Stochastic Variational Inference

Jing Zhao

Outline

- Variational Inference
  - Mean field
- Exponential Family
  - Natural parameter， sufficient statistics， base measure， log-normalizer
- Natural Gradient
- Stochastic Variational Inference
- SVI for LDA
- Distributed Variational Inference

Variational Inference

- The log marginal probability can be decomposed as

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

- Maximize the lower bound $L(q)$ is equivalent to minimizing the KL divergence

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

Variational Inference vs EM

- Often applied to $P(\boldsymbol{x}, \theta | \boldsymbol{y})$
  ($\boldsymbol{y}$ observed, $\boldsymbol{x}$ latent nuisance, θ latent parameters)

Expectation maximization	Variational Bayes
$\max_{\theta} \log \int P(\boldsymbol{y}, \boldsymbol{x} \mid \theta)\, d\boldsymbol{x}$	$\log \int P(\boldsymbol{y}, \boldsymbol{x} \mid \theta)\, d\boldsymbol{x}\, d\theta$
$\geq \max_{\theta, Q(\boldsymbol{x})} \Big\{ \mathrm{E}_Q[\log P(\boldsymbol{y}, \boldsymbol{x} \mid \theta)] + \mathrm{H}[Q(\boldsymbol{x})] \Big\}$	$\geq \max_{Q(\theta), Q(\boldsymbol{x})} \Big\{ \mathrm{E}_Q[\log P(\boldsymbol{y}, \boldsymbol{x} \mid \theta)] + \mathrm{H}[Q(\boldsymbol{x})] + \mathrm{H}[Q(\theta)] \Big\}$

Factorization assumption: $Q(\boldsymbol{x}, \theta) = Q(\boldsymbol{x}) Q(\theta)$

VI: Mean Field

- $q(\mathbf{Z})$ factorizes with respect to these groups, so that

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

- The lower bound is

$$\mathcal{L}(q) = \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z}$$

- A general expression for the optimal solution

$$\ln q_j^{\star}(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$$

Model Specification

- The joint distribution factorizes into a global term and a product of local terms

$$p(x, z, \beta \mid \alpha) = p(\beta \mid \alpha) \prod_{n=1}^{N} p(x_n, z_n \mid \beta)$$

$$p(x_n, z_n \mid x_{-n}, z_{-n}, \beta, \alpha) = p(x_n, z_n \mid \beta, \alpha)$$

- Our goal is to approximate the posterior distribution of the hidden variables given the observations $p(\beta, z \mid x)$

Model Specification

- Assumptions about the "complete conditionals" in the model

"the conditional distribution of a hidden variable given the other hidden variables and the observations"

$$p(\beta \mid x, z, \alpha) = h(\beta) \exp\{\eta_g(x, z, \alpha)^\top t(\beta) - a_g(\eta_g(x, z, \alpha))\},$$

$$p(z_{nj} \mid x_n, z_{n,-j}, \beta) = h(z_{nj}) \exp\{\eta_\ell(x_n, z_{n,-j}, \beta)^\top t(z_{nj}) - a_\ell(\eta_\ell(x_n, z_{n,-j}, \beta))\}$$

exponential family

Exponential Family

$$p(\beta \mid x, z, \alpha) = h(\beta) \exp\{\eta_g(x, z, \alpha)^\top t(\beta) - a_g(\eta_g(x, z, \alpha))\}$$

- base measure
- log-normalizer
- natural parameter
- sufficient statistics

Exponential Family

$$f_X(x \mid \boldsymbol{\eta}) = h(x) \exp\left(\boldsymbol{\eta} \cdot \mathbf{T}(x) - A(\boldsymbol{\eta})\right)$$

Distribution	Para	Natural parameter	Inverse parameter mapping	Base measure	Sufficient statistic	Log-partition	Log-partition
Normal	μ,σ^2	$\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$	$\begin{bmatrix} -\frac{\eta_1}{2\eta_2} \\ -\frac{1}{2\eta_2} \end{bmatrix}$	$\frac{1}{\sqrt{2\pi}}$	$\begin{bmatrix} x \\ x^2 \end{bmatrix}$	$-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)$	$\frac{\mu^2}{2\sigma^2} + \ln\sigma$
Bernoulli	p	$\ln\frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	1	x	$\ln(1+e^{\eta})$	$-\ln(1-p)$

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

$$\boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^{\mathrm{T}}$$

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

$$T(x) = \left(x, x^2\right)^{\mathrm{T}}$$

$$A(\boldsymbol{\eta}) = \frac{\mu^2}{2\sigma^2} + \log|\sigma| = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left|\frac{1}{2\eta_2}\right|$$

Other distributions

- "complete conditionals" imply a conjugacy
- The distribution of the local context given the global variables must be in an exponential family

$$p(x_n, z_n|\beta) = h(x_n, z_n)\exp\{\beta^\top t(x_n, z_n) - a_\ell(\beta)\}$$

