Variational inference and Latent Dirichlet allocation

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Introduction

Classic variational inference

Justification of classic VI algorithm

Latent Dirichlet allocation
- Background and motivation
- LDA model
- VI for LDA
- Applications and extensions

Outline

Introduction

Classic variational inference

Justification of classic VI algorithm

Latent Dirichlet allocation
- Background and motivation
- LDA model
- VI for LDA
- Applications and extensions

Introduction

- Variational inference (VI) is an approach to posterior inference based on approximating the posterior by a nicer distribution.

- The general idea is to:
  1. choose a nice family of distributions $\mathcal{Q}$,
  2. find a $q \in \mathcal{Q}$ that is as close as possible to the posterior, and
  3. use q to quantify uncertainty, as a proxy for the posterior.

- Different choices of $\mathcal{Q}$ and different definitions of "close" lead to different variational inference techniques.

- In Bayesian statistics, VI is also known as variational Bayes, but VI is also useful outside of Bayesian inference.

Introduction: VI versus MCMC

- Similarities with MCMC
  1. Approximate inference
  2. Tractable on complex high-dimensional posteriors
  3. Computation formulas work out nicely in many cases

- Differences with MCMC
  1. Accuracy of VI is limited, usually
  2. VI is (usually) deterministic, MCMC is stochastic
  3. VI is faster, usually, and it is clear when VI has converged
  4. VI formulas are usually a bit more complicated

Introduction: Variants of VI

- Classic VI is based on minimizing Kullback–Leibler divergence using factorized approximations.
- Classic VI is elegant and fast, but its accuracy is limited by the assumed factorization.
- More recently, VI techniques using more flexible approximating distributions have been developed.
- Additionally, other divergences/distances have been employed.
- Further, stochastic optimization and autodifferentiation have proven to be useful for VI.

Outline

Introduction

Classic variational inference

Justification of classic VI algorithm

Latent Dirichlet allocation
- Background and motivation
- LDA model
- VI for LDA
- Applications and extensions

Classic variational inference

- Let $\pi(\theta)$ denote the target distribution, e.g., $\pi(\theta) = p(\theta|x)$.

- Classic VI makes the following choices:
  1. the approximating family $\mathcal{Q}$ consists of all factorized distributions of the form

     $$q(\theta) = q_1(\theta_1) \cdots q_m(\theta_m)$$

     for some decomposition of θ into components $\theta_1, \ldots, \theta_m$, and

  2. we seek the $q \in \mathcal{Q}$ that minimizes the Kullback–Leibler divergence from q to π,

     $$q^{\text{opt}} \in \operatorname*{argmin}_{q \in \mathcal{Q}} D(q\|\pi).$$

- The *Kullback–Leibler (KL) divergence*, or *relative entropy*, is

$$D(q\|\pi) = \int q(\theta) \log \frac{q(\theta)}{\pi(\theta)} d\theta = \mathrm{E}_q\Big(\log \frac{q(\theta)}{\pi(\theta)} \Big).$$

Classic variational inference: Algorithm

- To derive the classic VI updates, we need only specify the target distribution π and the decomposition of θ into $\theta = (\theta_1, \ldots, \theta_m)$.

- The classic VI algorithm then proceeds as follows:
  1. Initialize $q_1, \ldots, q_m$.
  2. Repeat until convergence:
     For $j = 1, \ldots, m$, update q_j by setting it to be

$$q_j^{\text{new}}(\theta_j) \propto \exp(h_j(\theta_j))$$

     where

$$h_j(\theta_j) := \mathrm{E}_q\big(\log \pi(\theta) \mid \theta_j\big) = \int \big(\log \pi(\theta)\big) \prod_{i \neq j} q_i(\theta_i) d\theta_i.$$

  3. Use $q(\theta) = q_1(\theta_1) \cdots q_m(\theta_m)$ as an approximation to $\pi(\theta)$.

Classic variational inference: Comments

- Note that we do not assume the form of each $q_j(\theta)$.
- At each step j, this algorithm optimizes the KL divergence with respect to q_j, that is,

$$q_j^{\text{new}} \in \operatorname*{argmin}_{q_j} D(q_1 \cdots q_m \| \pi).$$

- Remarkably, q_j^{new} is often a well-known exponential family.
- However, the tractability and accuracy of the algorithm depends on:
  1. making a good choice of decomposition $\theta_1, \ldots, \theta_m$, and
  2. being in the lucky situation that your model is conducive to classic VI.
- Classic VI works well for some models, and not others.

Toy example: VI for univariate normal

- Consider the univariate normal model as a toy example:

$$X_1, \ldots, X_n | \mu, \lambda \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \lambda^{-1})$$

  and assume an improper uniform prior on μ and λ.

- Define the target distribution to be the posterior:

$$\begin{aligned} \pi(\mu, \lambda) = p(\mu, \lambda | x_{1:n}) &\propto p(x_{1:n} | \mu, \lambda) \\ &\underset{\mu,\lambda}{\propto} \lambda^{n/2} \exp\big(-\tfrac{1}{2}\lambda \textstyle\sum_{i=1}^n (x_i - \mu)^2 \big). \end{aligned}$$

- Thus, $\log \pi(\mu, \lambda) = \frac{n}{2} \log \lambda - \frac{1}{2}\lambda \sum_{i=1}^n (x_i - \mu)^2 + \text{const.}$

- A natural decomposition to try would be $\theta_1 = \mu$ and $\theta_2 = \lambda$, that is, to consider approximations of the form

$$q(\mu, \lambda) = q_1(\mu) q_2(\lambda).$$

  For convenience, we write $q(\mu, \lambda) = q(\mu) q(\lambda)$.

