Probabilistic Machine Learning
Lecture 25
Customizing Probabilistic Models

Philipp Hennig

19 July 2021

Eberhard Karls Universität Tübingen

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

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 (iterative maximum likelihood)
- variational inference / mean field

Variational Inference

- is a general framework to construct approximating **probability distributions** $q(z)$ to non-analytic posterior distributions $p(z \mid x)$ by minimizing the **functional**

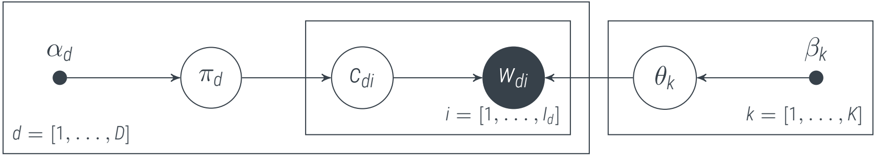
$$q^* = \arg\min_{q \in \mathcal{Q}} D_{KL}(q(z) \| p(z \mid x)) = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q)$$

- the beauty is that we get to *choose* q, so one can nearly always find a tractable approximation.
- If we impose the *mean field approximation* $q(z) = \prod_i q(z_i)$, get

$$\log q_j^*(z_j) = \mathbb{E}_{q, i \neq j}(\log p(x, z)) + \text{const.}.$$

- for Exponential Family p things are particularly simple: we only need the expectation under q of the sufficient statistics.

Variational Inference is an extremely flexible and powerful approximation method. Its downside is that constructing the bound and update equations can be tedious. For a quick test, variational inference is often not a good idea. But for a deployed product, it can be the most powerful tool in the box.



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

- 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_{ik} 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: $n_{dkv} = \#\{i : w_{di} = v, c_{ijk} = 1\}$. Write $n_{dk:} := [n_{dk1}, \dots, n_{dkV}]$ and $n_{dk\cdot} = \sum_v n_{dkv}$, etc.

The Variational Approximation

closing the loop

$$q(\boldsymbol{\pi}_d) = \mathcal{D}\left(\boldsymbol{\pi}_d; \tilde{\alpha}_{dk} := \left[\alpha_{dk} + \sum_{i=1}^{I_d} \tilde{\gamma}_{dik}\right]_{k=1,\dots,K}\right) \qquad \forall d = 1, \dots, D$$

$$q(\boldsymbol{\theta}_k) = \mathcal{D}\left(\boldsymbol{\theta}_k; \tilde{\beta}_{kv} := \left[\beta_{kv} + \sum_{d}^{D} \sum_{i=1}^{I_d} \tilde{\gamma}_{dik} \mathbb{I}(w_{di} = v)\right]_{v=1,\dots,V}\right) \qquad \forall k = 1, \dots, K$$

$$q(\boldsymbol{c}_{di}) = \prod_k \tilde{\gamma}_{dik}^{c_{dik}}, \qquad \forall d \;\; i = 1, \dots, I_d$$

where $\tilde{\gamma}_{dik} = \gamma_{dik} / \sum_k \gamma_{dik}$ and (note that $\sum_k \tilde{\alpha}_{dk} = \text{const.}$)

$$\gamma_{dik} = \exp\left(\mathbb{E}_{q(\pi_{dk})}(\log \pi_{dk}) + \mathbb{E}_{q(\theta_{di})}(\log \theta_{kw_{di}})\right)$$

$$= \exp\left(\digamma(\tilde{\alpha}_{jk}) + \digamma(\tilde{\beta}_{kw_{di}}) - \digamma\left(\sum_v \tilde{\beta}_{kv}\right)\right)$$

Building the Algorithm

updating and evaluating the bound

procedure LDA(W, α, β)

  $\tilde{\gamma}_{dik} \leftarrow$ DIRICHLET_RAND(α) // initialize

  $\mathcal{L} \leftarrow -\infty$

  while $\mathcal{L}$ not converged **do**

    for $d = 1, \dots, D; k = 1, \dots, K$ **do**

      $\tilde{\alpha}_{dk} \leftarrow \alpha_{dk} + \sum_i \tilde{\gamma}_{dik}$ // update document-topics distributions

    end for

    for $k = 1, \dots, K; v = 1, \dots, V$ **do**

      $\tilde{\beta}_{kv} \leftarrow \beta_{kv} + \sum_{d,i} \tilde{\gamma}_{dik} \mathbb{I}(w_{di} = v)$ // update topic-word distributions

    end for

    for $d = 1, \dots, D; k = 1, \dots, K; i = 1, \dots, I_d$ **do**

      $\tilde{\gamma}_{dik} \leftarrow \exp(\digamma(\tilde{\alpha}_{dk}) + \digamma(\tilde{\beta}_{kw_{di}}) - \digamma(\sum_v \tilde{\beta}_{kv}))$ // update word-topic assignments

