thus be loose
- We know how to construct a **Dirichlet** on Π from W given a Dirichlet prior, but not how to incorporate ϕ into the Dirichlet!.
- We know how to infer a **Gaussian** on y_d given **Gaussian** observations of y_d, but not given Dirichlets on π

Linking Gaussians and Dirichlets

It's all about the right basis

DJC MacKay, *Choice of Basis for Laplace Approximation.* Machine Learning, **33** (1):77–86, 1998





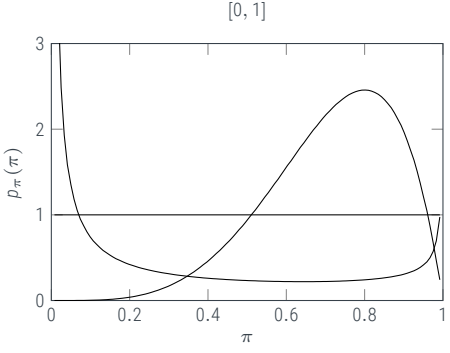
$$\mathcal{B}(\pi; a, b) = \frac{1}{B(a,b)} \pi^{a-1} \cdot (1-\pi)^{b-1} \qquad \pi \in [0,1]$$

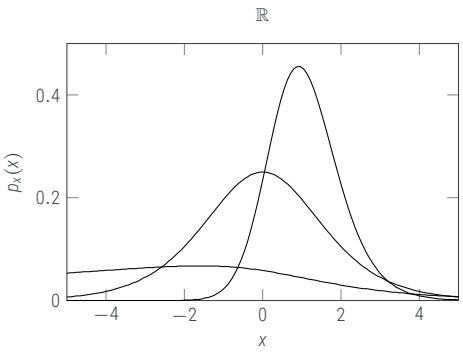
$$\arg\max_{\pi} \mathcal{B}(\pi; a, b) = \frac{a-1}{a+b-2} \qquad \mathbb{E}_{\mathcal{B}}(\pi) = \frac{a}{a+b}.$$

Linking Gaussians and Dirichlets

It's all about the right basis

DJC MacKay, *Choice of Basis for Laplace Approximation.* Machine Learning, **33** (1):77–86, 1998





consider $\pi(x) = \dfrac{\exp(x_k)}{\exp(x) + 1}, \quad x \in \mathbb{R},$

$p_x(x) = p_\pi(\pi(x)) \dfrac{\partial \pi_k}{\partial x_\ell} = p_\pi(\pi(x))\pi(1-\pi)$

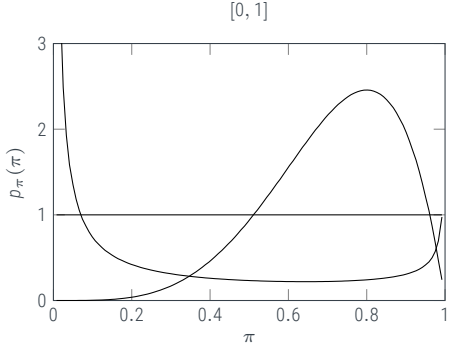
$\mathcal{B}_x(\pi(x); a, b) = \dfrac{1}{B(a,b)} \pi(x)^a \cdot (1 - \pi(x))^b$

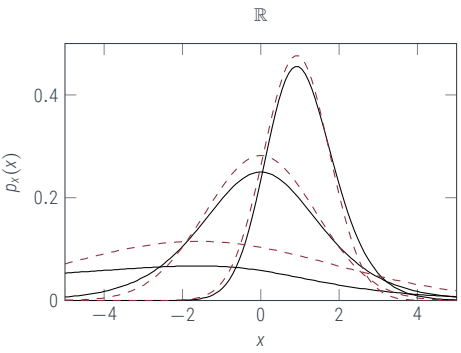
$\arg\max_{\pi(x)} \mathcal{B}_x(\pi(x); a, b) = \dfrac{a}{a+b} = \mathbb{E}_{\mathcal{B}}(\pi).$

Linking Gaussians and Dirichlets

It's all about the right basis

DJC MacKay, *Choice of Basis for Laplace Approximation.* Machine Learning, **33** (1):77–86, 1998