Normal example: Deriving the VI updates (1/3)

- Updating $q(\mu)$ given $q(\lambda)$:

$$h_1(\mu) = \text{(whiteboard exercise)}$$

- Therefore, according to the algorithm, we update $q(\mu)$ to be

$$q^{\text{new}}(\mu) \propto \text{(whiteboard exercise)}$$

Normal example: Deriving the VI updates (1/3)

- Updating $q(\mu)$ given $q(\lambda)$:

$$\begin{aligned} h_1(\mu) &= \int q(\lambda) \log \pi(\mu, \lambda) d\lambda \\ &= \frac{n}{2} \int q(\lambda) \log \lambda - \frac{1}{2} \left(\textstyle\sum_i (x_i - \mu)^2 \right) \int \lambda q(\lambda) d\lambda + \text{const.} \end{aligned}$$

- Therefore, according to the algorithm, we update $q(\mu)$ to be

$$\begin{aligned} q^{\text{new}}(\mu) &\propto \exp(h_1(\mu)) \propto \exp\left(-\tfrac{1}{2} \mathrm{E}(\lambda) \textstyle\sum_i (x_i - \mu)^2 \right) \\ &\underset{\mu}{\propto} \mathcal{N}\left(\mu \mid \bar{x}, (n\mathrm{E}(\lambda))^{-1}\right). \end{aligned}$$

- Computationally, we only need to compute and store $\bar{x}$ and $\mathrm{E}(\lambda)$.

- Here, $\mathrm{E}(\lambda) = \int \lambda q(\lambda) d\lambda$ is computed using the current $q(\lambda)$.

Normal example: Deriving the VI updates (2/3)

- Updating $q(\lambda)$ given $q(\mu)$:

$$h_2(\lambda) = \text{(whiteboard exercise)}$$

- Therefore, according to the algorithm, we update $q(\lambda)$ to be

$$q^{\text{new}}(\lambda) \propto \text{(whiteboard exercise)}$$

Normal example: Deriving the VI updates (2/3)

- Updating $q(\lambda)$ given $q(\mu)$:

$$\begin{aligned} h_2(\lambda) &= \int q(\mu) \log \pi(\mu, \lambda) d\mu \\ &= \frac{n}{2} \log \lambda - \frac{1}{2} \lambda \sum_i \int (x_i - \mu)^2 q(\mu) d\mu + \text{const.} \end{aligned}$$

- Therefore, according to the algorithm, we update $q(\lambda)$ to be

$$\begin{aligned} q^{\text{new}}(\lambda) &\propto \exp(h_2(\lambda)) \propto \lambda^{n/2} \exp(-\tfrac{1}{2} S \lambda) \\ &\underset{\lambda}{\propto} \text{Gamma}(\lambda \mid n/2 + 1,\ S/2) \end{aligned}$$

  where, after plugging in $q(\mu) = \mathcal{N}\left(\mu \mid \bar{x},\ (n\text{E}(\lambda))^{-1}\right)$ and simplifying,

$$S = \sum_i \int (x_i - \mu)^2 q(\mu) d\mu = n\hat{\sigma}^2 + 1/\text{E}(\lambda).$$

- Here, $\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$.

Normal example: Deriving the VI updates (3/3)

- Thus, the updates to $q(\mu)$ and $q(\lambda)$ are:

$$q^{\text{new}}(\mu) = \mathcal{N}\big(\mu \mid \bar{x},\, (n\mathrm{E}(\lambda))^{-1}\big),$$
$$q^{\text{new}}(\lambda) = \text{Gamma}\left(\lambda \mid n/2+1,\ \tfrac{1}{2}(n\hat{\sigma}^2 + 1/\mathrm{E}(\lambda))\right).$$

- The only thing we need to compute at each iteration is $\mathrm{E}(\lambda)$.

- From the form of $q^{\text{new}}(\lambda)$, we see that

$$\mathrm{E}^{\text{new}}(\lambda) = \frac{n/2+1}{\frac{1}{2}(n\hat{\sigma}^2 + 1/\mathrm{E}(\lambda))}.$$

- In this example, it turns out that we can analytically solve for the limiting value of $\mathrm{E}(\lambda)$, which is $\mathrm{E}(\lambda) = (n+1)/(n\hat{\sigma}^2)$.

- However, in more complex models, we will have to iterate until convergence.

Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/dNbCsQcM5f4ywixR8

(Three people per room, randomly assigned. 15 minutes.)

Outline

Introduction

Classic variational inference

Justification of classic VI algorithm

Latent Dirichlet allocation
- Background and motivation
- LDA model
- VI for LDA
- Applications and extensions

Justification of classic VI algorithm (1/2)

- Let $\theta = (\theta_1, \ldots, \theta_m)$ where $\theta_i \sim q_i$ independently for some $q_1, \ldots, q_m$. Let π be the target distribution. Then for any j,

$$
\begin{aligned}
D(q_1 \cdots q_m \,\|\, \pi) &= \mathrm{E}\Big(\log \frac{q_1(\theta_1) \cdots q_m(\theta_m)}{\pi(\theta)} \Big) \\
&= \sum_{i=1}^{m} \mathrm{E} \log q_i(\theta_i) - \mathrm{E} \log \pi(\theta) \\
&= \mathrm{E} \log q_j(\theta_j) - \mathrm{E}\big(\mathrm{E}(\log \pi(\theta) | \theta_j)\big) + (\text{const wrt } q_j) \\
&= \mathrm{E} \log q_j(\theta_j) - \mathrm{E} h_j(\theta_j) + (\text{const wrt } q_j) \\
&= \mathrm{E}\Big(\log \frac{q_j(\theta_j)}{c\, e^{h_j(\theta_j)}} \Big) + (\text{const wrt } q_j) \\
&= D(q_j \,\|\, c\, e^{h_j}) + (\text{const wrt } q_j)
\end{aligned}
$$

  where $h_j(\theta_j) := \mathrm{E}(\log \pi(\theta) \mid \theta_j)$ and $1/c = \int e^{h_j(\theta_j)} d\theta_j$.

