Optimization and Probability in Machine Learning

Neal Parikh

http://nparikh.org

August 19, 2019

Machine learning

- intersection of computer science and statistics
- algorithms that learn from data and/or improve with experience
- roots in AI, but also used in huge variety of other domains
- draws on a number of mathematical areas, but rests in particular on optimization theory and probability theory

- by framing various tasks that don't appear to involve optimization in that way, can bring the optimization toolbox to bear on a wider range of problems in machine learning

Outline

Regularized loss minimization

optimization-based framework for learning (parameter estimation)

$$\text{minimize} \quad l(w) + \lambda r(w)$$

- $w \in \mathbf{R}^n$ are the **model parameters** or **weights**
- $l : \mathbf{R}^n \to \mathbf{R}$ is a **loss function** measuring lack of fit on training data
- $r : \mathbf{R}^n \to \mathbf{R}$ is a **regularizer** measuring model complexity
- $\lambda > 0$ is a **regularization parameter** trading off between the two

if l and r are convex, problem can essentially be solved

Regularized loss minimization

- many models fall under this framework, and may or may not be probabilistic in nature
- ridge regression (Tikhonov regularized linear regression)

$$\text{minimize} \quad \|Xw - y\|_2^2 + \lambda \|w\|_2^2$$

- lasso (sparse linear regression)

$$\text{minimize} \quad \|Xw - y\|_2^2 + \lambda \|w\|_1$$

- support vector machine

$$\text{minimize} \quad \sum_{i=1}^{N} (1 - y_i w^T x_i)_+ + \lambda \|w\|_2^2$$

Regularized loss minimization

- also subsumes common ways of learning probabilistic models
- maximum likelihood estimation of model $p(x)$

$$\text{minimize} \quad -\sum_{i=1}^{N} \log p(x_i)$$

  - could also model $y \,|\, x$ and maximize $\sum_i \log p(y_i \,|\, x_i)$
  - includes, *e.g.*, linear regression, generalized linear models, . . .

- maximum a posteriori (MAP) estimation (finding posterior mode)

$$\text{minimize} \quad -\sum_{i=1}^{N} \log p(x_i \,|\, w) - \log p(w)$$

  - $w \sim$ Gaussian: Tikhonov regularization
  - $w \sim$ Laplacian: ℓ_1/lasso regularization

Naive Bayes

- joint model that factorizes as

$$p(\mathbf{x}, y) = p(y) \prod_{i=1}^{n} p(x_i \,|\, y)$$

  where x_i are words in an email and $y \in \{\text{spam}, \text{ham}\}$

- need to estimate $p(y = \text{spam})$ and $p(x = \text{bank} \,|\, y)$

- maximum likelihood estimates have a trivial closed form solution, *e.g.*:

$$\hat{p}(y = \text{spam}) = \frac{\#\text{ spam emails}}{\text{total }\#\text{ emails}}$$

$$\hat{p}(x = \text{bank} \,|\, \text{spam}) = \frac{\#\text{ times } x = \text{bank in spam mails}}{\#\text{ words in spam emails}}$$

Naive Bayes

- once parameters are fit, want to classify new example $\mathbf{x}^{\text{new}}$
- compute label posterior

$$p(y = \text{spam} \mid \mathbf{x}^{\text{new}}) = \frac{p(\mathbf{x}^{\text{new}} \mid \text{spam}) \cdot p(\text{spam})}{p(\mathbf{x}^{\text{new}})}$$

- pick whichever class has higher posterior probability
- in general, called a (probabilistic) **inference** problem

Gaussian discriminant analysis

- assume data comes from generative model

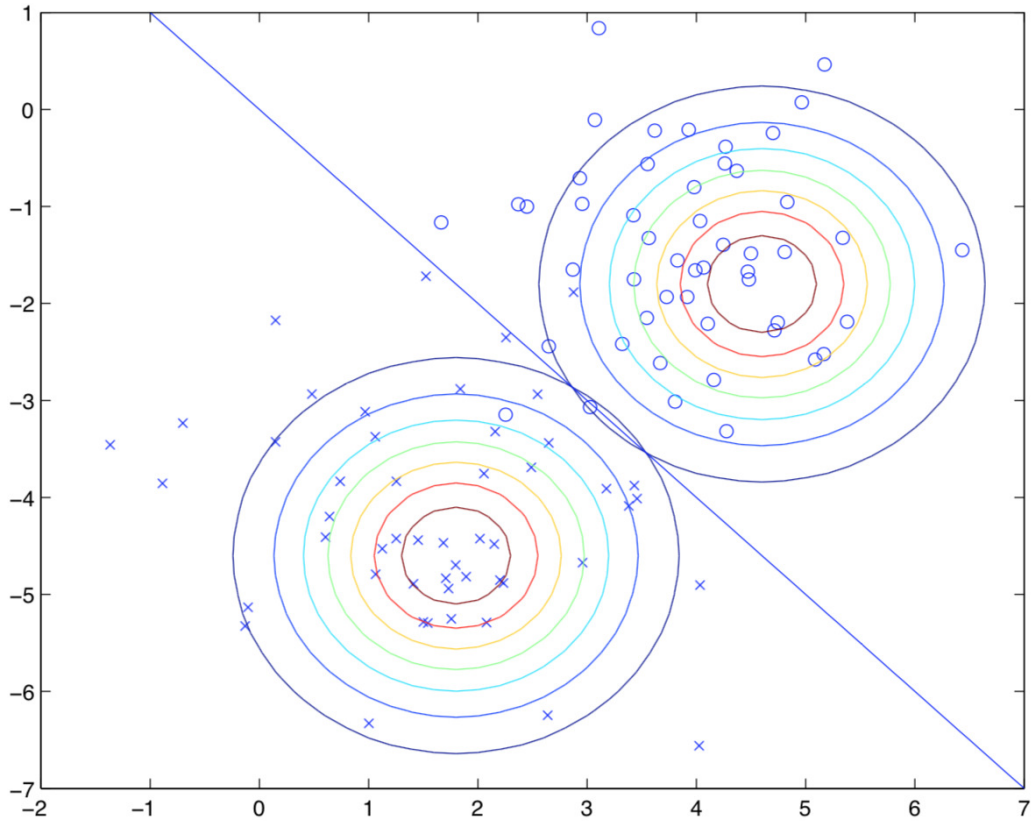
$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ x \,|\, y = 0 &\sim \text{N}(\mu_0, \Sigma) \\ x \,|\, y = 1 &\sim \text{N}(\mu_1, \Sigma) \end{aligned}$$

  i.e., data comes from one of two Gaussians chosen with a ϕ-coin flip

- estimate $w = (\phi, \mu_k, \Sigma)$ by maximizing

$$\begin{aligned} \ell(w) &= \log \prod_{i=1}^{N} p(x_i, y_i; w) \\ &= \sum_{i=1}^{N} \log p(x_i \,|\, y_i; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{N} \log p(y_i; \phi) \end{aligned}$$

Gaussian discriminant analysis



Maximum likelihood estimation

maximum likelihood estimates of parameters given by

$$\begin{aligned}
\hat{\phi} &= \frac{1}{N}\sum_{i=1}^{N}[y_i = 1] \\
\hat{\mu}_k &= \frac{\sum_{i=1}^{N}[y_i = k]x_i}{\sum_{i=1}^{N}[y_i = k]} \\
\hat{\Sigma} &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{y_i})(x_i - \mu_{y_i})^T
\end{aligned}$$

very natural interpretations:

- $\hat{\phi}$ is empirical proportion of positive label in $\mathcal{D}$
- $\hat{\mu}_k$ is empirical average of x_i with label k
- $\hat{\Sigma}$ is empirical covariance, with variance measured to relevant mean