      $\tilde{\gamma}_{dik} \leftarrow \tilde{\gamma}_{dik} / \tilde{\gamma}_{di\cdot}$

    end for

    $\mathcal{L} \leftarrow$ BOUND($\tilde{\gamma}, w, \tilde{\alpha}, \tilde{\beta}$) // update bound

  end while

end procedure

Exponential Family Approximations

The connection to EM is not accidental

- What has happened here? Why the connection to EM?
- Consider an **exponential family** joint distribution

$$p(x, z \mid \eta) = \prod_{n=1}^{N} \exp\left(\eta^\intercal \phi(x_n, z_n) - \log Z(\eta)\right)$$

with conjugate prior $p(\eta \mid \nu, v) = \exp\left(\eta^\intercal v - \nu \log Z(\eta) - \log F(\nu, v)\right)$

- and assume $q(z, \eta) = q(z) \cdot q(\eta)$. Then q is in the same exponential family, with

$$\log q^*(z) = \mathbb{E}_{q(\eta)}(\log p(x, z \mid \eta)) + \text{const.} = \sum_{n=1}^{N} \mathbb{E}_{q(\eta)}(\eta)^\intercal \phi(x_n, z_n)$$

$$q^*(z) = \prod_{n=1}^{N} \exp\left(\mathbb{E}(\eta)^\intercal \phi(x_n, z_n) - \log Z(\mathbb{E}(\eta))\right) \qquad \text{(note induced factorization)}$$

Exponential Family Approximations

The connection to EM is not accidental

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- What has happened here? Why the connection to EM?
- Consider an **exponential family** joint distribution

$$p(x, z \mid \eta) = \prod_{n=1}^{N} \exp\left(\eta^\intercal \phi(x_n, z_n) - \log Z(\eta)\right)$$

with conjugate prior $p(\eta \mid \nu, v) = \exp\left(\eta^\intercal v - \nu \log Z(\eta) - \log F(\nu, v)\right)$

- and assume $q(z, \eta) = q(z) \cdot q(\eta)$. Then q is in the same exponential family, with

$$\log q^*(\eta) = \log p(\eta \mid \nu, v) + \mathbb{E}_z(\log p(x, z \mid \eta)) + \text{const.}$$

$$= -\nu \log Z(\eta) + \eta^\intercal v + \sum_{n=1}^{N} - \log Z(\eta) + \eta^\intercal \mathbb{E}_z(\phi(x_n, z_n)) + \text{const.}$$

$$q^*(\eta) = \exp\left(\eta^\intercal \left(v + \sum_{n=1}^{N} \mathbb{E}_z(\phi(x_n, z_n))\right) - (\nu + N) \log Z(\eta) - \text{const.}\right)$$

Even, and especially if, you consider variational approximations,
using conjugate exponential family priors can make life much easier.

Reminder: Collapsed Gibbs Sampling

It pays off to look closely at the math!

T. L. Griffiths & M. Steyvers, *Finding scientific topics*, PNAS **101**/1 (4/2004), 5228–5235

Recall $\Gamma(x+1) = x \cdot \Gamma(x) \; \forall x \in \mathbb{R}_+$

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

$$= \left(\prod_{d=1}^{D} \frac{B(\alpha_d + n_{d:\cdot})}{B(\alpha_d)} \mathcal{D}(\pi_d; \alpha_d + n_{d:\cdot})\right) \cdot \left(\prod_{k=1}^{K} \frac{B(\beta_k + n_{\cdot k:})}{B(\beta_k)} \mathcal{D}(\theta_k; \beta_k + n_{\cdot k:})\right)$$

$$p(C,W) = \left(\prod_{d=1}^{D} \frac{B(\alpha_d + n_{d:\cdot})}{B(\alpha_d)}\right) \cdot \left(\prod_{k=1}^{K} \frac{B(\beta_k + n_{\cdot k:})}{B(\beta_k)}\right)$$

$$= \left(\prod_{d} \frac{\Gamma(\sum_{k'} \alpha_{dk'})}{\Gamma(\sum_{k'} \alpha_{dk'} + n_{dk'\cdot})} \prod_k \frac{\Gamma(\alpha_{dk}+n_{dk\cdot})}{\Gamma(\alpha_{dk})}\right) \left(\prod_{k} \frac{\Gamma(\sum_v \beta_{kv})}{\Gamma(\sum_v \beta_{kv} + n_{\cdot kv})} \prod_v \frac{\Gamma(\beta_{kv}+n_{\cdot kv})}{\Gamma(\beta_{kv})}\right)$$

$$p(c_{dik} = 1 \mid C^{\setminus di}, W) = \frac{(\alpha_{dk} + n_{dk\cdot}^{\setminus di})(\beta_{kw_{di}} + n_{\cdot kw_{di}}^{\setminus di})(\sum_v \beta_{kv} + n_{\cdot kv}^{\setminus di})^{-1}}{\sum_{k'}(\alpha_{dk'} + n_{dk'\cdot}^{\setminus di}) \cdot \sum_{w'}(\beta_{kw'} + n_{\cdot kw'}^{\setminus di}) \cdot \sum_{v'}(\beta_{kv'} + n_{\cdot kv'}^{\setminus di})^{-1}}$$

A Collapsed Gibbs Sampler for LDA

It pays off to look closely at the math!