Justification of classic VI algorithm (2/2)

- Hence,

$$\operatorname*{argmin}_{q_j} D(q_1 \cdots q_m \,\|\, \pi) = \operatorname*{argmin}_{q_j} D(q_j \,\|\, c\, e^{h_j}).$$

- By the properties of KL divergence, $D(q_j \,\|\, c\, e^{h_j}) \geq 0$ with equality if and only if $q_j = c\, e^{h_j}$ almost everywhere.

- Therefore, choosing $q_j \propto e^{h_j}$ minimizes $D(q_1 \cdots q_m \,\|\, \pi)$ with respect to q_j, given q_i for $i \neq j$.

- This shows that the classic VI algorithm performs coordinate descent, updating q_j to minimize KL at each step, holding q_i fixed for $i \neq j$.

Outline

Introduction

Classic variational inference

Justification of classic VI algorithm

Latent Dirichlet allocation
- Background and motivation
- LDA model
- VI for LDA
- Applications and extensions

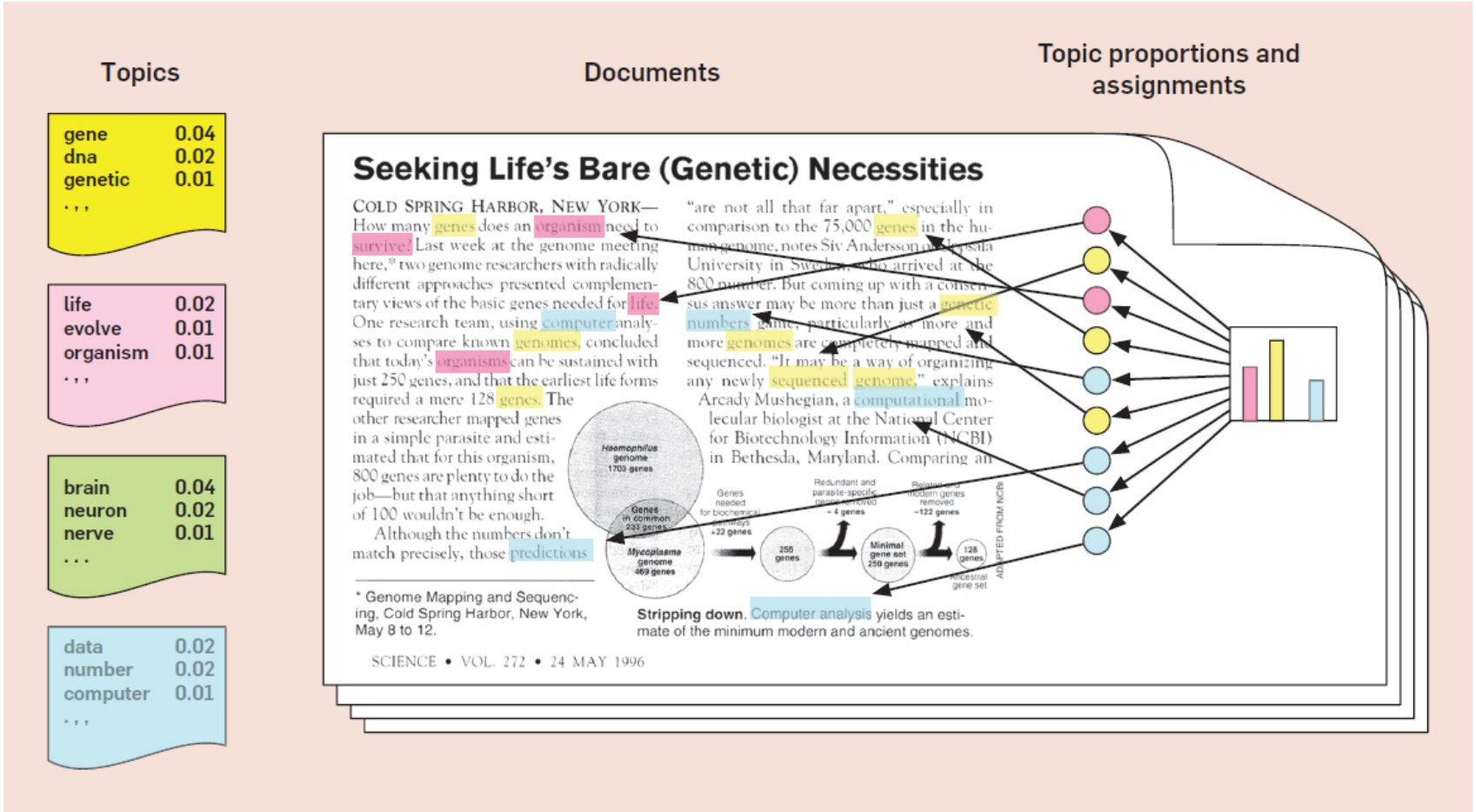
Latent Dirichlet allocation (LDA)

- LDA is a model for collections of discrete data such as documents.
- LDA is an admixture model in which each word in a document is drawn from a different topic.
- In LDA, topics are represented as distributions over words.
- The proportion of words coming from each topic varies from document to document.
- The words in each document are modeled as exchangeable, so LDA doesn't account for dependence on nearby words.