Regularized loss minimization

- subsumes a very wide class of machine learning models
- generative/discriminative, probabilistic/non-probabilistic, ...
- sometimes problem can just be solved in closed form
- sometimes need sophisticated optimization algorithms like L-BFGS, accelerated proximal gradient method, ...
- calculations, or subsequent model use, may involve probability operations

Outline

Mixture of Gaussians

- probabilistic model for clustering / density estimation
- consider data $\mathcal{D} = \{x_1, \ldots, x_N\}$
- generative model

$$\begin{aligned} z &\sim \text{Multinomial}(\phi) \\ x \,|\, z = k &\sim \text{N}(\mu_k, \Sigma_k) \end{aligned}$$

- *i.e.*, each x_i generated by sampling a **unobserved** (hidden, latent) $z_i \in [K]$ and then drawing x_i from the corresponding Gaussian
- presence of these latent variables is the key new wrinkle
- model parameters are ϕ, μ_k, Σ_k

Maximum likelihood estimation

- model parameters are ϕ, μ_k, Σ_k
- as usual, write down likelihood for $w = (\phi, \mu_k, \Sigma_k)$

$$\begin{aligned} \ell(w) &= \sum_{i=1}^{N} \log p(x_i; w) \\ &= \sum_{i=1}^{N} \log \sum_{z_i=1}^{K} p(x_i \mid z_i) p(z_i) \end{aligned}$$

- this function is *nonconvex* due to sum over values of z_i
- can no longer easily solve the relevant optimization problem

Maximum likelihood estimation

- if z_i were known, problem is easy and becomes

$$\ell(w) = \sum_{i=1}^{N} \log p(x_i \mid z_i) + \sum_{i=1}^{N} \log p(z_i)$$

- maximizing with respect to ϕ, μ, Σ gives

$$\phi_j = \frac{1}{N} \sum_{i=1}^{N} [z_i = j], \quad \mu_j = \frac{\sum_{i=1}^{N} [z_i = j] x_i}{\sum_{i=1}^{N} [z_i = j]}$$

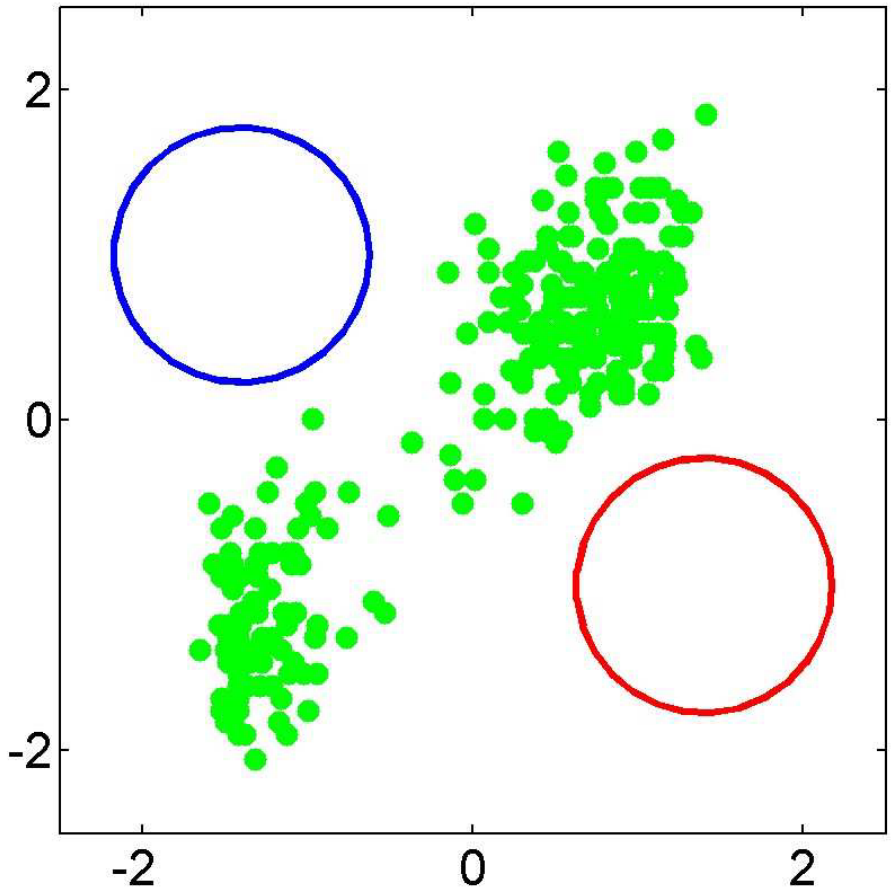
  similar expression for Σ

- *i.e.*, if z_i were known, nearly identical to maximum likelihood estimates in GDA (with z_i's as class labels)

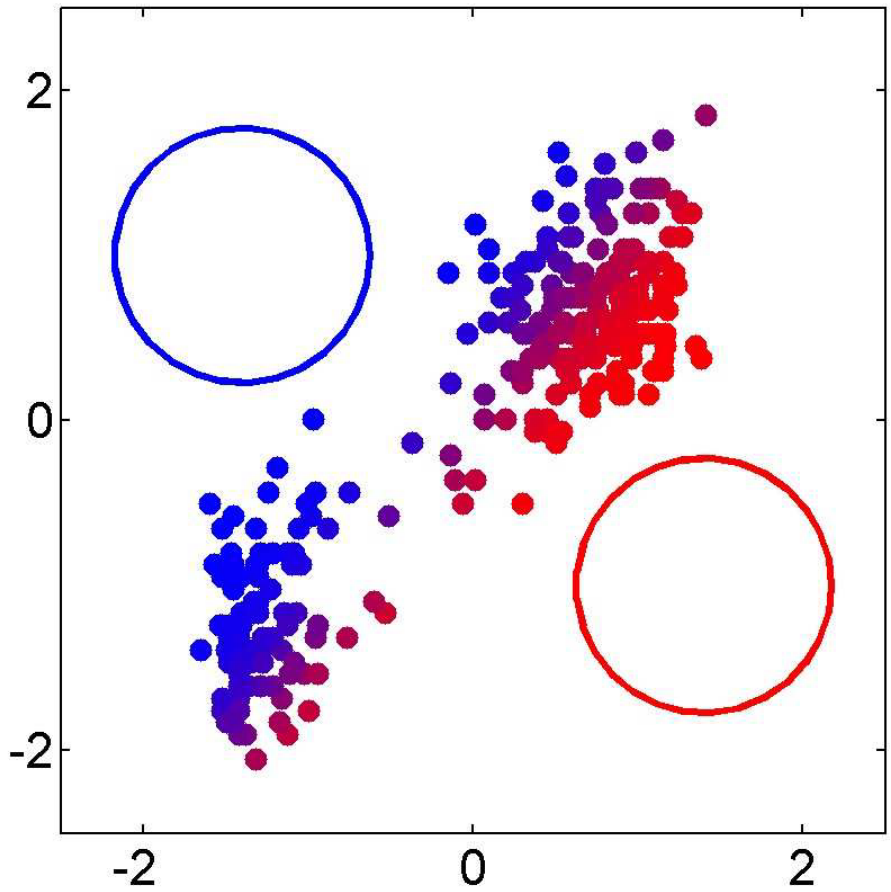
EM algorithm

- **idea**: iteratively guess the z_i and then use formulas above:
  1. E-step (probability): compute $\rho_{ij} = p(z_i = j \mid x_i; \theta, \mu, \Sigma)$
  2. M-step (optimization): use formulas above with ρ_{ij} in place of $[z_i = j]$
- E-step is an inference task: compute posterior probability of z_i's, given data and current setting of parameters; 'soft guesses' for values of z_i
- M-step is 'regular' maximum likelihood estimation, but there is uncertainty around the value of the z_i and that's incorporated in estimates
- yields a 'soft' version of k-means for this model