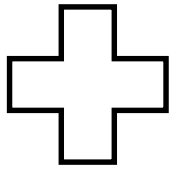
- The prior distribution $p(\beta)$ must also be in an exponential family

$$p(\beta) = h(\beta)\exp\{\alpha^\top t(\beta) - a_g(\alpha)\}$$

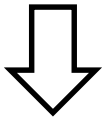
$$t(\beta) = (\beta, -a_\ell(\beta)) \qquad \alpha = (\alpha_1, \alpha_2)$$

About Natural Paramters

- $p(\beta \mid x, z, \alpha) = h(\beta) \exp\{\eta_g(x, z, \alpha)^\top t(\beta) - a_g(\eta_g(x, z, \alpha))\}$



- $p(x_n, z_n|\beta)$ and $p(\beta)$



- Natural parameter

$$\eta_g(x, z, \alpha) = (\alpha_1 + \sum_{n=1}^{N} t(z_n, x_n), \alpha_2 + N)$$

Variational Inference for Exponential

- The lower bound on the log marginal

$$\mathbb{E}_q[\log p(x, z, \beta)] - \mathbb{E}_q[\log q(z, \beta)]$$

- Variational distribution

$$q(z, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj} \mid \phi_{nj})$$

- In the same exponential family as the complete conditional distributions

$$q(\beta \mid \lambda) = h(\beta) \exp\{\lambda^\top t(\beta) - a_g(\lambda)\},$$
$$q(z_{nj} \mid \phi_{nj}) = h(z_{nj}) \exp\{\phi_{nj}^\top t(z_{nj}) - a_\ell(\phi_{nj})\}$$

Variational Inference for Exponential

- Optimize each coordinate in closed form
- Rewrite the objective w.r.t. global parameter as

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta \mid x, z)] - \mathbb{E}_q[\log q(\beta)] + \text{const.}$$

$$\mathbb{E}_q[t(\beta)] = \nabla_\lambda a_g(\lambda)$$

$$q(\beta \mid \lambda) = h(\beta)\exp\{\lambda^\top t(\beta) - a_g(\lambda)\},$$

$$p(\beta \mid x, z, \alpha) = h(\beta)\exp\{\eta_g(x, z, \alpha)^\top t(\beta) - a_g(\eta_g(x, z, \alpha))\}$$

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\eta_g(x, z, \alpha)]^\top \nabla_\lambda a_g(\lambda) - \lambda^\top \nabla_\lambda a_g(\lambda) + a_g(\lambda) + \text{const}$$

- Setting the gradient to zero results in

$$\lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)]$$

recall : $\eta_g(x, z, \alpha) = (\alpha_1 + \sum_{n=1}^{N} t(z_n, x_n), \alpha_2 + N)$

Variational Inference for Exponential

- Turn to the local parameters

$$\nabla_{\phi_{nj}} \mathcal{L} = \nabla^2_{\phi_{nj}} a_\ell(\phi_{nj})(\mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)] - \phi_{nj})$$