Latent Dirichlet allocation: Motivation

- LDA performs *topic modeling*, that is, it finds recurring themes in a large, unstructured collection of documents.
- It can be adapted to many kinds of data, such as images, omics data, and social networks.
- It has also been extended to handle streaming collections such as from a Web API.
- LDA has become widely used due to the need to organize an ever growing number of digital documents, along with its flexibility and utility.

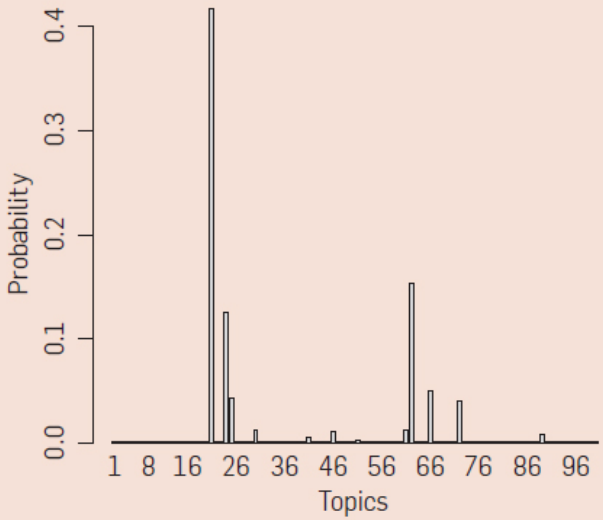
LDA: Application to Science articles



(figure from Blei, 2012)

- 17,000 articles from Science magazine, using $K = 100$ topics.

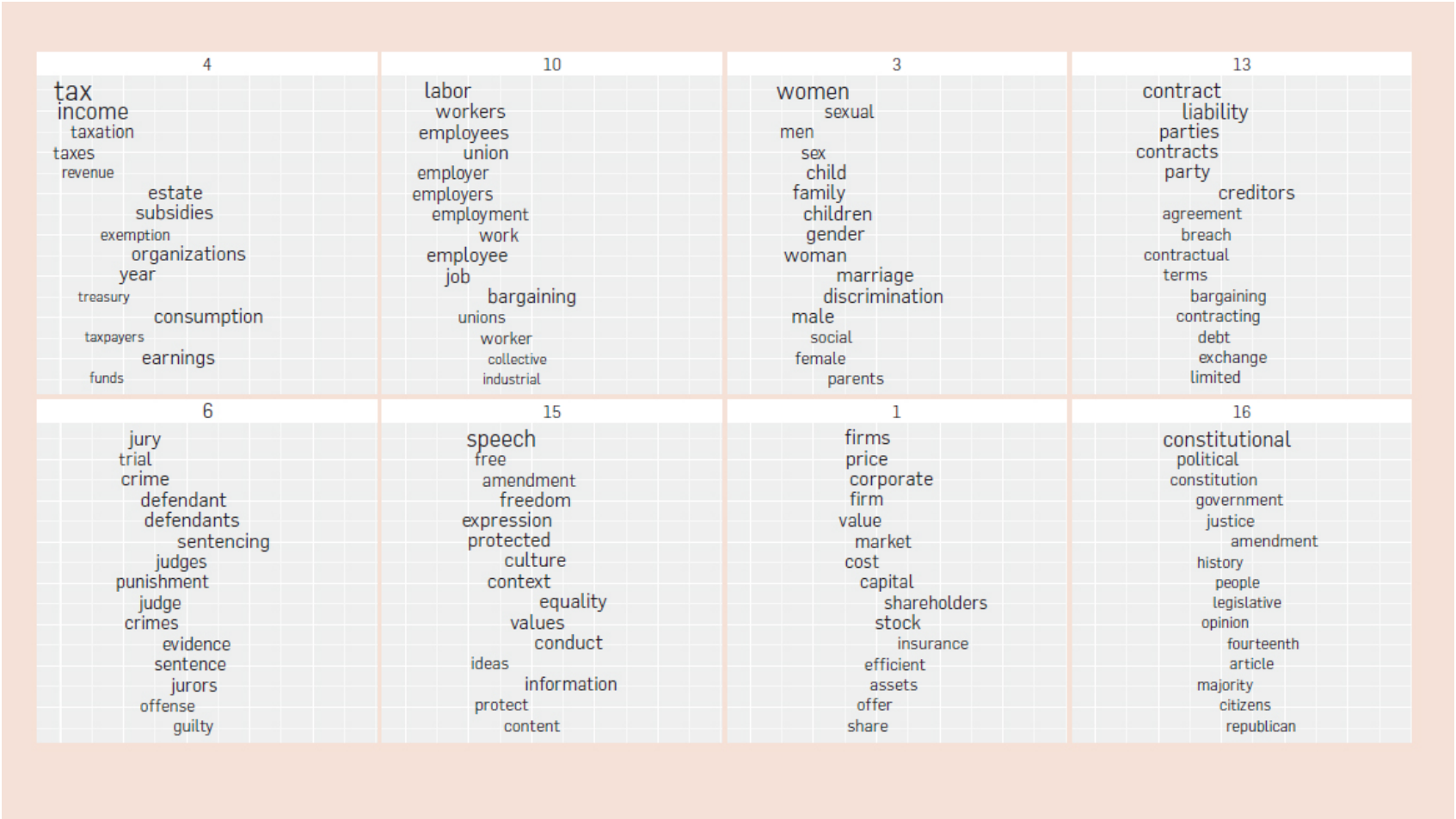
LDA: Application to Science articles



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

(figure from Blei, 2012)

LDA: Application to Yale law journal articles



(figure from Blei, 2012)

- x-axis position of each term indicates specificity to the article (more general $\leftrightarrow$ more specific).