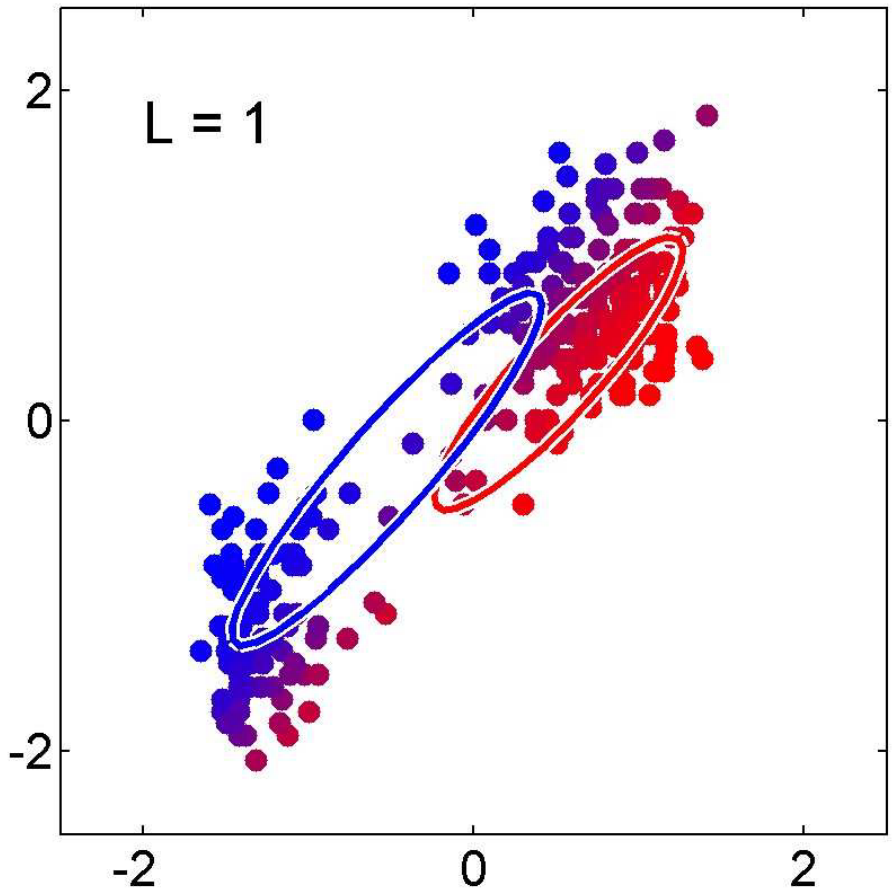
Gaussian mixture model



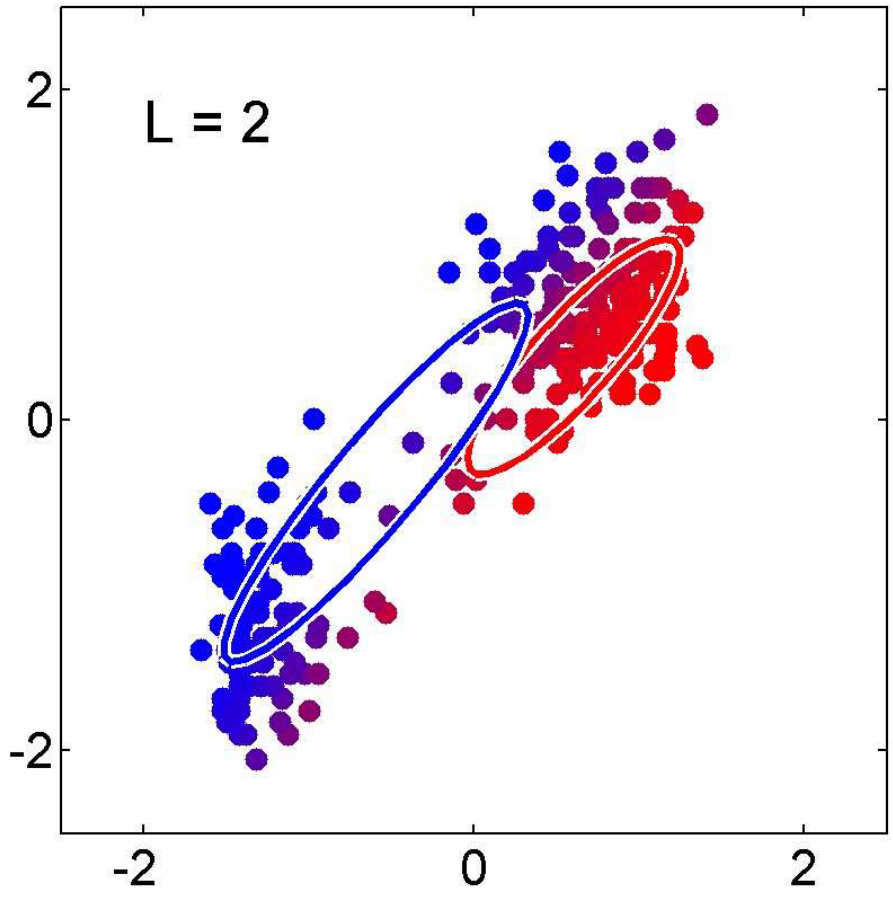
Gaussian mixture model



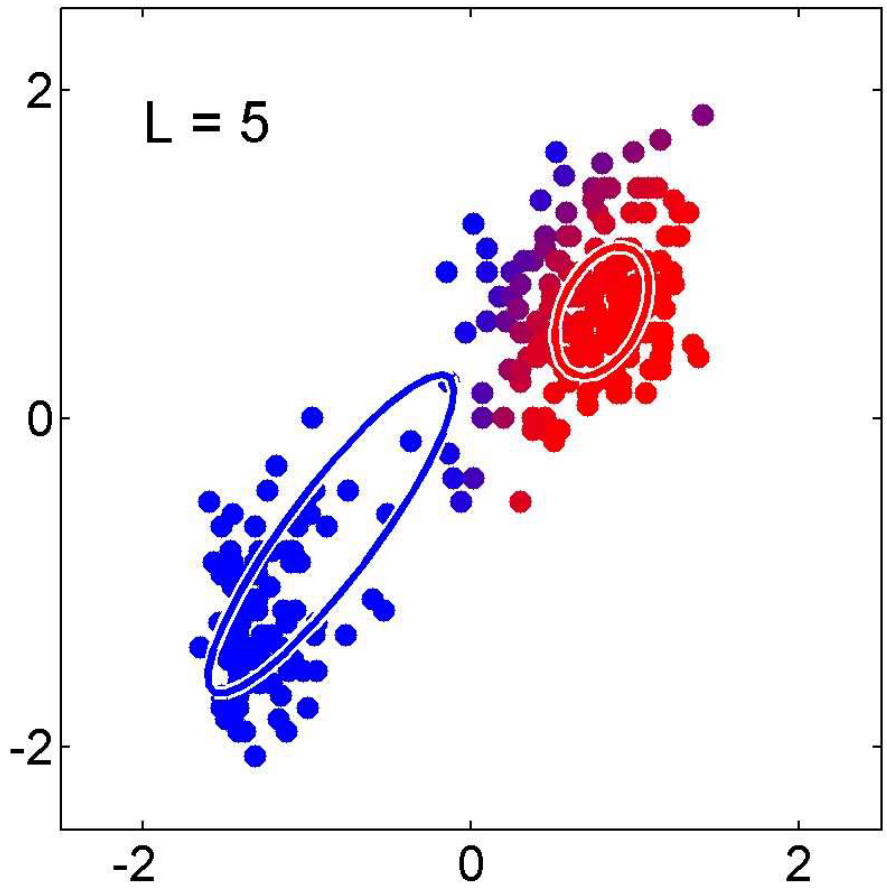
Gaussian mixture model



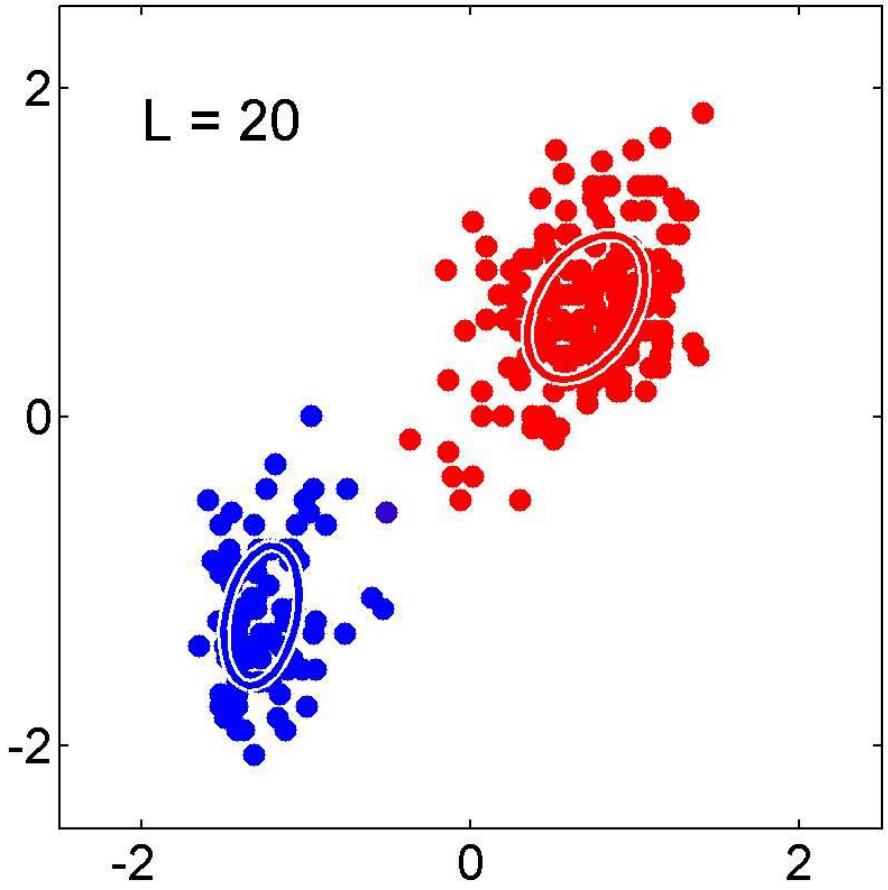
Gaussian mixture model



Gaussian mixture model



Gaussian mixture model



EM algorithm

- in general, EM algorithm is standard approach to maximum likelihood estimation with latent variable models
- data $\mathcal{D} = \{x_1, \ldots, x_N\}$
- want to fit model $p(x, z)$ with z hidden
- likelihood is given by

$$\ell(w) = \sum_{i=1}^{N} \log p(x; w) = \sum_{i=1}^{N} \log \sum_{z} p(x, z; w)$$

- often the case that maximum likelihood estimation of x would be easy if z were known, so alternate the two steps

EM algorithm

- iteratively lower bound ℓ, then maximize that lower bound
- for each i, let q_i be a distribution over z's

$$\begin{aligned}
\sum_{i=1}^{N} \log p(x_i) &= \sum_{i=1}^{N} \log \sum_{z_i} p(x_i, z_i) \\
&= \sum_{i=1}^{N} \log \sum_{z_i} q_i(z_i) \frac{p(x_i, z_i)}{q_i(z_i)} \\
&\geq \sum_{i=1}^{N} \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i)}{q_i(z_i)}
\end{aligned}$$

by Jensen's inequality

EM algorithm

- previous formula gives lower bound for *any* q_i; ideally, have the lower bound be tight (inequality holds with equality) for current value of w
- can show that this is the case when $q_i(z_i) = p(z_i \mid x_i; w)$
- E-step: lower bound ℓ via computing $p(z \mid x)$
- M-step: maximize this lower bound

$$\mathrm{E}_q[\log p(z, x; w)]$$

with respect to w

EM algorithm

- previous motivation is as a 'majorization-minimization' algorithm
- can also be viewed as coordinate ascent on

$$F(q, w) = \sum_{i=1}^{N} \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i; w)}{q_i(z_i)}$$

- E-step: maximization with respect to q
- M-step: maximization with respect to w
- (note: can also be modified for MAP estimation)
- this perspective suggests/justifies many variations

Outline

Inference

comes up in several places:

- using a trained model, *e.g.*, to predict out-of-sample outcomes
- E-step of EM for MLE in partially observed model
- Bayesian learning (work with joint model $p(x, z, w)$ with w random)