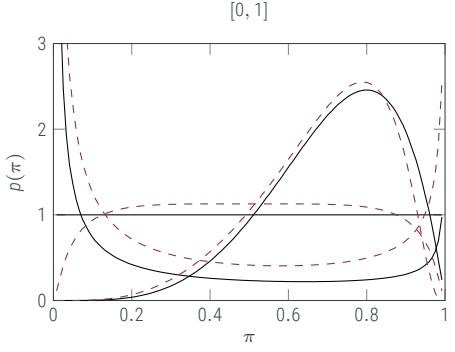
$$\frac{\partial \log \mathcal{B}_x(\pi(x), a, b)}{\partial x} = 0 \quad \Rightarrow \quad x = \log a - \log b =: \mu$$

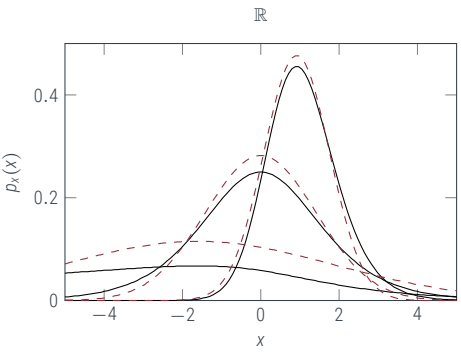
$$-\left(\frac{\partial^2 \log \mathcal{B}_x(\pi(x), a, b)}{\partial x^2}\bigg|_{x=\mu} \right)^{-1} = \frac{a+b}{a \cdot b} =: \sigma^2 \quad \Rightarrow \quad \mathcal{B}_x(x, a, b) \approx \mathcal{N}\left(x; \log\left(\frac{a}{b}\right), \frac{a+b}{a \cdot b}\right)$$

Linking Gaussians and Dirichlets

It's all about the right basis

DJC MacKay, *Choice of Basis for Laplace Approximation.* Machine Learning, **33** (1):77–86, 1998





$$\mathcal{N}(x;\mu,\sigma^2) \approx \mathcal{B}_\pi\left(\pi(x); a, b = \frac{\exp(\pm\mu)+1}{\sigma^2}\right) \quad \Leftrightarrow \quad \mathcal{B}_x(x,a,b) \approx \mathcal{N}\left(x; \log\left(\frac{a}{b}\right), \frac{a+b}{a\cdot b}\right)$$

The General Case

DJC MacKay, *Choice of Basis for Laplace Approximation.* Machine Learning, **33** (1):77–86, 1998

$$\mathcal{D}(\pi;\alpha) = \frac{1}{B(\alpha)} \prod_k \pi_k^{\alpha_k - 1} \qquad \pi \in [0,1], \quad \sum_k \pi_k = 1$$

$$\arg\max_\pi \mathcal{D}(\pi;\alpha) = \frac{\alpha - 1}{\sum_{k'} \alpha_{k'} - K} \qquad \mathbb{E}_{\mathcal{D}}(\pi_k) = \frac{\alpha_k}{\sum_\ell \alpha_\ell}$$

$$\text{consider} \quad \pi_k(x) = \frac{\exp(x_k)}{\sum_{k'} \exp(x_{k'})} \qquad p_x(x) = p_\pi(\pi(x)) \left| \left[\frac{\partial \pi_k}{\partial x_\ell} \right]_{x\ell} \right| = p_\pi(\pi(x)) \prod_k \pi_k$$

leaving out some book-keeping details to ensure $\sum_k \pi_k = 1$.

$$\mathcal{D}_x(\pi(x);\alpha) = \frac{1}{B(\alpha)} \prod_k \pi(x)_k^{\alpha_k} \qquad x \in [0,1]$$

$$\arg\max_{\pi(x)} \mathcal{D}_x(\pi(x);\alpha) = \frac{\alpha}{\sum_{k'} \alpha_{k'}} = \mathbb{E}_{\mathcal{D}}(\pi_k)$$