- It equals zero when

$$\phi_{nj} = \mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)]$$

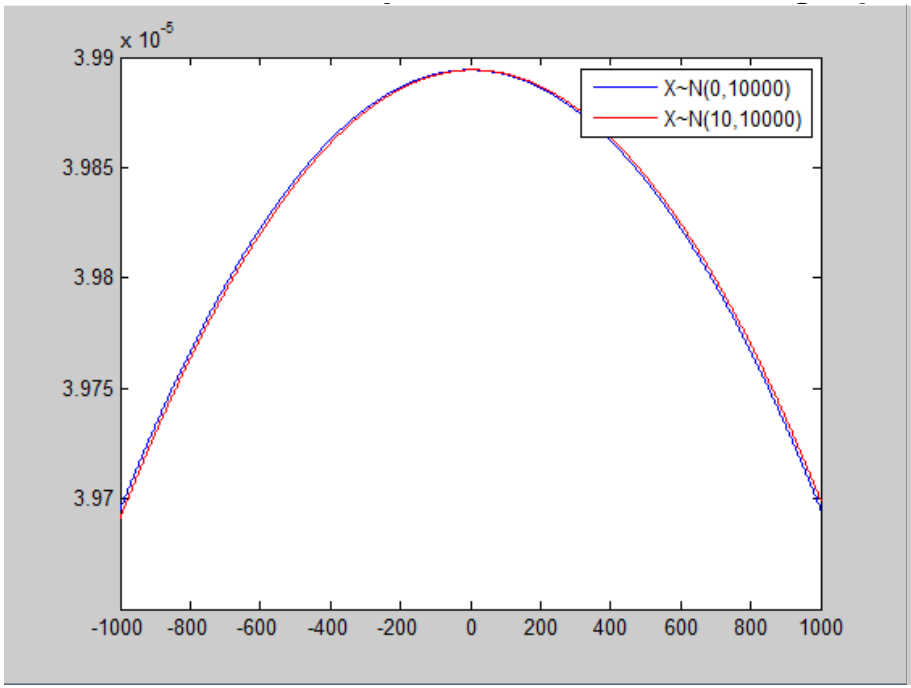
- Recall

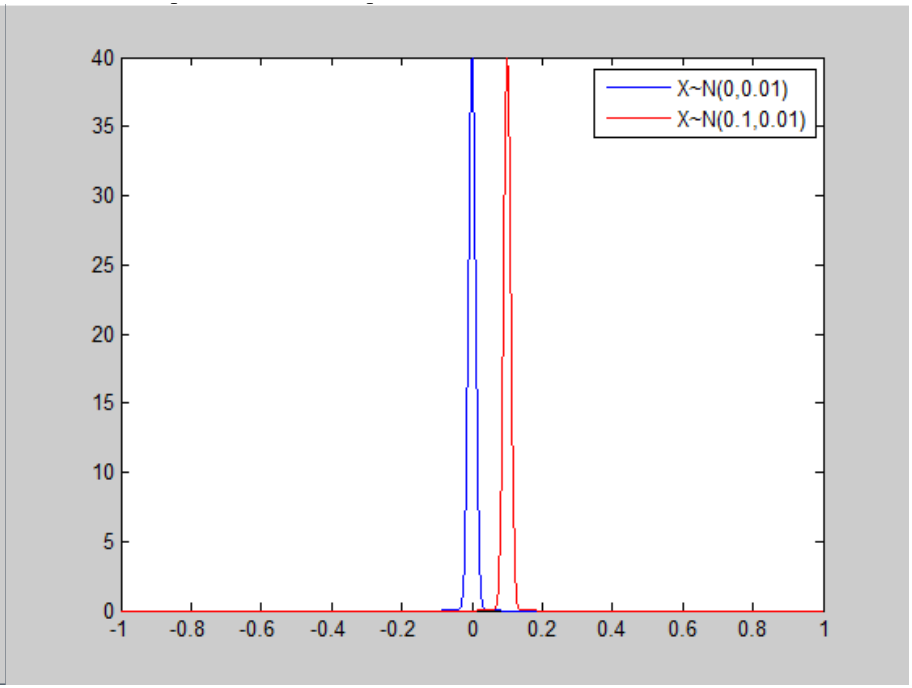
$$\lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)]$$

$$\eta_g(x, z, \alpha) = (\alpha_1 + \textstyle\sum_{n=1}^{N} t(z_n, x_n), \alpha_2 + N)$$

Natural Gradient

- N(0,10000) and N(10,10000)
- N(0,0.01) and N(0.1,0.01)





Natural Gradient

- N(0,10000) and N(10,10000)
- N(0,0.01) and N(0.1,0.01)
- A natural measure of dissimilarity between probability distributions is the symmetrized KL divergence

$$D_{KL}^{\text{sym}}(\lambda, \lambda') = \mathbb{E}_{\lambda}\left[\log \frac{q(\beta|\lambda)}{q(\beta|\lambda')}\right] + \mathbb{E}_{\lambda'}\left[\log \frac{q(\beta|\lambda')}{q(\beta|\lambda)}\right]$$

$$\arg\max_{d\lambda} f(\lambda + d\lambda) \quad \text{subject to } D_{KL}^{\text{sym}}(\lambda, \lambda + d\lambda) < \varepsilon.$$

As $\varepsilon \to 0$, the solution to this problem points in the same direction as the *natural gradient*. While the Euclidean gradient points in the direction of steepest ascent in Euclidean space, the natural gradient points in the direction of steepest ascent in the Riemannian space, that is, the space where local distance is defined by KL divergence rather than the L^2 norm.

Natural Gradient and Gradient

- Transformation G: Fisher information matrix

$$\hat{\nabla}_\lambda f(\lambda) \triangleq G(\lambda)^{-1} \nabla_\lambda f(\lambda)$$

$$d\lambda^T G(\lambda) d\lambda = D_{KL}^{\text{sym}}(\lambda, \lambda + d\lambda)$$

$$d^* = \arg\min_d \mathcal{L}(\lambda + d) + \alpha(\mathrm{KL}[p_\lambda \| p_{\lambda+d}] - c)$$

$$\approx \arg\min_d \mathcal{L}(\lambda) + \nabla_\lambda \mathcal{L}(\lambda)^{\mathrm{T}} d + \frac{1}{2}\alpha d^{\mathrm{T}} \bar{G} d - \lambda c.$$

- How to calculate G?

$$G(\lambda) = \mathbb{E}_\beta \left[(\nabla_\lambda \log q(\beta \mid \lambda))(\nabla_\lambda \log q(\beta \mid \lambda))^\top \right] \qquad F = -E_{p(\beta \mid \lambda)} \left[\frac{\partial^2}{\partial \lambda \partial \lambda^T} \log p(\beta \mid \lambda) \right]$$