T. L. Griffiths & M. Steyvers, *Finding scientific topics*, PNAS **101**/1 (4/2004), 5228–5235

$$p(C, W) = \left(\prod_d \frac{\Gamma(\sum_k \alpha_{dk})}{\Gamma(\sum_k \alpha_{dk} + n_{dk\cdot})} \prod_k \frac{\Gamma(\alpha_{dk} + n_{dk\cdot})}{\Gamma(\alpha_{dk})} \right) \left(\prod_k \frac{\Gamma(\sum_v \beta_{kv})}{\Gamma(\sum_v \beta_{kv} + n_{\cdot kv})} \prod_v \frac{\Gamma(\beta_{kv} + n_{\cdot kv})}{\Gamma(\beta_{kv})} \right)$$

A **collapsed** sampling method can converge much faster by eliminating the latent variables that mediate between individual data.

procedure LDA(W, α, β)
- $\gamma_{dkv} \leftarrow 0 \ \forall d, k, v$ // initialize counts
- **while** true **do**
  - **for** $d = 1, \dots, D; i = 1, \dots, I_d$ **do** // can be parallelized
    - $c_{di} \propto (\alpha_{dk} + n_{dk\cdot}^{\backslash di})(\beta_{kw_{di}} + n_{\cdot kw_{di}}^{\backslash di})(\sum_v \beta_{kv} + n_{\cdot kv}^{\backslash di})^{-1}$ // sample assignment
    - $n \leftarrow$ UPDATECOUNTS(c_{di}) // update counts (check whether first pass or repeat)
  - **end for**
- **end while**

end procedure

Can we do the same for variational inference?

Why don't we use the mean field in our variational bound?

Yee Whye Teh, David Newman & Max Welling, NeurIPS 2017

- Deriving our variational bound, we previously imposed the factorization

$$q(\Pi, \Theta, C) = q(\Pi, \Theta) \cdot \prod_{di} q(c_{di}), \quad \text{but can we get away with less? Like,}$$

$$q(\Pi, \Theta, C) = q(\Theta, \Pi \mid C) \cdot \prod_{di} q(c_{di})$$

- Note $p(C, \Theta, \Pi \mid W) = p(\Theta, \Pi \mid C, W)p(C \mid W)$. So when we minimize

$$D_{\text{KL}}(q(\Pi, \Theta, C) \| p(\Pi, \Theta, C \mid W)) = \int q(\Pi, \Theta \mid C)q(C) \log \left(\frac{q(\Pi, \Theta|C)q(C)}{p(\Pi, \Theta \mid C, W)p(C \mid W)} \right) dC \, d\Pi \, d\Theta$$

$$= \int q(\Pi, \Theta \mid C)q(C) \left[\log \left(\frac{q(\Pi, \Theta \mid C)}{p(\Pi, \Theta \mid C, W)} \right) + \log \left(\frac{q(C)}{p(C \mid W)} \right) \right] dC \, d\Pi \, d\Theta$$

$$= D_{\text{KL}}(q(\Pi, \Theta \mid C) \| p(\Pi, \Theta \mid C, W)) + D_{\text{KL}}(q(C) \| p(C \mid W))$$

we will just get $q(\Theta, \Pi) = p(\Theta, \Pi \mid C, W)$ and the bound will be *tight* in Π, Θ.

A Collapsed Variational Bound

Why don't we use the mean field in our variational bound?

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Yee Whye Teh, David Newman & Max Welling, NeurIPS 2007

$$p(C,W) = \left(\prod_d \frac{\Gamma(\sum_k \alpha_{dk})}{\Gamma(\sum_k \alpha_{dk} + n_{dk\cdot})} \prod_k \frac{\Gamma(\alpha_{dk}+n_{dk\cdot})}{\Gamma(\alpha_{dk})}\right) \left(\prod_k \frac{\Gamma(\sum_v \beta_{kv})}{\Gamma(\sum_v \beta_{kv} + n_{\cdot kv})} \prod_v \frac{\Gamma(\beta_{kv}+n_{\cdot kv})}{\Gamma(\beta_{kv})}\right)$$

- The remaining **collapsed variational bound** (ELBO) becomes

$$\mathcal{L}(q) = \int q(C) \log p(C,W)\, dC + \mathbb{H}(q(C))$$

- because we make strictly less assumptions about q than before, we will get a strictly better approximation to the true posterior!
- The bound is maximized for c_{di} if

$$\log q(c_{di}) = \mathbb{E}_{q(C^{\setminus di})}(\log p(C,W)) + \text{const.}$$

Constructing the Algorithm

Why didn't we do this earlier?