The Laplace Bridge

MacKay, 1998 / Hennig, Stern, Herbrich, Graepel, 2012

$\mathcal{D}(\pi;\alpha) \twoheadrightarrow \mathcal{N}(x;\mu,\Sigma)$

$$\mu_k = \log \alpha_k - \frac{1}{K}\sum_{\ell=1}^{K} \log \alpha_\ell$$

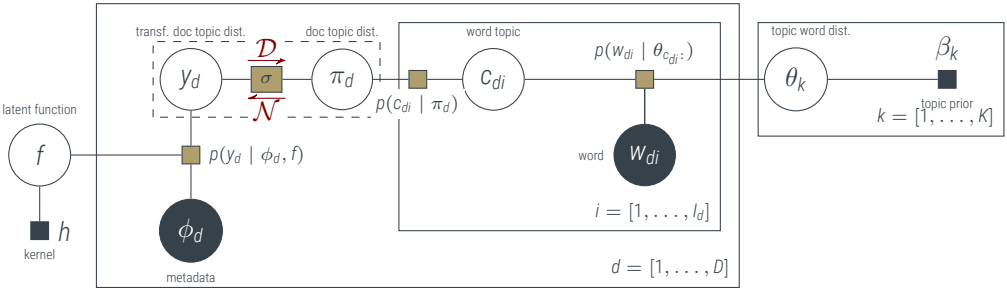
$$\Sigma_{k\ell} = \frac{\delta_{k\ell}}{\alpha_k} - \frac{1}{K}\left(\frac{1}{\alpha_k} + \frac{1}{\alpha_\ell} - \frac{1}{K}\sum_{u=1}^{K}\frac{1}{\alpha_u}\right), \quad \text{and in particular}$$

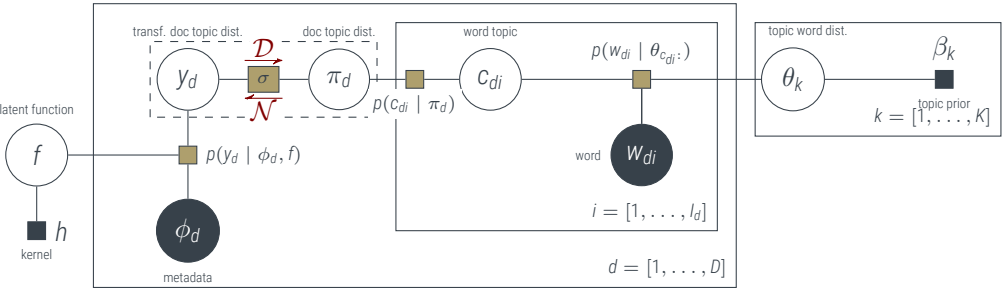
$$\Sigma_{kk} = \frac{1}{\alpha_k}\left(1 - \frac{2}{K}\right) + \frac{1}{K^2}\sum_{\ell=1}^{K}\frac{1}{\alpha_\ell}$$

nb: off-diagonal elements suppressed by $\mathcal{O}(K^{-1})$, so diagonal approximation is fine for $K \gg 1$.

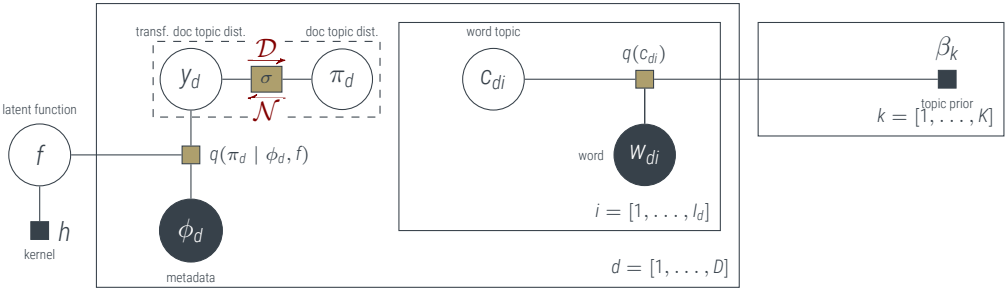
$\mathcal{N}(x;\mu,\Sigma) \twoheadrightarrow \mathcal{D}(\pi;\alpha)$

$$\alpha_k = \frac{1}{\Sigma_{kk}}\left(1 - \frac{2}{K} + \frac{\exp(-\mu_k)}{K^2}\sum_{\ell=1}^{K}\exp(-\mu_\ell)\right)$$


transf. doc topic dist.
doc topic dist.
word topic
topic word dist.
latent function
$\mathcal{D}$
σ
$\mathcal{N}$
y_d
π_d
c_{di}
$p(w_{di} \mid \theta_{c_{di}:})$
θ_k
β_k
topic prior
$k = [1, \ldots, K]$
$p(c_{di} \mid \pi_d)$
f
$p(y_d \mid \phi_d, f)$
word
w_{di}
h
kernel
ϕ_d
metadata
$i = [1, \ldots, l_d]$
$d = [1, \ldots, D]$



Unfortunately, this isn't a tree.