trivial in very simple cases, but expensive or intractable in complex models

(note: in some models, like hidden Markov models or Kalman filters, inference can be carried out exactly but requires use of an algorithm)

Variational inference

- let x be observed and z be hidden
- interested in computing the posterior distribution

$$p(z \mid x) = \frac{p(z, x)}{p(x)} = \frac{p(z, x)}{\int_z p(z, x)}$$

- denominator ('evidence') is hard to compute and makes this difficult
- main idea is to pick family $\mathcal{Q}$ of distributions over the latent variables indexed by *variational parameters*

$$q(z \mid \nu)$$

  and set ν *via optimization* to make q close to $p(z \mid x)$

- *i.e.*, turn a probability problem into an optimization problem

Variational inference

- recall following lower bound from EM

$$\log p(x) \geq L(q) = \mathrm{E}_q[\log p(x, z)] - \mathrm{E}_q[\log q(z)]$$

  where RHS is called the **evidence lower bound**

- here, choose a parametrized family of distributions for q such that these expectations are computable, then maximize lower bound L with respect to these 'variational parameters'

- can show that

$$\mathrm{KL}(q(z) \,\|\, p(z \,|\, x)) + L(q) = \log p(x)$$

- minimizing KL divergence is equivalent to maximizing L, plus obtain a lower bound on $\log p(x)$

Mean field family

- in mean field variational inference, assume that the family factorizes

$$q(z_1, \ldots, z_m) = \prod_{j=1}^{m} q(z_j)$$

  i.e., all variables are independent

- typically, this family does not contain the true posterior because the hidden variables are dependent (and these dependencies are what make the posterior difficult to work with)

- in 'coordinate ascent variational inference', iteratively optimize each variational distribution while holding others fixed

- computations end up being simple when relevant parts of original model are exponential family distributions

Additional topics

- **variational EM**
  - use variational inference to compute approximate posterior $p(z \mid x)$ in E-step, *i.e.*, do inexact maximization of F with respect to q
- **variational Bayes**
  - parameters w are random and model is $p(x, z, w)$
  - use lower bound