Latent Dirichlet allocation: Model

- Suppose there are K topics, n documents, L_i words in document i, and $V =$ words in the vocabulary.

- For document i, define
  - $w_{ik} =$ proportion of the document that originates from topic k.
  - $z_{i\ell} =$ topic of origin for word ℓ.
  - $x_{i\ell} =$ the ℓth word in the document.

- $\beta_{kv} =$ frequency of word v in topic k.

- Differences from the population structure model:
  1. the topic distributions don't depend on the word location,
  2. there is only one word per location (not two allele copies), and
  3. each document has a different number of words.

Latent Dirichlet allocation: Model

- Consider the following model:

$$\begin{aligned} Z_{i\ell} \mid w &\sim \text{Categorical}(w_i) \\ X_{i\ell} \mid \beta,\, Z_{i\ell} = k &\sim \text{Categorical}(\beta_k) \end{aligned}$$

  independently for $i \in \{1, \ldots, n\}$, $\ell \in \{1, \ldots, L_i\}$.

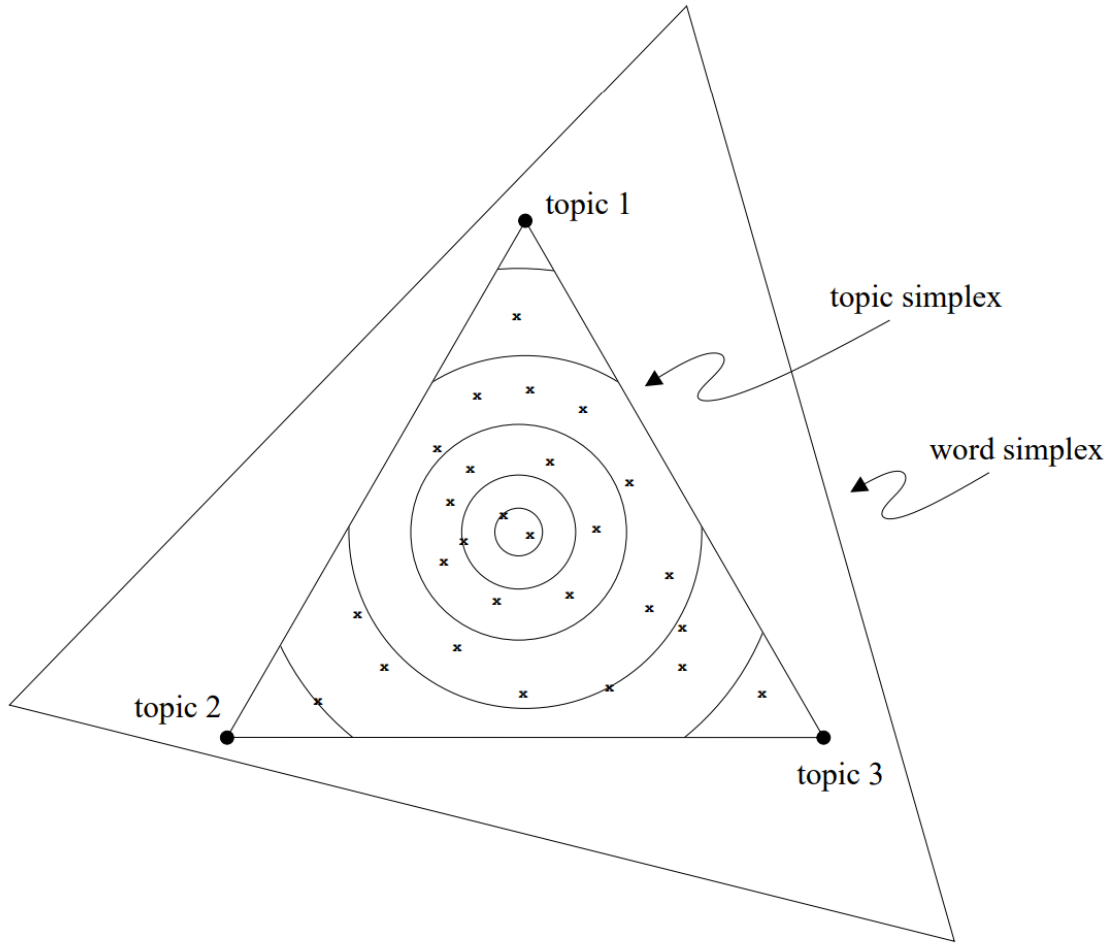
- Here, $w_i := (w_{i1}, \ldots, w_{iK})$ and $\beta_k := (\beta_{k1}, \ldots, \beta_{kV})$.