- Using properties of exponential family:

$$\begin{aligned} G(\lambda) &= \mathbb{E}_\beta \left[(\nabla_\lambda \log q(\beta \mid \lambda))(\nabla_\lambda \log q(\beta \mid \lambda))^\top \right] \\ &= \mathbb{E}_\beta \left[(t(\beta) - \mathbb{E}_\beta[t(\beta)])(t(\beta) - \mathbb{E}_\beta[t(\beta)])^\top \right] \\ &= \nabla_\lambda^2 a_g(\lambda). \end{aligned}$$

- Recall:

$$\nabla_{\phi_{nj}} \mathcal{L} = \nabla^2_{\phi_{nj}} a_\ell(\phi_{nj})(\mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)] - \phi_{nj})$$

$$\nabla_\lambda \mathcal{L} = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda).$$

$$\hat{\nabla}_\lambda f(\lambda) \triangleq G(\lambda)^{-1} \nabla_\lambda f(\lambda) \qquad G(\lambda) = \nabla^2_\lambda a_g(\lambda)$$

- Natural gradient

$$\hat{\nabla}_\lambda \mathcal{L} = \mathbb{E}_\phi[\eta_g(x, z, \alpha)] - \lambda$$

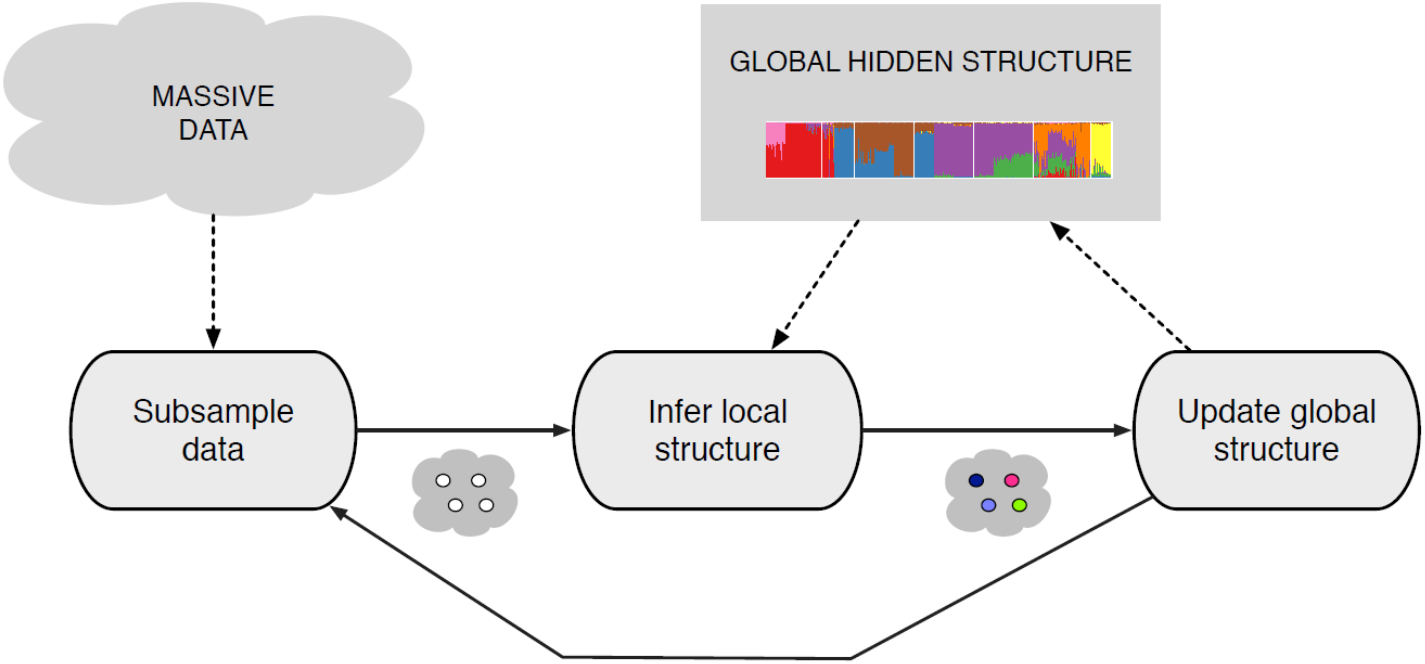
$$\hat{\nabla}_{\phi_{nj}} \mathcal{L} = \mathbb{E}_{\lambda, \phi_{n,-j}}[\eta_\ell(x_n, z_{n,-j}, \beta)] - \phi_{nj}$$

- gradient set to 0:

$$\lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)]$$

$$\phi_{nj} = \mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)]$$

Stochastic Variational Inference



"Stochastic variational inference" [Hoffman et al., 2013, JMLR]

Stochastic Variational Inference

- Stochastic optimization

the gradient can be written as a sum of terms (one for each data point) and we can compute a fast noisy approximation by subsampling the data

- Stochastic variational inference
  - Sample a data, optimize local variational parameters
  - Form intermediate global variational parameters
  - Update the global variational parameters