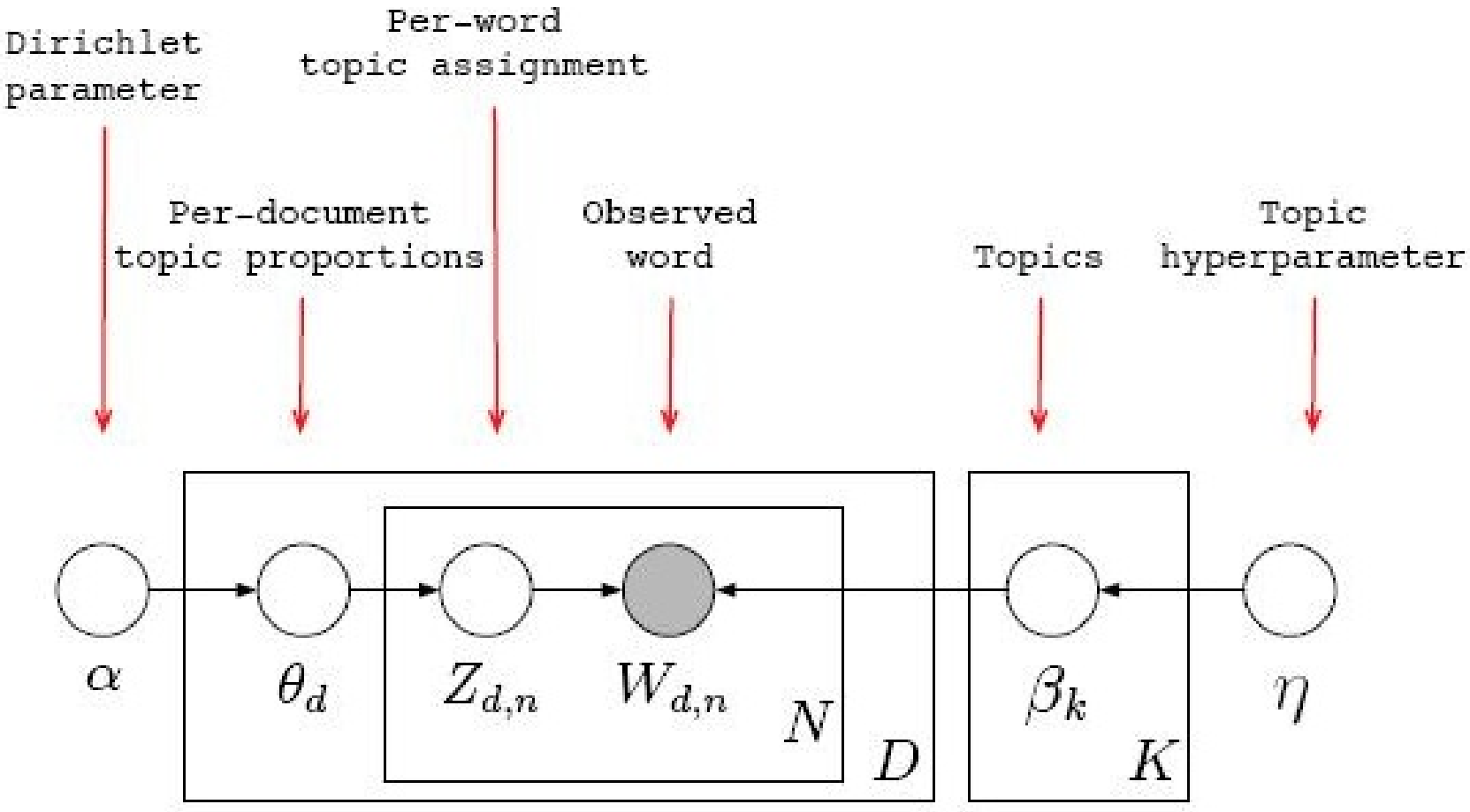
$$\log p(x) = \iint p(x, z, w)\, dz\, dw \geq \mathrm{E}[\log p(x, z, w)] - \mathrm{E}[\log q(z, w)]$$

    with factorized approximation $q(z, w) = q(z)q(w)$ and do alternating maximization w.r.t. z, w
  - yields EM-like algorithm sometimes called 'variational Bayesian EM'
- **stochastic variational inference**: use stochastic optimization to scale optimization carried out in variational inference

Latent Dirichlet Allocation

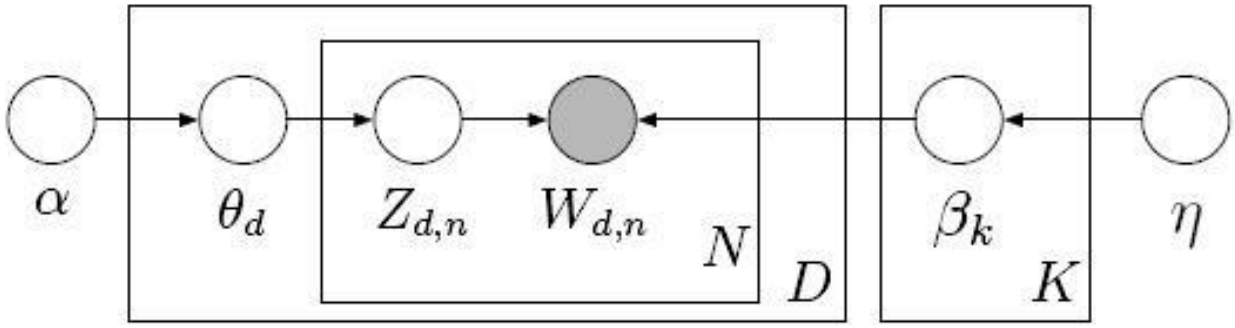
- **words** are multinomial random variables w
- **documents** are sequences of N words $\mathbf{w} = (w_1, \ldots, w_N)$
- **topics** are (multinomial) distributions over words
- model document as a random mixture θ over K latent topics
- from an *unlabeled* collection of documents, infer
  - per-word topic assignments in each document
  - per-document topic proportions
  - per-corpus topic distributions

Latent Dirichlet Allocation



Topic models

models for discovering thematic structure in document collections



joint distribution of topic mixture θ, topic distributions $\mathbf{z}$, and words $\mathbf{w}$ is:

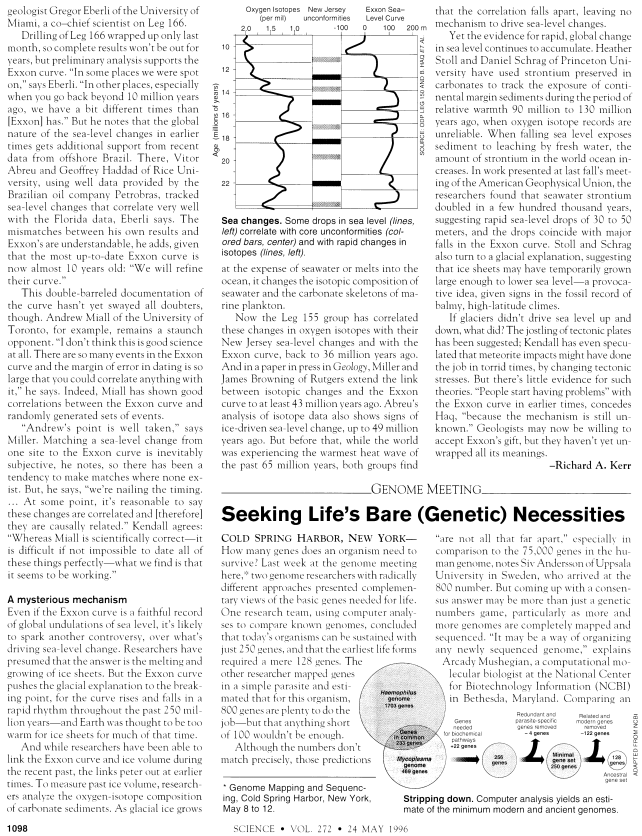
$$p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) = p(\theta \,|\, \alpha) \prod_{n=1}^{N} p(z_n \,|\, \theta) p(w_n \,|\, z_n, \beta)$$

goal: fit parameters and compute posterior $p(\theta, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta)$

A 100 topic model of Science 1980-2000

sound	quantum	brain	computer	ice
speech	laser	memory	data	climate
acoustic	light	human	information	ocean
language	optical	visual	problem	sea
sounds	electron	cognitive	computers	temperature
stars	research	materials	fossil	volcanic
universe	national	organic	species	years
galaxies	science	molecules	evolution	fig
astronomers	new	molecular	birds	deposits
star	funding	polymer	evolutionary	rocks

Topic proportions in documents



geologist Gregor Eberli of the University of Miami, a co-chief scientist on Leg 166.