- Note that $c_{di} \in \{0;1\}^K$ and $\sum_k c_{dik} = 1$. So $q(c_{di}) = \prod_k \gamma_{dik}$ with $\sum_k \gamma_{dik} = 1$
- Also: $\Gamma(\alpha + n) = \prod_{\ell=0}^{n-1}(\alpha+\ell)$, thus $\log\Gamma(\alpha+n) = \sum_{\ell=0}^{n-1}\log(\alpha+\ell)$

$$p(C,W) = \left(\prod_d \frac{\Gamma(\sum_k \alpha_{dk})}{\Gamma(\sum_k \alpha_{dk} + n_{dk\cdot})} \prod_k \frac{\Gamma(\alpha_{dk}+n_{dk\cdot})}{\Gamma(\alpha_{dk})}\right)\left(\prod_k \frac{\Gamma(\sum_v \beta_{kv})}{\Gamma(\sum_v \beta_{kv} + n_{\cdot kv})} \prod_v \frac{\Gamma(\beta_{kv}+n_{\cdot kv})}{\Gamma(\beta_{kv})}\right)$$

$$\log q(c_{di}) = \mathbb{E}_{q(C^{\setminus di})}(\log p(C,W)) + \text{const.}$$

$$\begin{aligned}
\log \gamma_{dik} &= \log q(c_{dik} = 1) \\
&= \mathbb{E}_{q(C^{\setminus di})}\left[\log\Gamma(\alpha_{dk} + n_{dk\cdot}) + \log\Gamma(\beta_{kw_{di}} + n_{\cdot kw_{di}}) - \log\Gamma\left(\sum_v \beta_{kv} + n_{\cdot kv}\right)\right] + \text{const.} \\
&= \mathbb{E}_{q(C^{\setminus di})}\left[\log(\alpha_{dk} + n_{dk\cdot}^{\setminus di}) + \log(\beta_{kw_{di}} + n_{\cdot kw_{di}}^{\setminus di}) - \log\left(\sum_v \beta_{kv} + n_{\cdot kv}^{\setminus di}\right)\right] + \text{const.}
\end{aligned}$$

(note all terms in $p(C,W)$ that don't involve c_{dik} can be moved into the constant, as can all sums over k. We can also *add* terms to const., such as $\sum_{\ell=0}^{n^{\setminus di}-1}\log(\alpha+\ell)$, effectively cancelling terms in $\log\Gamma$)

A Complication

Ah, that's why!

$$\gamma_{dik} \propto \exp\left(\mathbb{E}_{q(C^{\setminus di})}\left[\log(\alpha_{dk} + n_{dk\cdot}^{\setminus di}) + \log(\beta_{kw_{di}} + n_{\cdot kw_{di}}^{\setminus di}) - \log\left(\sum_v \beta_{kv} + n_{\cdot kv}^{\setminus di}\right)\right]\right)$$

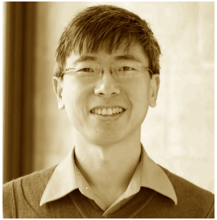
- Under $q(C) = \prod_{di} c_{di}$, the counts $n_{dk\cdot}$ are sums of independent Bernoulli variables (i.e. they have a **multinomial** distribution). Computing their expected logarithm is tricky ($\mathcal{O}(n_{d\cdot\cdot}^2)$):

$$\mathbb{H}(q(n_{dk\cdot})) = \mathbb{E}[\log n_{dk\cdot}] = -\log(I_d!) - I_d \sum_{k}^{K} \gamma_{dk\cdot} \log(\gamma_{dk\cdot}) + \sum_{k=1}^{K} \sum_{n_{dk\cdot}=1}^{I_d} \binom{I_d}{n_{dk\cdot}} \gamma_{dk\cdot}^{n_{dk\cdot}} (1-\gamma_{dk\cdot})^{I_d - n_{dk\cdot}} \log(n_{dk\cdot}!)$$

- That's likely why the original paper (and `scikit-learn`) don't do this.

If arithmetic doesn't work, try creativity!