- For the prior, LDA uses:

$$\begin{aligned} w_i &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K), \\ \beta_k &\sim \text{Dirichlet}(\lambda_1, \ldots, \lambda_V). \end{aligned}$$

Latent Dirichlet allocation: Geometric interpretation

Visualization of document distributions over words



(Blei et al., 2003)

Latent Dirichlet allocation: Comments on model

- LDA is invariant to the order of words in a document, that is, you could permute the words and it would appear the same to the model.
- This is referred to as a "bag of words"-type model.
- This is clearly not a realistic model for language, but it is good enough to provide valuable insights into topics and documents.
- Other models account for word order using a Markov model, n-grams, or probabilistic context free grammars (PCFGs).
- Similarly, the model is invariant to the order of documents, which could be a limitation for collections of documents that span time.

Latent Dirichlet allocation: Posterior computation

- Gibbs sampling is straightforward, since everything is semi-conjugate.
- However, for very large datasets, Gibbs tends to be slow.
- Blei et al. (2003) proposed a variational inference algorithm that is much faster.
- The basic idea is just the classic VI algorithm applied to the LDA model.

Latent Dirichlet allocation: Variational inference (1/2)

- The target is the posterior, $\pi(z, w, \beta) := p(z, w, \beta \mid x)$.

- Consider approximations to π that factorize as

$$q(z, w, \beta) = q(z)q(w)q(\beta).$$

- The classic VI algorithm yields (after several pages of math):

$$q(w) = \prod_{i=1}^{n} \text{Dirichlet}(w_i \mid r_{i1}, \ldots, r_{iK})$$

$$q(\beta) = \prod_{k=1}^{K} \text{Dirichlet}(\beta_k \mid s_{k1}, \ldots, s_{kV})$$

$$q(z) = \prod_{i=1}^{n} \prod_{\ell=1}^{L_i} \text{Categorical}(z_{i\ell} \mid t_{i\ell})$$

where r, s, and t are computed as follows...

Latent Dirichlet allocation: Variational inference (2/2)

To compute the parameters r, s, and t, randomly initialize them and iterate the following steps until convergence:

1. For all i, k, update $r_{ik} \leftarrow \alpha_k + \sum_{\ell=1}^{L_i} t_{i\ell k}$.
2. For all k, v, update $s_{kv} \leftarrow \lambda_v + \sum_{i=1}^{n} \sum_{\ell=1}^{L_i} \mathrm{I}(x_{i\ell} = v) t_{i\ell k}$.
3. For all i, ℓ, k, update $t_{i\ell k} \leftarrow \frac{\exp(u_{i\ell k})}{\sum_{k'=1}^{K} \exp(u_{i\ell k'})}$, where

$$u_{i\ell k} = \psi(r_{ik}) - \psi\left(\textstyle\sum_{k'} r_{ik'}\right) + \sum_{v=1}^{V} \mathrm{I}(x_{i\ell} = v)\left(\psi(s_{kv}) - \psi\left(\textstyle\sum_{v'} s_{kv'}\right)\right)$$

and $\psi(\cdot)$ is the digamma function and $\mathrm{I}(\cdot)$ is the indicator function.

Latent Dirichlet allocation: Comments about VI algorithm

- Note that although we only assumed the factorization $q(z, w, \beta) = q(z)q(w)q(\beta)$, the KL optimal updates turn out to factorize further as

$$q(z, w, \beta) = \Big(\prod_{i,\ell} q(z_{i\ell})\Big)\Big(\prod_{i} q(w_i)\Big)\Big(\prod_{k} q(\beta_k)\Big).$$

  This is fairly common in classic VI.

- Further, the KL optimal distributions for $q(w_i)$ and $q(\beta_k)$ are Dirichlet, even though we didn't put any constraints on their functional form.

- As is typical with classic VI, a fair amount of effort is required to derive the updates, but the algorithm itself ends up being relatively simple.

LDA: Associated Press (AP) corpus example

- Example from original LDA article (Blei et al., 2003).
- TREC AP corpus: $n = 16{,}333$ newswire articles.
- Vocabulary size: $V = 23{,}075$ unique terms (unique words).
- It is necessary to remove *stop words*, that is, very common words like "a", "the", "and", etc.
- Number of topics: Chose to use $K = 100$ topics.
- For posterior computation, they used VI for z and w, and used expectation-maximization for β.