Drilling of Leg 166 wrapped up only last month, so complete results won't be out for years, but preliminary analysis supports the Exxon curve. "In some places we were spot on," says Eberli. "In other places, especially when you go back beyond 10 million years ago, we have a bit different times than [Exxon] has." But he notes that the global nature of the sea-level changes in earlier times gets additional support from recent data from offshore Brazil. There, Vitor Abreu and Geoffrey Haddad of Rice University, using well data provided by the Brazilian oil company Petrobras, tracked sea-level changes that correlate very well with the Florida data, Eberli says. The mismatches between his own results and Exxon's are understandable, he adds, given that the most up-to-date Exxon curve is now almost 10 years old: "We will refine their curve."

This double-barreled documentation of the curve hasn't yet swayed all doubters, though. Andrew Miall of the University of Toronto, for example, remains a staunch opponent. "I don't think this is good science at all. There are so many events in the Exxon curve and the margin of error in dating is so large that you could correlate anything with it," he says. Indeed, Miall has shown good correlations between the Exxon curve and randomly generated sets of events.

"Andrew's point is well taken," says Miller. Matching a sea-level change from one site to the Exxon curve is inevitably subjective, he notes, so there has been a tendency to make matches where none exist. But, he says, "we're nailing the timing. ... At some point, it's reasonable to say these changes are correlated and [therefore] they are causally related." Kendall agrees: "Whereas Miall is scientifically correct—it is difficult if not impossible to date all of these things perfectly—what we find is that it seems to be working."

A mysterious mechanism

Even if the Exxon curve is a faithful record of global undulations of sea level, it's likely to spark another controversy, over what's driving sea-level change. Researchers have presumed that the answer is the melting and growing of ice sheets. But the Exxon curve pushes the glacial explanation to the breaking point, for the curve rises and falls in a rapid rhythm throughout the past 250 million years—and Earth was thought to be too warm for ice sheets for much of that time.

And while researchers have been able to link the Exxon curve and ice volume during the recent past, the links peter out at earlier times. To measure past ice volume, researchers analyze the oxygen-isotope composition of carbonate sediments. As glacial ice grows

Oxygen Isotopes (per mil) 2.0 1.5 1.0 | New Jersey unconformities -100 0 | Exxon Sea-Level Curve 100 200 m

Age (millions of years) 10 12 14 16 18 20 22

SOURCE: OGP LEG 150 AND B. HAQ ET AL.

Sea changes. Some drops in sea level *(lines, left)* correlate with core unconformities *(colored bars, center)* and with rapid changes in isotopes *(lines, left)*.

at the expense of seawater or melts into the ocean, it changes the isotopic composition of seawater and the carbonate skeletons of marine plankton.

Now the Leg 155 group has correlated these changes in oxygen isotopes with their New Jersey sea-level changes and with the Exxon curve, back to 36 million years ago. And in a paper in press in *Geology*, Miller and James Browning of Rutgers extend the link between isotopic changes and the Exxon curve to at least 43 million years ago. Abreu's analysis of isotope data also shows signs of ice-driven sea-level change, up to 49 million years ago. But before that, while the world was experiencing the warmest heat wave of the past 65 million years, both groups find that the correlation falls apart, leaving no mechanism to drive sea-level changes.

Yet the evidence for rapid, global change in sea level continues to accumulate. Heather Stoll and Daniel Schrag of Princeton University have used strontium preserved in carbonates to track the exposure of continental margin sediments during the period of relative warmth 90 million to 130 million years ago, when oxygen isotope records are unreliable. When falling sea level exposes sediment to leaching by fresh water, the amount of strontium in the world ocean increases. In work presented at last fall's meeting of the American Geophysical Union, the researchers found that seawater strontium doubled in a few hundred thousand years, suggesting rapid sea-level drops of 30 to 50 meters, and the drops coincide with major falls in the Exxon curve. Stoll and Schrag also turn to a glacial explanation, suggesting that ice sheets may have temporarily grown large enough to lower sea level—a provocative idea, given signs in the fossil record of balmy, high-latitude climes.

If glaciers didn't drive sea level up and down, what did? The jostling of tectonic plates has been suggested; Kendall has even speculated that meteorite impacts might have done the job in torrid times, by changing tectonic stresses. But there's little evidence for such theories. "People start having problems" with the Exxon curve in earlier times, concedes Haq, "because the mechanism is still unknown." Geologists may now be willing to accept Exxon's gift, but they haven't yet unwrapped all its meanings.

–Richard A. Kerr

GENOME MEETING

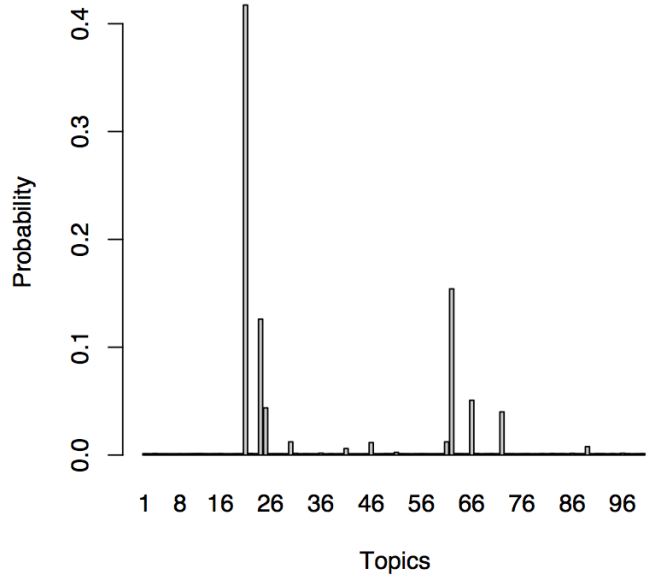
Seeking Life's Bare (Genetic) Necessities

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.