Yee Whye Teh, David Newman & Max Welling, NeurIPS 2007



Yee Whye Teh

image: Oxford U

Max Welling

image: U v Amsterdam

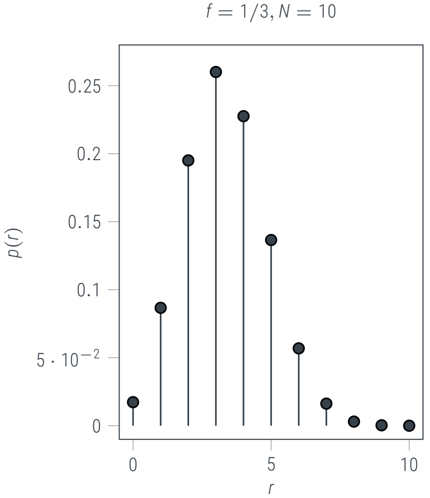
$$\gamma_{dik} \propto \exp\left(\mathbb{E}_{q(C^{\setminus di})}\left[\log(\alpha_{dk} + n_{dk\cdot}^{\setminus di}) + \log(\beta_{kw_{di}} + n_{\cdot kw_{di}}^{\setminus di}) - \log\left(\sum_v \beta_{kv} + n_{\cdot kv}^{\setminus di}\right)\right]\right)$$

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Yee Whye Teh, David Newman & Max Welling, NeurIPS 2007

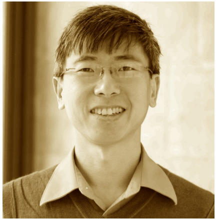
The probability measure of $R = \sum_i^N x_i$ with discrete x_i of probablity f is

$$
\begin{aligned}
P(R = r \mid f, N) &= \frac{N!}{(N-r)! \cdot r!} \cdot f^r \cdot (1-f)^{N-r} \\
&= \binom{N}{r} \cdot f^r \cdot (1-f)^{N-r} \\
&\approx \mathcal{N}(r; Nr, Nr(1-r))
\end{aligned}
$$



If arithmetic doesn't work, try creativity!

Yee Whye Teh, David Newman & Max Welling, NeurIPS 2007



Yee Whye Teh

image: Oxford U

Max Welling

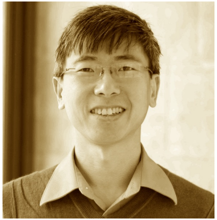
image: U v Amsterdam

but the CLT applies! So a Gaussian approximation should be good:

$$p(n_{dk\cdot}^{\setminus di}) \approx \mathcal{N}(n_{dk\cdot}^{\setminus di}; \mathbb{E}_q[n_{dk\cdot}^{\setminus di}], \operatorname{var}_q[n_{dk\cdot}^{\setminus di}]) \quad \text{with} \quad \mathbb{E}_q[n_{dk\cdot}^{\setminus di}] = \sum_{j \neq i} \gamma_{dkj}, \quad \operatorname{var}_q[n_{dk\cdot}^{\setminus di}] = \sum_{j \neq i} \gamma_{dkj}(1 - \gamma_{dkj})$$

If arithmetic doesn't work, try creativity!

Yee Whye Teh, David Newman & Max Welling, NeurIPS 2007



Yee Whye Teh

image: Oxford U

Max Welling

image: U v Amsterdam

$$\log(\alpha + n) \approx \log(\alpha + \mathbb{E}(n)) + (n - \mathbb{E}(n)) \cdot \frac{1}{\alpha + \mathbb{E}(n)} - \frac{1}{2}(n - \mathbb{E}(n))^2 \cdot \frac{1}{(\alpha + \mathbb{E}(n))^2}$$

$$\mathbb{E}_q[\log(\alpha_{dk} + n_{dk\cdot}^{\setminus di})] \approx \log(\alpha_{dk} + \mathbb{E}_q[n_{dk\cdot}^{\setminus di}]) - \frac{\operatorname{var}_q[n_{dk\cdot}^{\setminus di}]}{2(\alpha_{dk} + \mathbb{E}_q[n_{dk\cdot}^{\setminus di}])^2}$$

Approximate Approximate Inference

probabilistic machine learning often involves creative tweaks

$$\gamma_{dik} \propto \exp\left(\mathbb{E}_{q(C^{\setminus di})}\left[\log(\alpha_{dk} + n_{dk\cdot}^{\setminus di}) + \log(\beta_{kw_{di}} + n_{\cdot kw_{di}}^{\setminus di}) - \log\left(\sum_v \beta_{kv} + n_{\cdot kv}^{\setminus di}\right)\right]\right)$$

$$\mathbb{E}_q[\log(\alpha_{dk} + n_{dk\cdot}^{\setminus di})] \approx \log(\alpha_{dk} + \mathbb{E}_q[n_{dk\cdot}^{\setminus di}]) - \frac{\operatorname{var}_q[n_{dk\cdot}^{\setminus di}]}{2(\alpha_{dk} + \mathbb{E}_q[n_{dk\cdot}^{\setminus di}])^2}$$