- Decompose Linto a global term and a sum of local terms

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + \sum_{n=1}^{N} \max_{\phi_n} (\mathbb{E}_q[\log p(x_n, z_n \mid \beta)] - \mathbb{E}_q[\log q(z_n)]).$$

Now consider a variable that chooses an index of the data uniformly at random, $I \sim \text{Unif}(1, \ldots, N)$. Define $\mathcal{L}_I(\lambda)$ to be the following random function of the variational parameters,

$$\mathcal{L}_I(\lambda) \triangleq \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + N \max_{\phi_I} (\mathbb{E}_q[\log p(x_I, z_I \mid \beta) - \mathbb{E}_q[\log q(z_I)]). \quad (25)$$

$$\hat{\nabla} \mathcal{L}_i = \mathbb{E}_q \left[\eta_g \left(x_i^{(N)}, z_i^{(N)}, \alpha \right) \right] - \lambda$$

Stochastic Variational Inference

1: Initialize $\lambda^{(0)}$ randomly.
2: Set the step-size schedule ρ_t appropriately.
3: **repeat**
4: Sample a data point x_i uniformly from the data set.
5: Compute its local variational parameter,

$$\text{Set local parameter } \phi \leftarrow \mathbb{E}_\lambda\left[\eta_\ell(\beta, x_j)\right].$$

6: Compute intermediate global parameters as though x_i is replicated N times,

$$\hat{\lambda} = \mathbb{E}_\phi[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

7: Update the current estimate of the global variational parameters,

$$\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}.$$

8: **until** forever

Topic Model

Motivation: Topic Modeling

1. **Discover** the thematic structure in a large collection of documents
2. **Annotate** the documents
3. **Use** the annotations to visualize, organize, summarize, ...

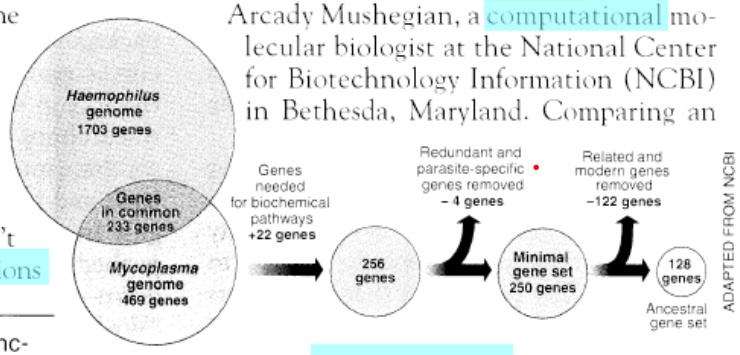
Topic Model

Example: Latent Dirichlet allocation (LDA)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

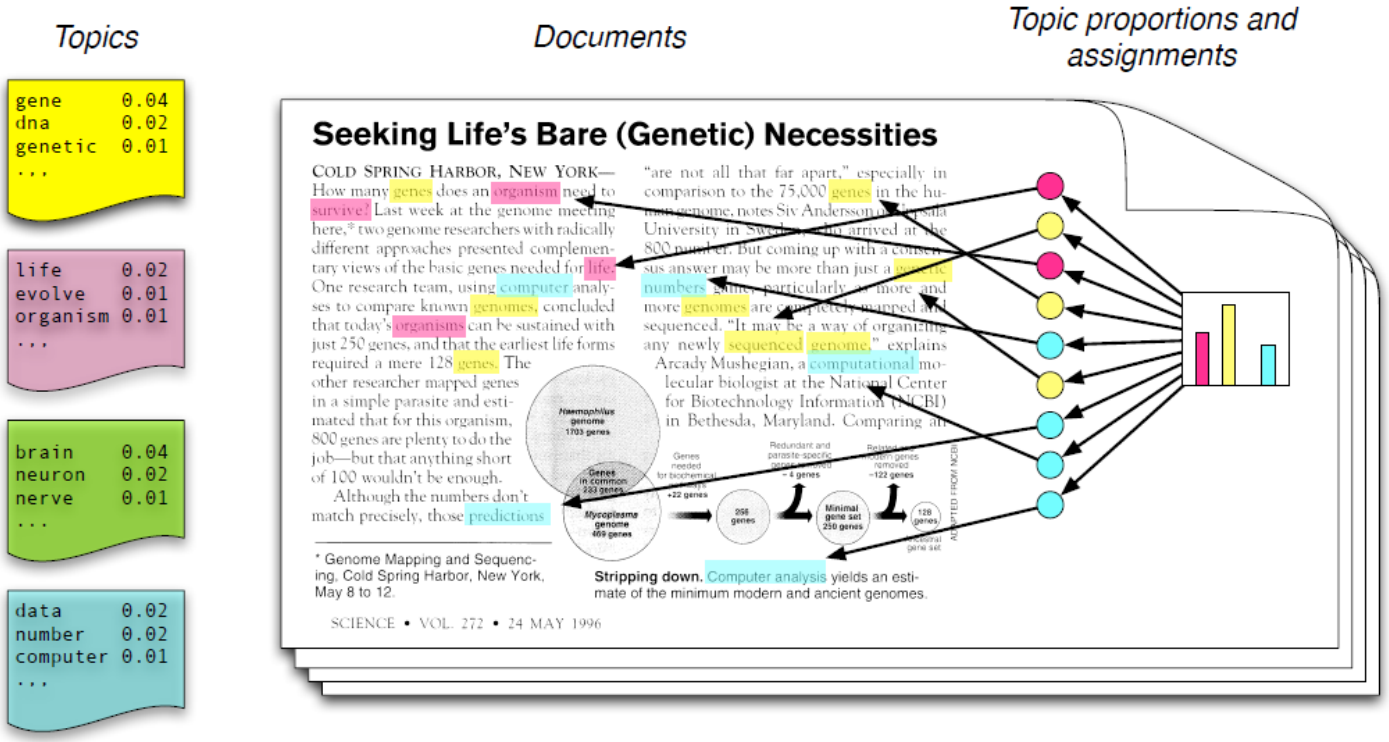
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Documents exhibit multiple topics.