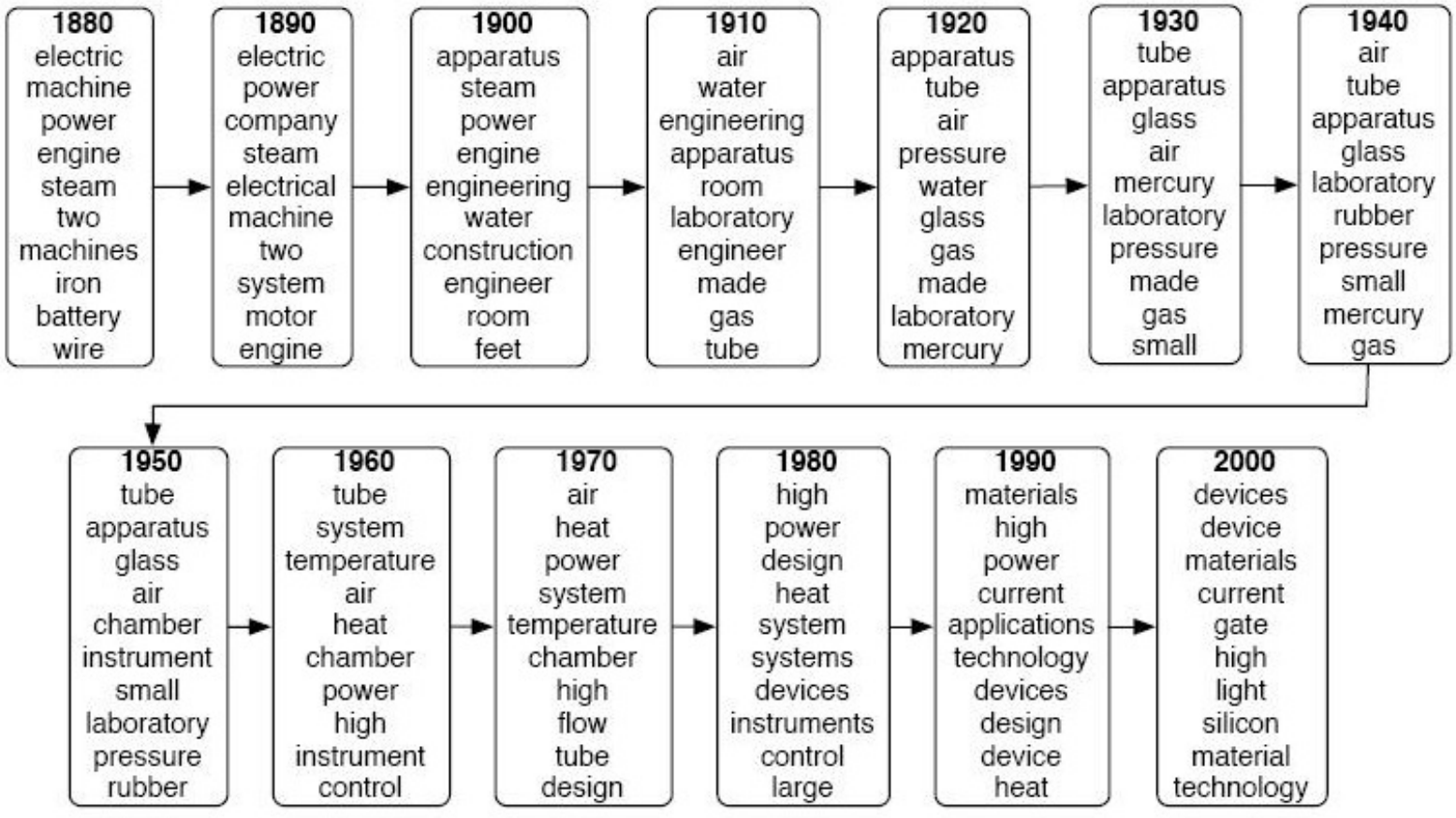
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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.

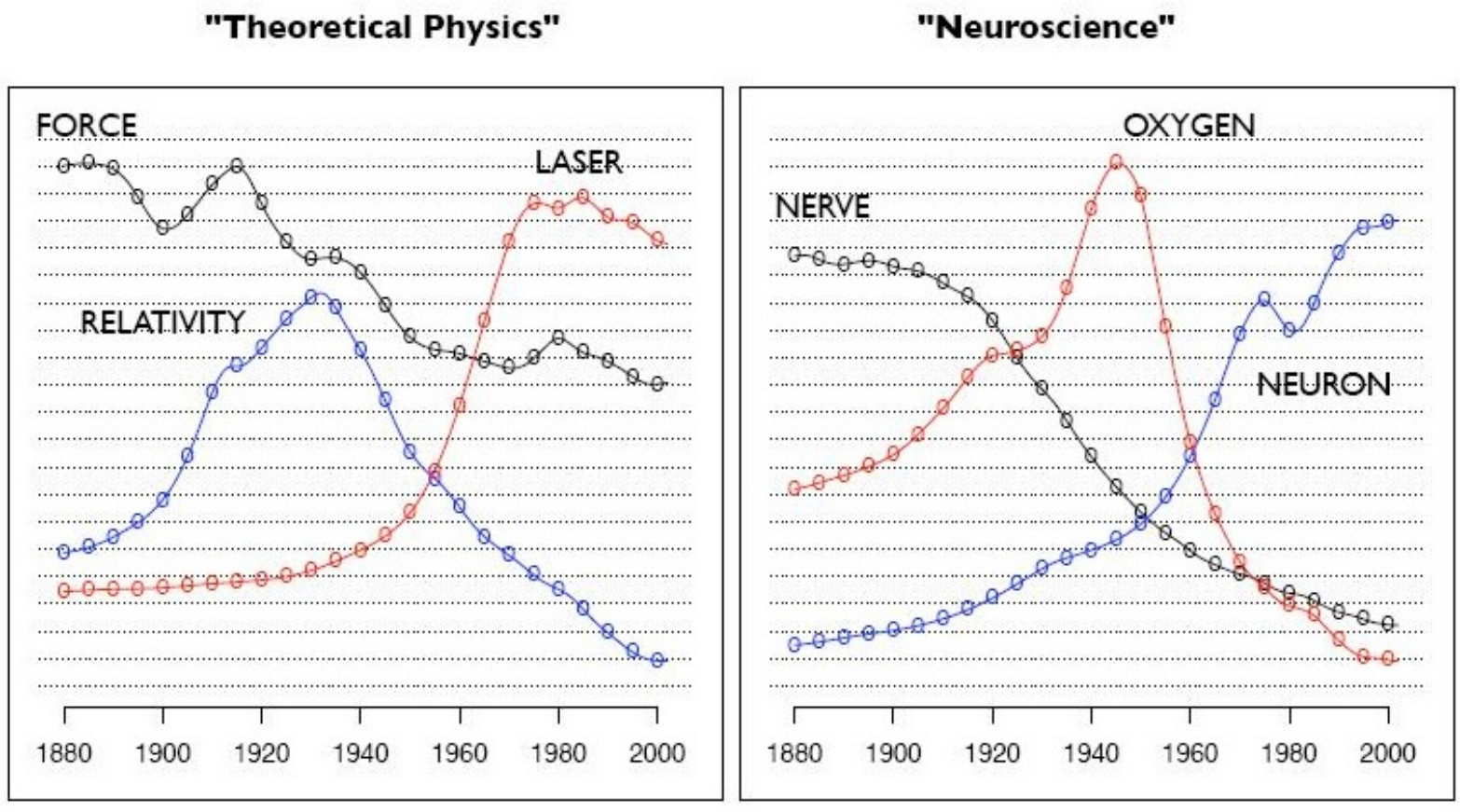
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