1. In the real-valued basis, perform **Gaussian process** inference

$$p(\boldsymbol{y} \mid \Pi, \Phi) \propto \mathcal{GP}(\boldsymbol{y} \mid \Phi) \cdot \mathcal{N}_\pi(\boldsymbol{y}; m_\pi, \Lambda_\pi) = \mathcal{GP}(y(t); h_{tT}(h_{TT}+\Lambda_\pi)^{-1} m_\pi, h_{tt} - h_{tT}(h_{TT}+\Lambda_\pi)^{-1} h_{Tt})$$

2. in the count basis, fit a **semi-collapsed variational bound** $q(\Pi, C) = \prod_d q(\pi_d) \prod_i q(c_{di})$ to

$$p(W, C, \Pi \mid \beta, \alpha) = \left(\prod_d \mathcal{D}(\pi_d; \alpha_{\phi,d}) \prod_k \pi_{dk}^{n_{dk\cdot}} \right) \left(\prod_k \frac{\Gamma(\sum_v \beta_{kv})}{\Gamma(n_{\cdot k\cdot} + \sum_v \beta_{kv})} \prod_v \frac{\Gamma(\beta_{kv} + n_{\cdot kv})}{\Gamma(\beta_{kv})} \right)$$

The Algorithm

[Hennig, Stern, Herbrich, Graepel. *Kernel Topic Models*. AISTATS 2012]

Variables in memory:

- $m_d \in \mathbb{R}^D, h_{dd'} \in \mathbb{R}^{D\times D}$ for Gaussian process on $\boldsymbol{y}$
- $\nu_d \in \mathbb{R}_+^K$ for Dirichlet on π_d
- $\gamma_{di} \in \mathbb{R}_+^K$ for discrete distribution on c_{di} (and pseudocounts n_{dkv}, $\mathbb{E}_q[n]$, $\mathrm{var}_q[n]$ constructed hence)

1. propagate meta-data information between documents by **Gaussian process** inference

$$p(\boldsymbol{y} \mid \Pi, \Phi) = \mathcal{GP}(y(t); h_{tT}(h_{TT}+\Lambda_\pi)^{-1}m_\pi, h_{tt} - h_{tT}(h_{TT}+\Lambda_\pi)^{-1}h_{Tt})$$

2. map to count-basis using the **Laplace bridge** $\alpha_d = \texttt{Laplace}(m_d^{\setminus d}, \lambda_d^{\setminus d})$
3. update **variational approximation** (use Gaussian approximation for $\mathbb{E}_q[\log \beta + n]$, cf. Lecture 23)

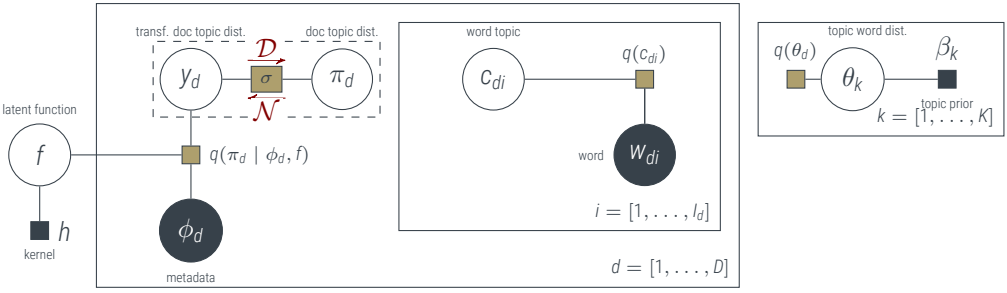
$$q(C) = \prod_{di}\prod_k \gamma_{dik} \quad \text{with} \quad \gamma_{dik} \propto \exp\left(F(\nu_{dk}) + \mathbb{E}_q(\log(\beta_{kw_{di}} + n_{\cdot kw_{di}}^{\setminus di})) - \mathbb{E}_q\left(\sum_v \beta_{kv} + n_{\cdot k\cdot}^{\setminus di}\right)\right)$$

$$q(\Pi) = \prod_d \mathcal{D}(\pi_d; \nu_d) \quad \text{with} \quad \nu_{dk} = \alpha_{dk} + \mathbb{E}(n_{dk\cdot}) = \alpha_{dk} + \gamma_{d\cdot k}$$