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

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

Complexity

The algorithm

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

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

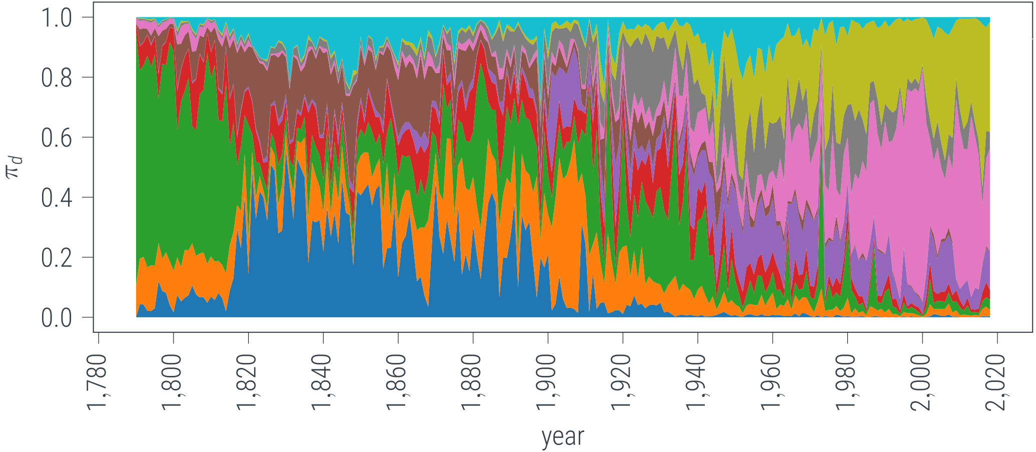
Note that γ_{dik} doesn't depend on $i \in 1, \dots, I_d$, it's the same for all w_{di} in d with $w_{di} = v$!

- memory requirement: $\mathcal{O}(DKV)$, since we have to store γ_{dik} for each value of $i \in 1, \dots, V$ and
  - $\mathbb{E}[n_{dk\cdot}], \mathrm{var}[n_{dk\cdot}] \in \mathbb{R}^{D\times K}$
  - $\mathbb{E}[n_{\cdot kv}], \mathrm{var}[n_{\cdot kv}] \in \mathbb{R}^{K\times V}$
  - $\mathbb{E}[n_{\cdot k\cdot}], \mathrm{var}[n_{\cdot k\cdot}] \in \mathbb{R}^{K}$
- computational complexity: $\mathcal{O}(DKV)$ We can loop over V rather than I_d (good for long documents!) Often, a document will be sparse in V, so iteration cost can be much lower.

Because machine learning involves real-world data, every problem is unique.
Thinking hard about both your *model* and your *algorithm*
can make a **big** difference in *predictive* and *numerical* performance.

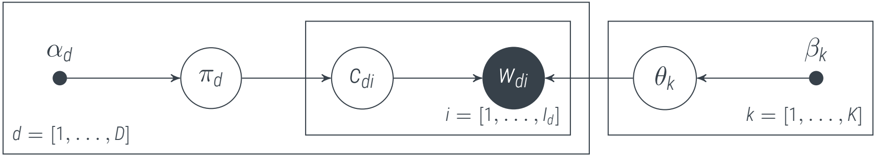
Some Output

can we be happy with this?



Can we be happy with this model?

Have we described all we know about the data?



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

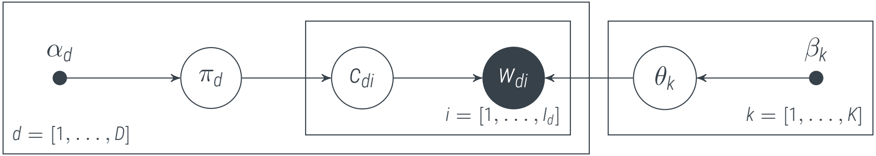
Meta-Data

It's right there!

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

What about the hyperparameters?

EM-style point estimates from the ELBO



$$\log p(W \mid \alpha, \beta) = \mathcal{L}(q, \alpha, \beta) + D_{\mathrm{KL}}(q \| p(C \mid W, \alpha, \beta))$$

$$\mathcal{L}(q, \alpha, \beta) = \int q(C, \Theta, \Pi) \log \left(\frac{p(W, \Pi, \Theta, C \mid \alpha, \beta)}{q(C, \Theta, \Pi)} \right)$$

$$\log p(\alpha, \beta \mid W) \geq \mathcal{L}(q, \alpha, \beta) + \log p(\alpha, \beta)$$

$$\nabla_{\alpha,\beta} \log p(\alpha, \beta \mid W) = \nabla_{\alpha,\beta} \mathcal{L}(q, \alpha, \beta) + \nabla_{\alpha,\beta} \log p(\alpha, \beta) + \underbrace{\nabla_{\alpha,\beta} D_{\mathrm{KL}}(q \| p(C \mid W, \alpha, \beta))}_{\approx 0}$$