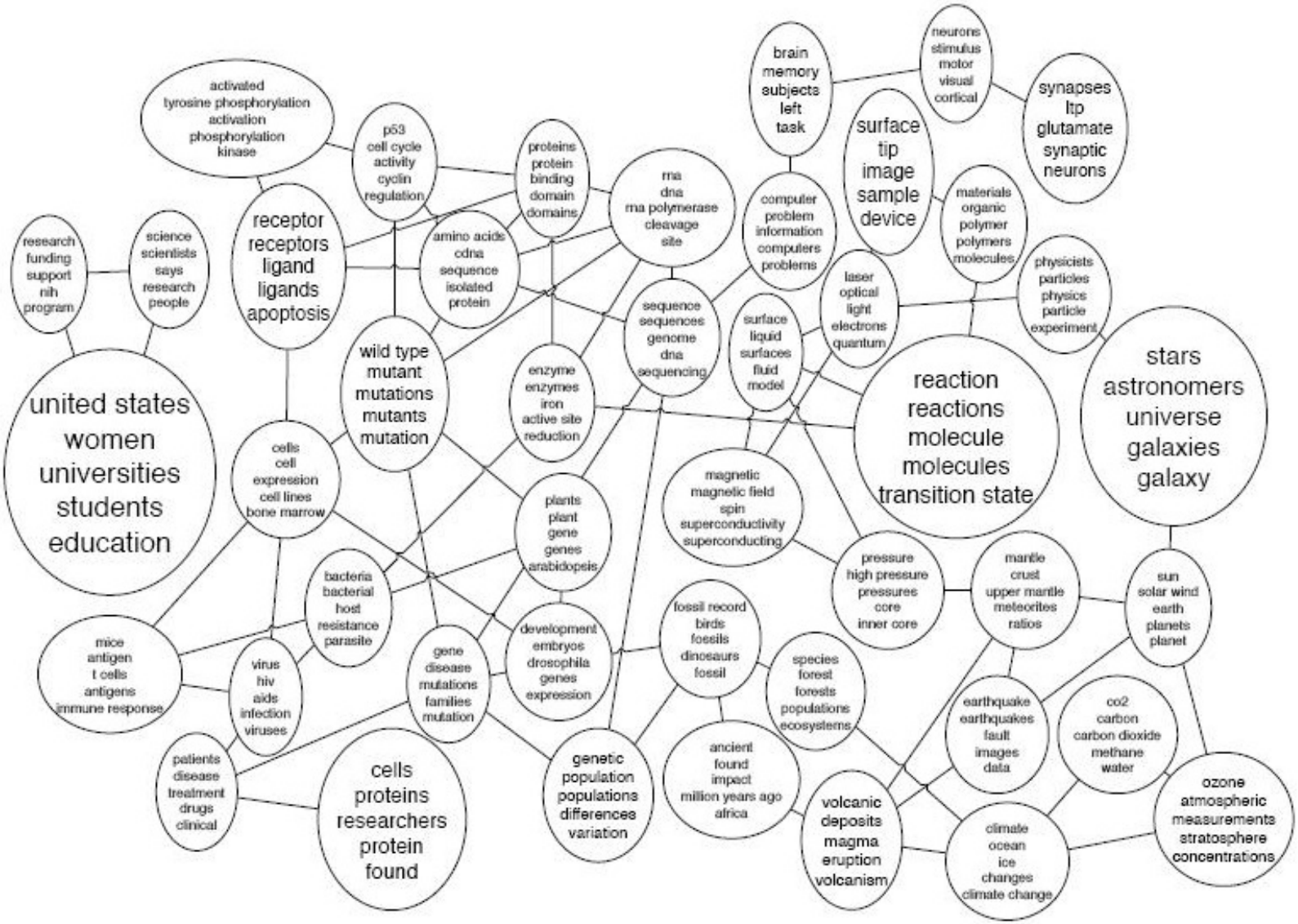
Model Evolution of Topics over Time



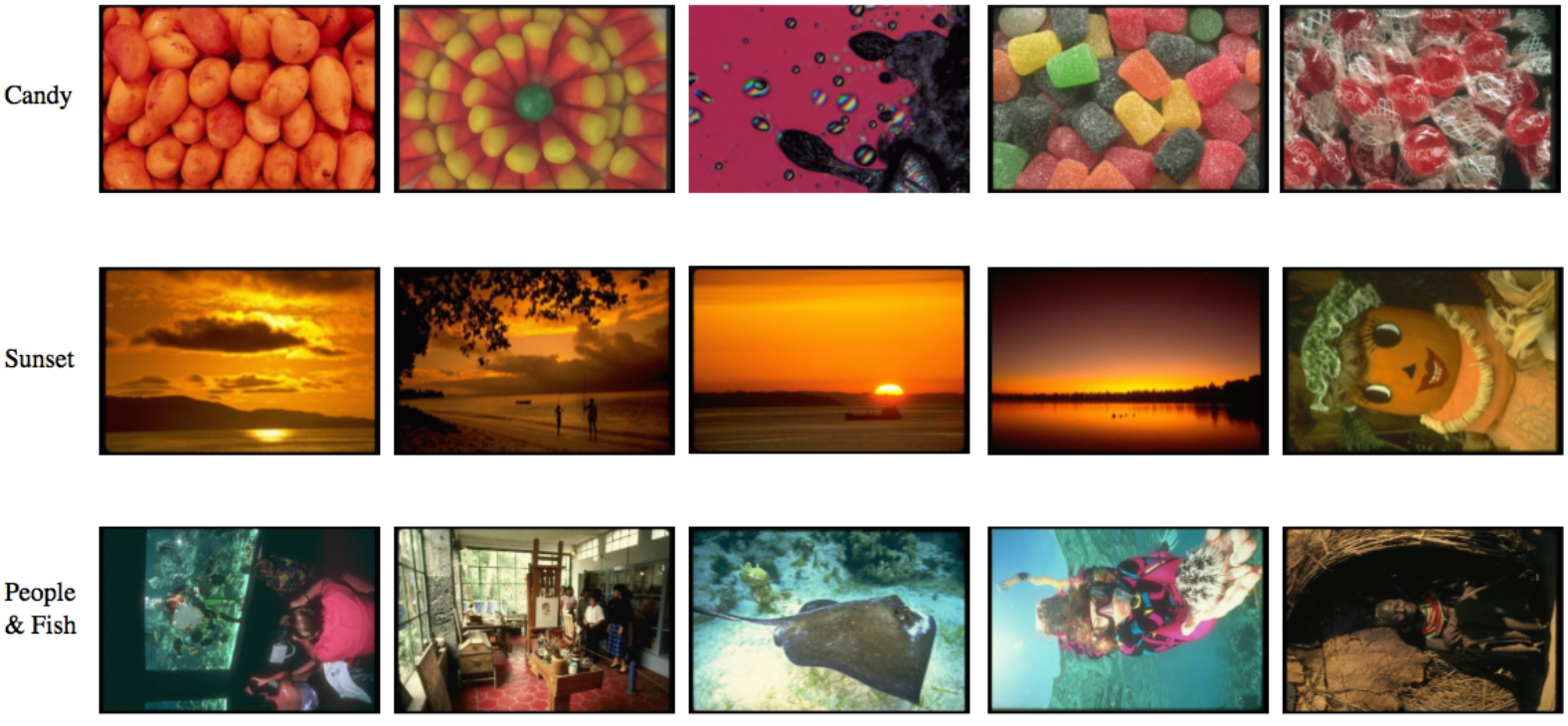
Visualizing Trends Within Topics



Model Connections Between Topics



Matching Words and Pictures



Matching Words and Pictures

True caption
market people
Corr–LDA
people market pattern textile display

True caption
scotland water
Corr–LDA
scotland water flowers hills tree

True caption
bridge sky water
Corr–LDA
sky water buildings people mountain

True caption
sky tree water
Corr–LDA
tree water sky people buildings

True caption
birds tree
Corr–LDA
birds nest leaves branch tree

True caption
fish reefs water
Corr–LDA
fish water ocean tree coral

True caption
mountain sky tree water
Corr–LDA
sky water tree mountain people

True caption
clouds jet plane
Corr–LDA
sky plane jet mountain clouds

Variational EM for LDA

- computing evidence $p(\mathbf{w} \mid \alpha, \beta)$ is intractable, but evidence lower bound

$$\mathrm{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - \mathrm{E}_q[\log q(\theta, \mathbf{z} \mid \gamma, \phi)]$$

  gives lower bound on $\log p(\mathbf{w} \mid \alpha, \beta)$

- plug in form of p and use family of approximate posteriors given by

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$

  where γ is a variational Dirichlet parameter and ϕ_n are variational multinomial parameters

- variational E-step: maximize lower bound with respect to γ and ϕ_n via alternating maximization (both simple closed form expressions)

- M-step: maximize with respect to hyperparameters α (simple numerical method), β (closed form)

Outline

Conclusions

- interactions of optimization and probability in machine learning
- using probabilistic structure can ease optimization
- framing probabilistic computations in variational form can help bring full optimization toolbox to bear on wider range of problems
- can lead to fast and scalable algorithms that enable working with very complex probabilistic models on huge datasets, beyond the reach of other methods
- many analogies between probabilistic methods (*e.g.*, Gibbs sampling) and optimization-based methods (*e.g.*, coordinate ascent)

Acknowledgements

- David Blei & collaborators
- Michael Jordan & collaborators
- Andrew Ng (diagrams)