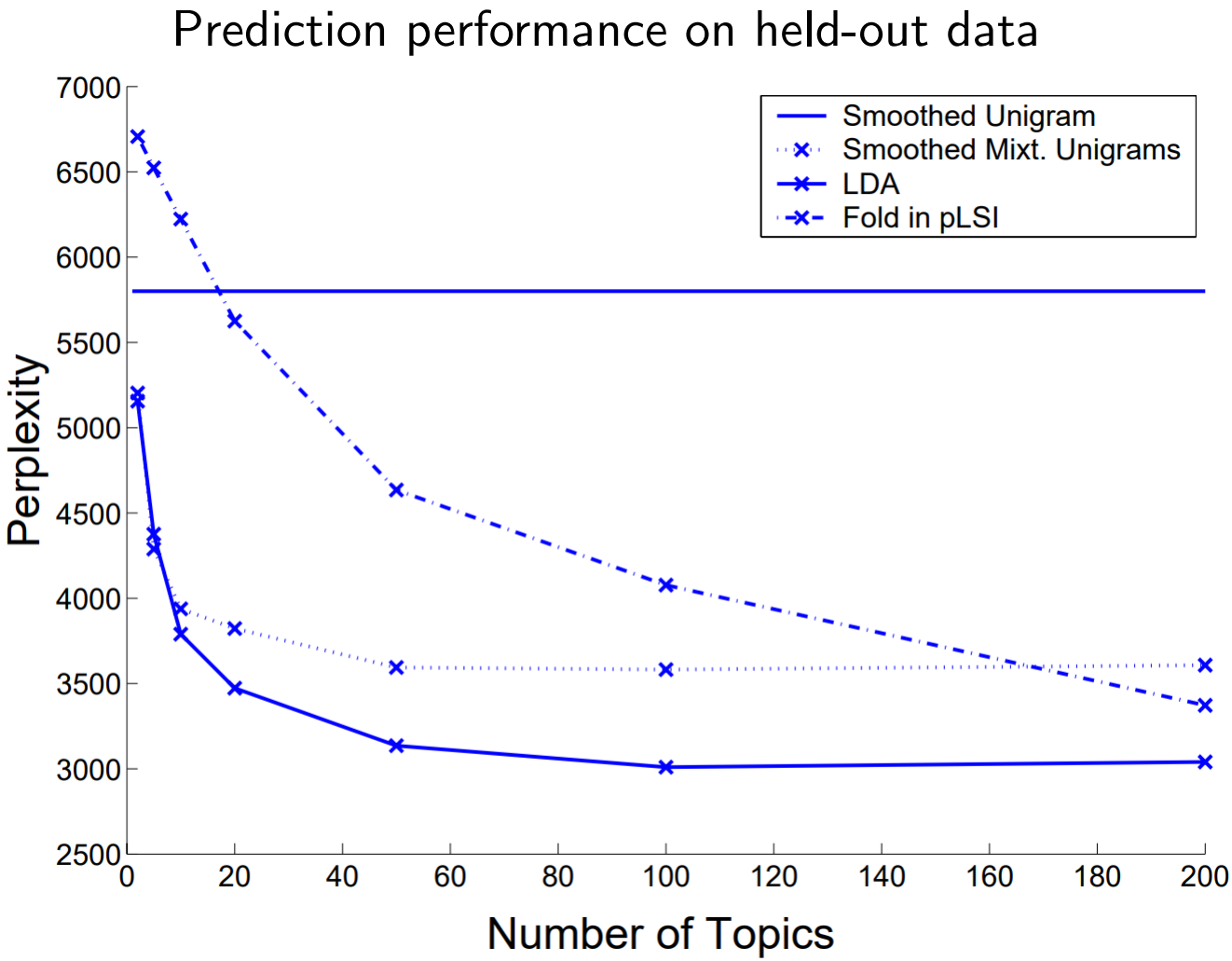
Latent Dirichlet allocation: AP corpus example

"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.

(figure from Blei et al., 2003)