The ELBO

useful for monitoring / bug-fixing

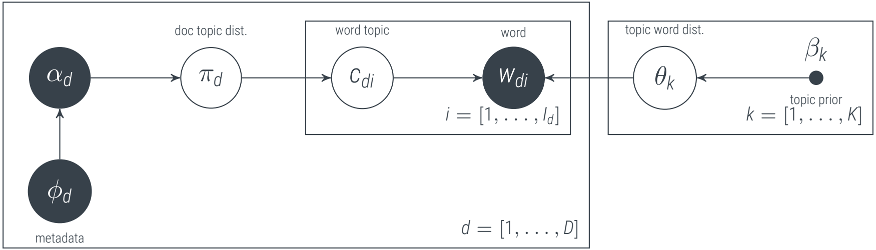
$$p(C,\Pi,\Theta,W) = \left(\prod_{d=1}^{D} \frac{\Gamma(\sum_k \alpha_{dk})}{\prod_k \Gamma(\alpha_{dk})} \prod_{k=1}^{K} \pi_{dk}^{\alpha_{dk}-1+n_{dk\cdot}}\right) \cdot \left(\prod_{k=1}^{K} \frac{\Gamma(\sum_v \beta_{kv})}{\prod_v \Gamma(\beta_{kv})} \prod_{v=1}^{V} \theta_{kv}^{\beta_{kv}-1+n_{\cdot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

We need

$$\begin{aligned}
\mathcal{L}(q,W) &= \mathbb{E}_q(\log p(W,C,\Theta,\Pi)) + \mathbb{H}(q) \\
&= \int q(C,\Theta,\Pi) \log p(W,C,\Theta,\Pi)\, dC\, d\Theta\, d\Pi - \int q(C,\Theta,\Pi) \log q(C,\Theta,\Pi)\, dC\, d\Theta\, d\Pi \\
&= \int q(C,\Theta,\Pi) \log p(W,C,\Theta,\Pi)\, dC\, d\Theta\, d\Pi + \sum_k \mathbb{H}(\mathcal{D}(\theta_k\ \tilde{\beta}_k)) + \sum_d \mathbb{H}(\mathcal{D}(\pi_d\ \tilde{\alpha}_d)) + \sum_{di} \mathbb{H}(\tilde{\gamma}_{di})
\end{aligned}$$

Adding more Information

a model for document metadata



The Price of Packaged Solutions

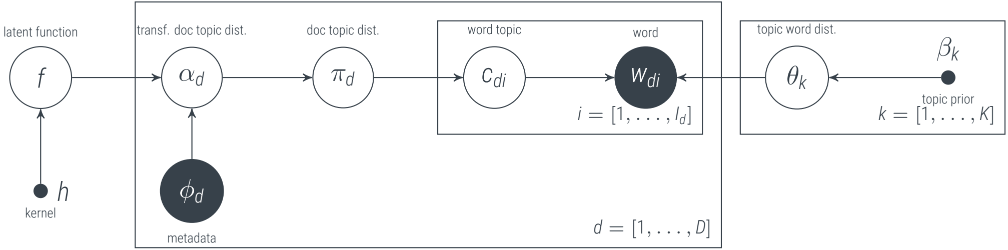
Toolboxes speed up development, but also make it brittle

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*

A Generalized Linear Model

Latent Topic Dynamics

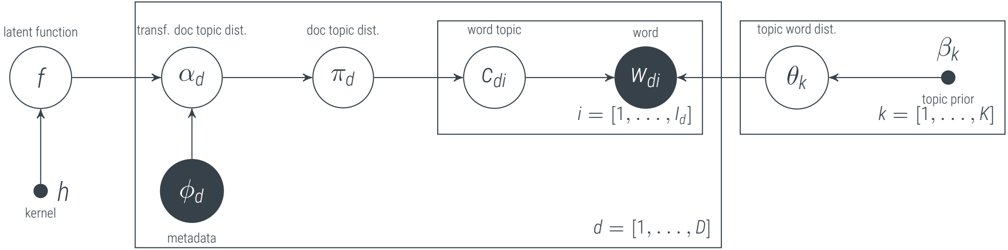


To generate the words W of documents $d = 1, \dots, D$ with features $\phi_d \in \mathbb{F}$:

- draw function $f : \mathbb{F} \to \mathbb{R}^K$ from $p(f \mid h) = \mathcal{GP}(f; 0, h)$
- draw document topic distribution π_d from $\mathcal{D}(\alpha_d = \exp(f(\phi_d)))$