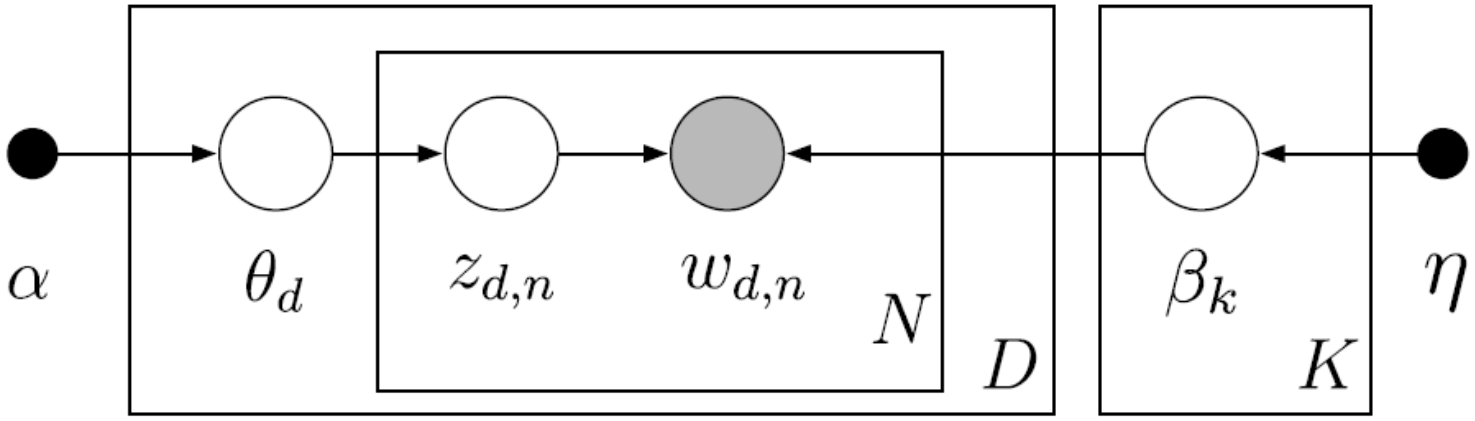
Latent Dirichlet Allocation

Example: Latent Dirichlet allocation (LDA)



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

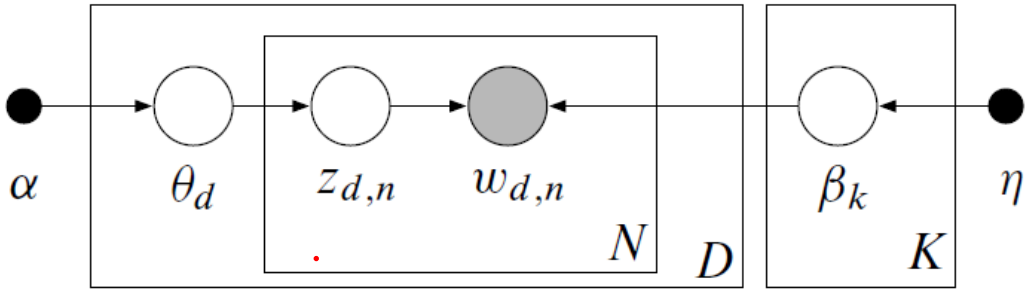
Latent Dirichlet Allocation



- 1. Draw topics $\beta_k \sim \text{Dirichlet}(\text{n}.\ldots.\text{n})$

  A *topic* β_k is a distribution over the vocabulary.
- 2. For each document :
  - (a) Draw topic proportions $\theta \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$
  - (b) For each word :
    - Draw topic assignment $z_{dn} \sim \text{Multinomial}(\theta_d)$
    - Draw word $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$

Posterior Inference For LDA



- The posterior of the latent variables given the documents is

$$p(\beta, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}) = \frac{p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{\int_{\beta} \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}.$$

- We can't compute the denominator, the marginal $p(\mathbf{w})$.
- We use approximate inference.