4. map back to real-valued basis using inverse **Laplace bridge** $(m_d, \lambda_d) = \texttt{Laplace}(\nu_d - \alpha_d)$

The Algorithm — Simplified

Kernel Topic Models with explicit variational approximation on Θ



1. In the real-valued basis, perform **Gaussian process** inference

$$p(\boldsymbol{y} \mid \Pi, \Phi) \propto \mathcal{GP}(\boldsymbol{y} \mid \Phi) \cdot \mathcal{N}_\pi(\boldsymbol{y}; m_\pi, \Lambda_\pi) = \mathcal{GP}(y(t); h_{tT}(h_{TT}+\Lambda_\pi)^{-1} m_\pi, h_{tt} - h_{tT}(h_{TT}+\Lambda_\pi)^{-1} h_{Tt})$$

2. in the count basis, fit an **explicit variational bound** $q(\Pi, C, \Theta) = \prod_k q(\theta_k) \prod_d q(\pi_d) \prod_i q(c_{di})$ to

$$p(W, C, \Pi, \Theta \mid \beta, \alpha) = \left(\prod_d \mathcal{D}(\pi_d; \alpha_{\phi,d}) \prod_k \pi_{dk}^{n_{dk\cdot}} \right) \left(\prod_k \mathcal{D}(\theta_k; \beta_k) \prod_v \theta_{kv}^{n_{\cdot kv}} \right)$$

The Algorithm — Simplified

Kernel Topic Models with explicit variational approximation on Θ

Variables in memory:

- $m_d \in \mathbb{R}^D, h_{dd'} \in \mathbb{R}^{D\times D}$ for Gaussian process on $\boldsymbol{y}$
- $\nu_d \in \mathbb{R}^K_+$ for Dirichlet on π_d
- $\gamma_{di} \in \mathbb{R}^K_+$ for discrete distribution on c_{di} (and pseudocounts n_{dkv} constructed hence)

1. propagate meta-data information between documents by **Gaussian process** inference

$$p(\boldsymbol{y} \mid \Pi, \Phi) = \mathcal{GP}(y(t); h_{tT}(h_{TT} + \Lambda_\pi)^{-1} m_\pi, h_{tt} - h_{tT}(h_{TT} + \Lambda_\pi)^{-1} h_{Tt})$$

2. map to count-basis using the **Laplace bridge** $\alpha_d = \texttt{Laplace}(m_d^{\setminus d}, \lambda_d^{\setminus d})$
3. update **variational approximation**

$$q(C) = \prod_{di}\prod_k \gamma_{dik} \quad \text{with} \qquad \gamma_{dik} \propto \exp\left(F(\nu_{dk}) + F(\xi_{kw_{di}}) - F\left(\sum_v \xi_{kv}\right)\right)$$

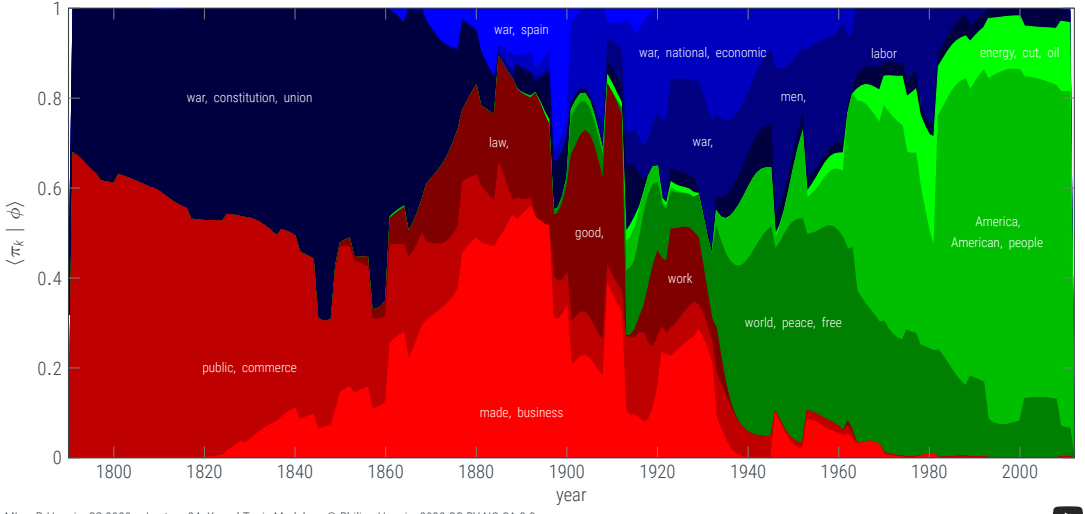
$$q(\Pi)q(\Theta) = \prod_d \mathcal{D}(\pi_d; \nu_d) \prod_k \mathcal{D}(\theta_k; \xi_k) \quad \text{with} \quad \nu_{dk} = \alpha_{dk} + n_{d\cdot k}, \quad \xi_{kv} = \beta_{kv} + n_{\cdot kv}$$