A Generalized Linear Model

Latent Topic Dynamics

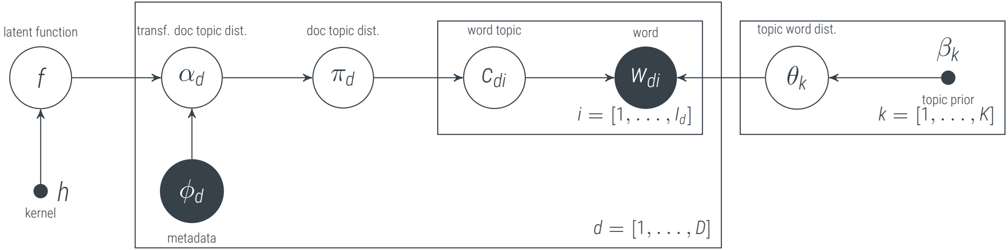


To generate the words W of documents $d = 1, \dots, D$ with features $\phi_d \in \mathbb{F}$:

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

A change in prior

EM-style point estimates from the ELBO



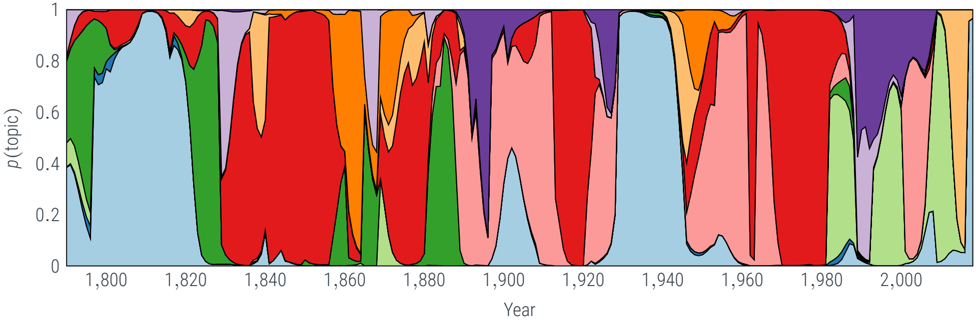
$$\log p(\alpha, \beta \mid W) \geq \mathcal{L}(q, \alpha, \beta) + \log p(\alpha, \beta)$$

$$\nabla_{\alpha,\beta} \log p(\alpha, \beta \mid W) = \nabla_{\alpha,\beta} \mathcal{L}(q, \alpha, \beta) + \nabla_{\alpha,\beta} \log p(\alpha, \beta) + \underbrace{\nabla_{\alpha,\beta} D_{\mathrm{KL}}(q \| p(C \mid W, \alpha, \beta))}_{\approx 0}$$

$$\log p(f = \log \alpha) = -\frac{1}{2}\|f_d\|_k^2 = -\frac{1}{2} f_d^\intercal k_{DD}^{-1} f_d$$

A GP prior on topics

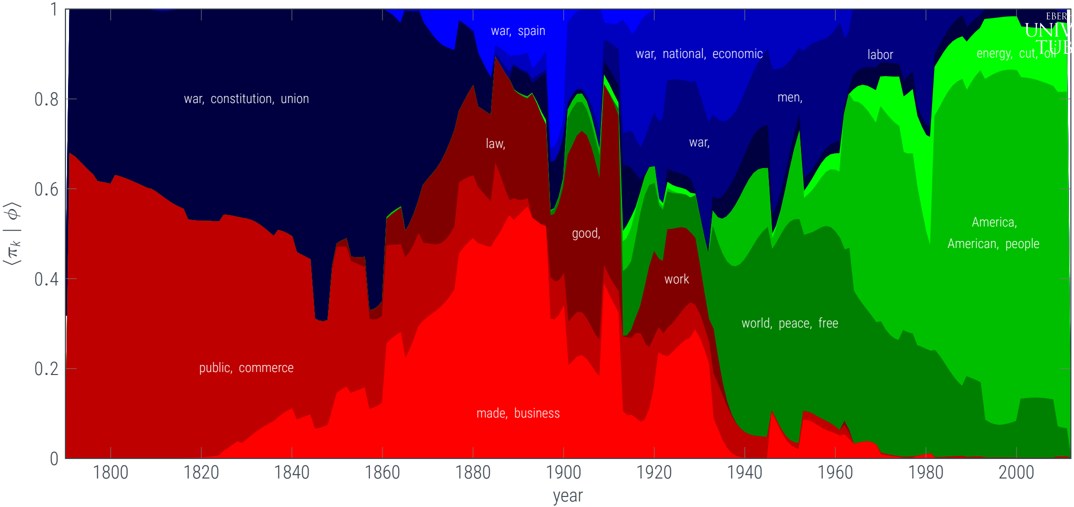
Lectures 3 – 13



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

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

S



1
0.8
0.6
0.4
0.2
0
$\langle \pi_k \mid \phi \rangle$
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
year
war, spain
war, national, economic
labor
energy, cut, oil
war, constitution, union
men,
law,
war,
good,
America,
American, people
work
world, peace, free
public, commerce
made, business

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)

The Topics of American History

1980

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.