Latent Dirichlet Allocation

- Complete conditionals

$$p(z_{dn} = k \mid \theta_d, \beta_{1:K}, w_{dn}) \propto \exp\{\log \theta_{dk} + \log \beta_{k,w_{dn}}\}$$

$$p(\theta_d \mid z_d) = \text{Dirichlet}\left(\alpha + \sum_{n=1}^{N} z_{dn}\right)$$

$$p(\beta_k \mid z, w) = \text{Dirichlet}\left(\eta + \sum_{d=1}^{D} \sum_{n=1}^{N} z_{dn}^{k} w_{dn}\right)$$

- Variational distribution

$$q(z_{dn}) = \text{Multinomial}(\phi_{dn})$$

$$q(\beta_k) = \text{Dirichlet}(\lambda_k)$$

$$q(\theta_d) = \text{Dirichlet}(\gamma_d)$$

Inference For LDA

- Recall coordinate update or one step natural gradient

$$\phi_{nj} = \mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)] \quad \text{(local)}$$

$$\lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)] \quad \text{(global)}$$

- We have

$$\phi_{dn}^k \propto \exp\{\Psi(\gamma_{dk}) + \Psi(\lambda_{k,w_{dn}}) - \Psi(\textstyle\sum_v \lambda_{kv})\} \quad \text{for } n \in \{1, \ldots, N\},$$

$$\gamma_d = \alpha + \textstyle\sum_{n=1}^N \phi_{dn}.$$

(local)

$$\lambda_k = \eta + \textstyle\sum_{d=1}^D \sum_{n=1}^N \phi_{dn}^k w_{dn} \quad \text{(global)}$$

SVI for LDA

1: Initialize $\lambda^{(0)}$ randomly.
2: Set the step-size schedule ρ_t appropriately.
3: **repeat**
4: Sample a document w_d uniformly from the data set.
5: Initialize $\gamma_{dk} = 1$, for $k \in \{1, \ldots, K\}$.
6: **repeat**
7: For $n \in \{1, \ldots, N\}$ set

$$\phi_{dn}^k \propto \exp\{\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{k,w_{dn}}]\}, k \in \{1, \ldots, K\}.$$

8: Set $\gamma_d = \alpha + \sum_n \phi_{dn}$.
9: **until** local parameters ϕ_{dn} and γ_d converge.
10: For $k \in \{1, \ldots, K\}$ set intermediate topics

$$\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi_{dn}^k w_{dn}.$$

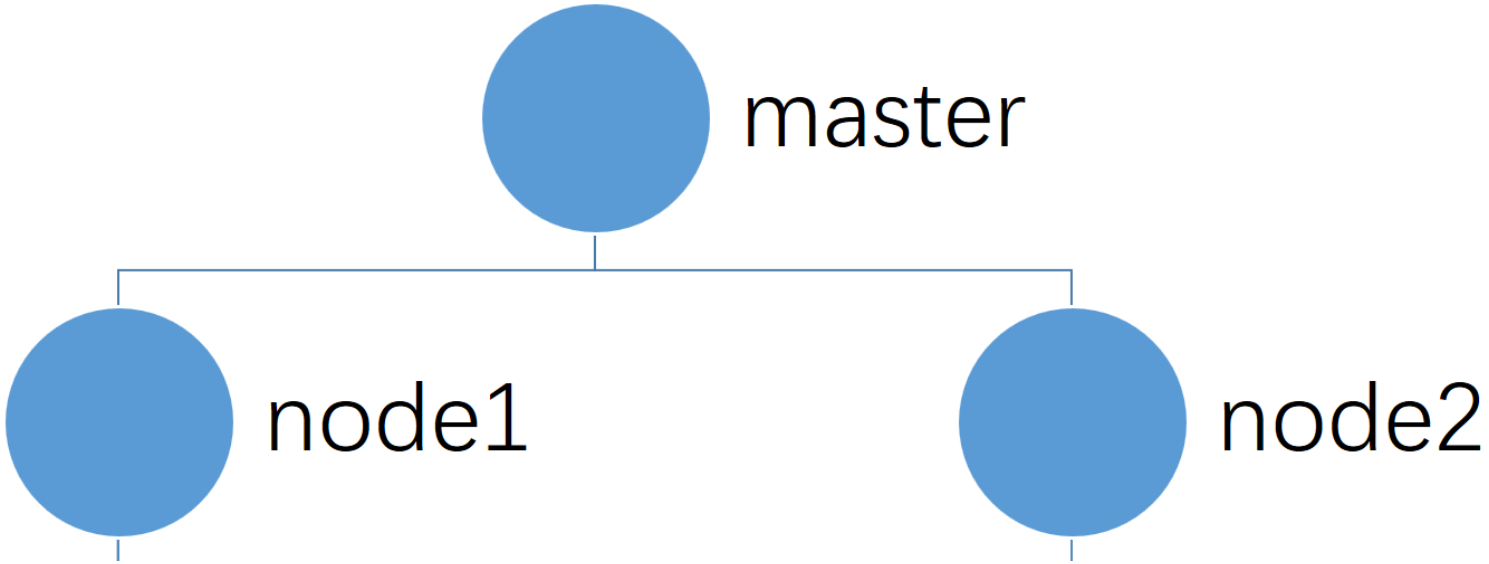
11: Set $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$.
12: **until** forever

Question

- How about non exponential family?
- SVI for GP?
- SVI for HMM?
- How to improve?