4. map back to real-valued basis using inverse **Laplace bridge** $(m_d, \lambda_d) = \texttt{Laplace}(\nu_d - \alpha_d)$

Results

Kernel Topic Models

[Hennig, Stern, Herbrich, Graepel. *Kernel Topic Models*. AISTATS 2012]



The Topics of American History

1897

The most important problem with which this Government is now called upon to deal pertaining to its foreign relations concerns its duty toward Spain and the Cuban insurrection.

(William McKinley, 1897)

Spanish–American War

Part of the Philippine Revolution and the Cuban War of Independence

(clockwise from top left)
Signal Corps extending telegraph lines from trenches · USS *Iowa* · Filipino soldiers wearing Spanish pith helmets outside Manila · The defeated side signing the Treaty of Paris · Roosevelt and his Rough Riders at the captured San Juan Hill · Replacing of the Spanish flag at Fort Malate

Date April 21, 1898[b] – August 13, 1898
(3 months, 3 weeks and 2 days)

Three basic developments have helped to shape our challenges: the steady growth and increased projection of Soviet military power beyond its own borders; the overwhelming dependence of the Western democracies on oil supplies from the Middle East; and the press of social and religious and economic and political change in the many nations of the developing world, exemplified by the revolution in Iran. (Jimmy Carter, 1980)

1979 oil crisis

From Wikipedia, the free encyclopedia

Further information: 1979 world oil market chronology

The **1979** (or **second**) **oil crisis** or **oil shock** occurred in the world due to decreased oil outpu
months, and long lines once again appeared at gas stations, as they had in the 1973 oil crisis.[

In 1980, following the outbreak of the Iran–Iraq War, oil production in Iran nearly stopped, and

After 1980, oil prices began a 20-year decline, except for a brief rebound during the Gulf War,
the top world producer; North Sea and Alaskan oil flooded the market. It seemed that the Unite

Can we do even better?

Intra-Document Structure! Bags of Bags of Words

Mr. Speaker, Mr. Vice President, Members of Congress, my fellow Americans:

We are 15 years into this new century. Fifteen years that dawned with terror touching our shores, that unfolded with a new generation fighting two long and costly wars, that saw a vicious recession spread across our Nation and the world. It has been and still is a hard time for many.

But tonight we turn the page. Tonight, after a breakthrough year for America, our economy is growing and creating jobs at the fastest pace since 1999. Our unemployment rate is now lower than it was before the financial crisis. More of our kids are graduating than ever before. More of our people are insured than ever before. And we are as free from the grip of foreign oil as we've been in almost 30 years.

Tonight, for the first time since 9/11, our combat mission in Afghanistan is over. Six years ago, nearly 180,000 American troops served in Iraq and Afghanistan. Today, fewer than 15,000 remain. And we salute the courage and sacrifice of every man and woman in this 9/11 generation who has served to keep us safe. We are humbled and grateful for your service.

America, for all that we have endured, for all the grit and hard work required to come back, for all the tasks that lie ahead, know this: The shadow of crisis has passed, and the State of the Union is strong.

Barack H. Obama, 2015

Each document is actually pre-structured into sequential sub-documents, typically of one topic each.

Designing a probabilistic machine learning method:

1. get the **data**
   1.1 try to collect as much meta-data as possible
2. build the **model**
   2.1 identify quantities and datastructures; assign names
   2.2 design a generative process (graphical model)
   2.3 assign (conditional) distributions to factors/arrows (use exponential families!)
3. design the **algorithm**
   3.1 consider conditional independence
   3.2 try standard methods for early experiments
   3.3 run unit-tests and sanity-checks
   3.4 identify bottlenecks, find customized approximations and refinements

Packaged solutions can give great first solutions, fast.
Building a tailormade solution requires creativity and mathematical stamina.