Latent Dirichlet allocation: AP corpus example

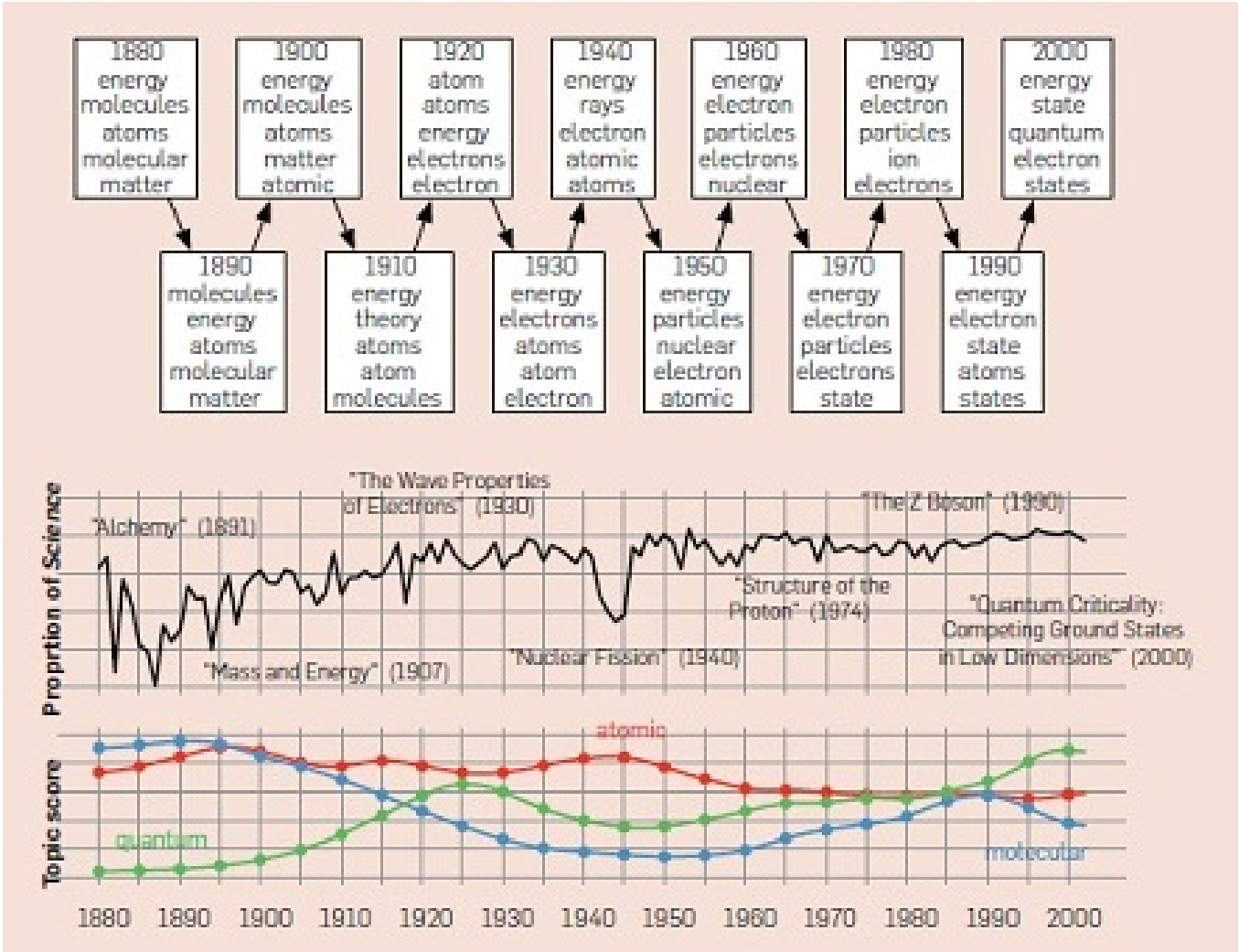


(figure from Blei et al., 2003)

Latent Dirichlet allocation: Model extensions

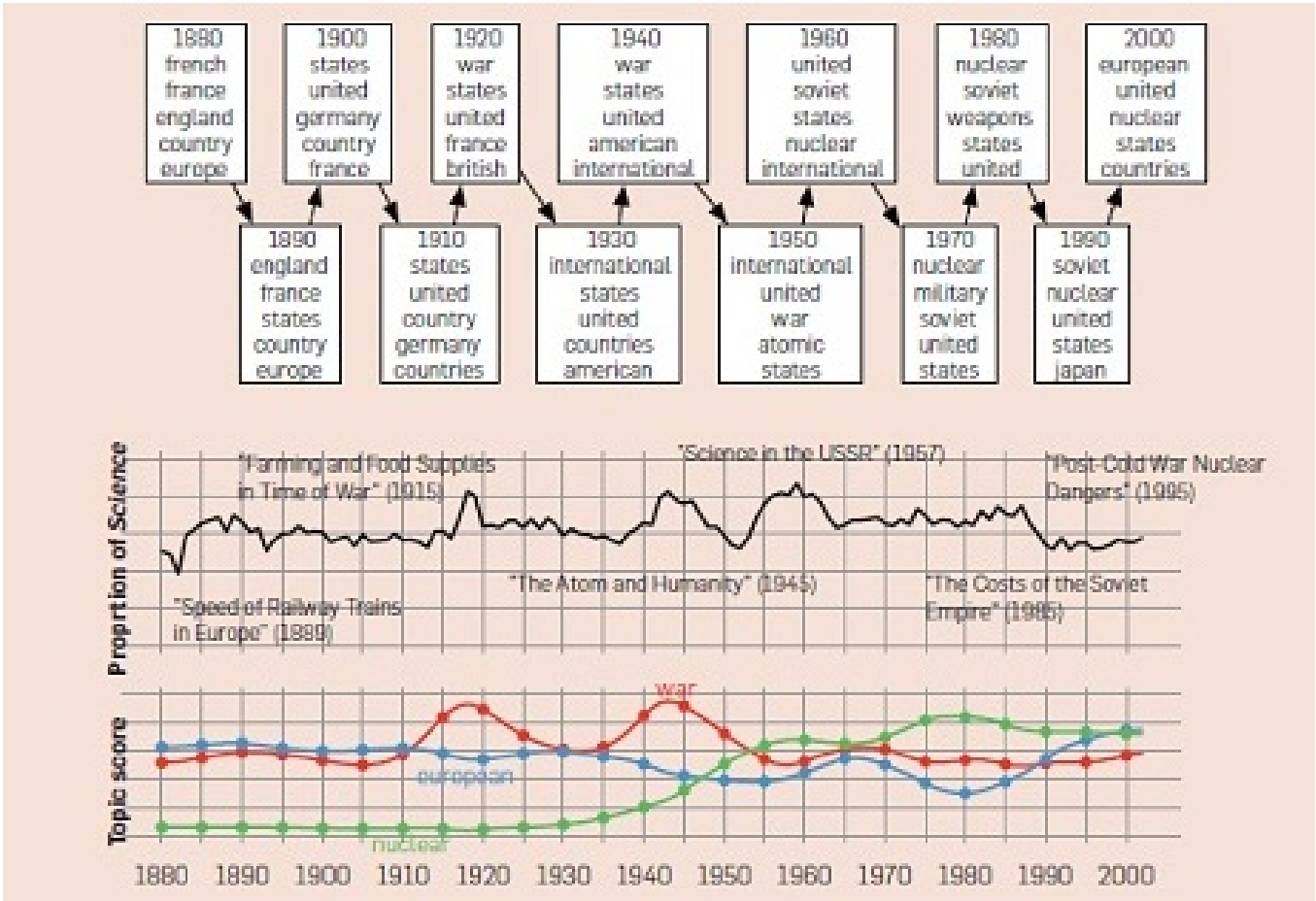
- LDA + Hidden Markov model, capturing dependencies among nearby words.
- Nonparametric topic model based on Dirichlet process, using infinitely many topics ($K \to \infty$).
- Dynamic topic model, capturing change in topics that evolve over time.
- Hierarchical topic model, using a tree of topics, from more general to more concrete.
- Extensions to account for metadata such as author, title, location, etc.

Dynamic topic model applied to Science articles



(figure from Blei, 2012)

Dynamic topic model applied to Science articles



(figure from Blei, 2012)

References and supplements

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.
- Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Individual activity: Exit ticket

Answer these questions individually:
https://forms.gle/VfoMKU1bGHDRXSyy9