Distributed Variational Inference

Distributed framework



Different ways of optimizing lower bound

- 1 . Try to eliminate partial
  - Collapsed variational inference
- 2. try to find the optimal variational distribution
  - Lagrange multipliers to optimize q(.)
- 3. Optimize the variational distribution
  - Define the formulation of q(.|\theta) and optimize theta

(GP for big data use 3 for SVI)

Distributed inference formulation

- F = sum $f(x_i)$ =>SVI and DVI
  - Select mini-batch of X_s
  - Calculate objective F_s and its gradient on partial data

- F = g(sum $f(x_i)$) =>DVI
  - g is calculated on the master
  - $f(x_i)$ is calculated on the nodes

Rewrite the lower bound

- $$\mathcal{L}_2 = \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} + \beta^{-1}\mathbf{I}\right) - \frac{1}{2}\beta \mathrm{tr}\left(\widetilde{\mathbf{K}}\right)$$

  $$\widetilde{\mathbf{K}} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$$

- $$\log p(Y|X) \geq \frac{d}{2}\log|K_{mm}| - \frac{d}{2}\log\left|K_{mm} + \beta\sum_{i=1}^{n} K_{mi}K_{im}\right| - \frac{nd}{2}\log 2\pi\beta^{-1}$$

  $$- \frac{\beta}{2}\sum_{i=1}^{n}\left(Y_i Y_i^T + d\cdot k(X_i, X_i) - d\cdot \mathrm{Tr}(K_{mm}^{-1}K_{mi}K_{im})\right)$$

  $$+ \frac{\beta^2}{2}\mathrm{Tr}\left(\left(\sum_{i=1}^{n} K_{mi}Y_i\right)^T\left(K_{mm} + \beta\sum_{i=1}^{n} K_{mi}K_{im}\right)^{-1}\left(\sum_{i=1}^{n} K_{mi}Y_i\right)\right)$$

Rewrite the lower bound

- $$\log p(Y) \geq \log \left[\frac{(\beta)^{\frac{N}{2}} |K_{MM}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}} |\beta\Psi_2 + K_{MM}|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{y}_d^T W \mathbf{y}_d} \right]$$
$$- \frac{\beta\psi_0}{2} + \frac{\beta}{2}\mathrm{Tr}\left(K_{MM}^{-1}\Psi_2\right),$$
- $$- \sum_{i=1}^{n} KL(q(X_i)||p(X_i))$$

$$\log p(Y) \geq \frac{d}{2}\log|K_{mm}| - \frac{d}{2}\log\left|K_{mm} + \beta\sum_{i=1}^{n}\left\langle K_{mi}^{X_t} K_{im}^{X_t}\right\rangle_{q(X_t)}\right| - \frac{nd}{2}\log 2\pi\beta^{-1} \tag{4.19}$$

$$- \frac{\beta}{2}\sum_{i=1}^{n}\left(Y_i Y_i^T + d\left\langle K_{ii}^{X_t}\right\rangle_{q(X_t)} - d\mathrm{Tr}\left(K_{mm}^{-1}\left\langle K_{mi}^{X_t} K_{im}^{X_t}\right\rangle_{q(X_t)}\right)\right) \tag{4.20}$$

$$+ \frac{\beta^2}{2}\mathrm{Tr}\left(\left(\sum_{i=1}^{n}\left\langle K_{mi}^{X_t}\right\rangle_{q(X_t)} Y_i\right)^T \left(K_{mm} + \beta\sum_{i=1}^{n}\left\langle K_{mi}^{X_t} K_{im}^{X_t}\right\rangle_{q(X_t)}\right)^{-1} \cdot \left(\sum_{i=1}^{n}\left\langle K_{mi}^{X_t}\right\rangle_{q(X_t)} Y_i\right)\right) \tag{4.21}$$

$$- \sum_{i=1}^{n} KL(q(X_i)||p(X_i)) \tag{4.22}$$

Reference

Hoffman M D, Blei D M, Wang C, et al. Stochastic variational inference[J]. The Journal of Machine Learning Research, 2013, 14(1): 1303-1347.

Hensman J, Fusi N, Lawrence N D. Gaussian Processes for Big Data[C]// Uncertainty in Artificial Intelligence. 2013: 282.

Gal Y, Van Der Wilk M, Rasmussen C E. Distributed variational inference in sparse Gaussian process regression and latent variable models[C]//Advances in neural information processing systems. 2014: 3257-3265.