Going beyond map–reduce and going beyond maximum-likelihood

David W. Hogg

Center for Cosmology and Particle Physics, New York University

2012 September 11

punchlines

- The map–reduce framework (or something like it) does important tasks in $\log N$ time; it is the "only" framework for big data operations at the present day.
  - good news: We can do maximum-likelihood problems in map–reduce!
- bad news: The next generation of astronomy projects must go beyond maximum-likelihood methods to deliver.
  - *Gaia*, *LSST*, *Euclid*, *etc.*
- We don't know how to do this "at scale".
  - call to arms
  - (and get rich too!)

principal collaborators

- Jo Bovy (IAS)
- Brendon Brewer (Auckland)
- Rob Fergus (NYU)
- Dan Foreman-Mackey (NYU)
- Jonathan Goodman (NYU)
- Dustin Lang (CMU)

map–reduce or die

- *"We won't even consider any algorithms that can't be written in the map–reduce framework."*
- map:
  - at each "data point" (on the distributed system), do an operation on that datum, produce output
  - think: *Search document for "kittens"; return DocumentID and PageRank if it hits.*
  - *distributed data* is the key: Every datum is near a CPU.
- reduce:
  - between each pair of outputs, do an operation and return one new output, recurse up the tree
  - think: *Compare two PageRanks and return DocumentID and PageRank of the better.*
  - tree structure of the data center is the key: There are only $\log_2 N$ branches to any datum.
- Brilliant. And a huge opportunity.

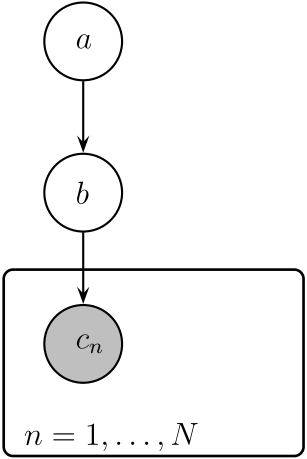
maximum-likelihood and map–reduce

- full-data likelihood: $p(D \mid \theta) = \prod_n p(d_n \mid \theta)$
- Find the *maximum with respect to* θ of this likelihood.
- map:
  - compute $\dfrac{\mathrm{d} \ln p(d_n \mid \theta)}{\mathrm{d}\theta}$
- reduce:
  - pairwise sum
- Go uphill. Repeat as necessary; each iteration only takes $\log N$ time.
  - (use L-BFGS or whatever you like)

astronomical scale

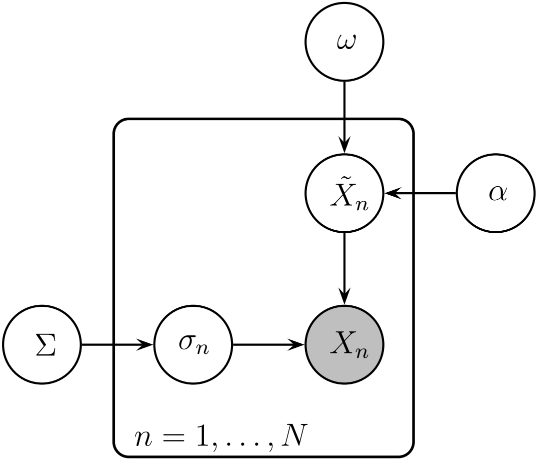
- *LSST*: 10^{10} galaxies in 10^{15} pixels
  - get the cosmic shear map
  - and then the cosmological parameters
- *Gaia*: 10^{9} stars in 10^{12} pixels
  - infer the dynamics of the Milky Way
  - but also—necessarily—the distribution function of stars in that potential
- *non-parametric* shear map or distribution function
  - "non-parametric" means the model *gets bigger as the data set gets bigger* (or better)
  - think: *As you observe more and more galaxies, with better redshift estimates, you increase the angular and redshift resolution of your shear map.*
  - importantly, non-parametric models are *never* inferred at high signal-to-noise

probabilistic graphical models

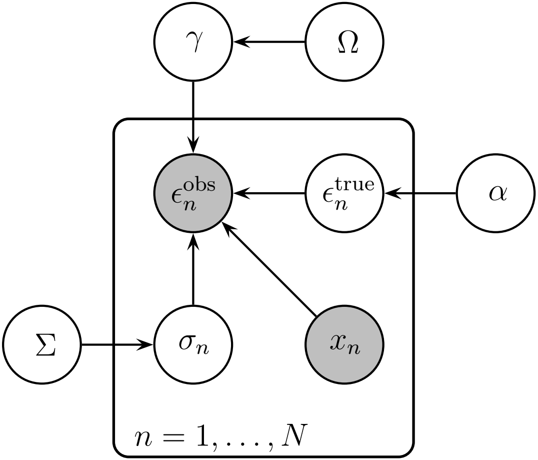


$$p(a, b, \{c_n\}) = p(a)\, p(b \mid a) \prod_{n=1}^{N} p(c_n \mid b)$$

astrophysics problems are hierarchical



astrophysics problems are hierarchical



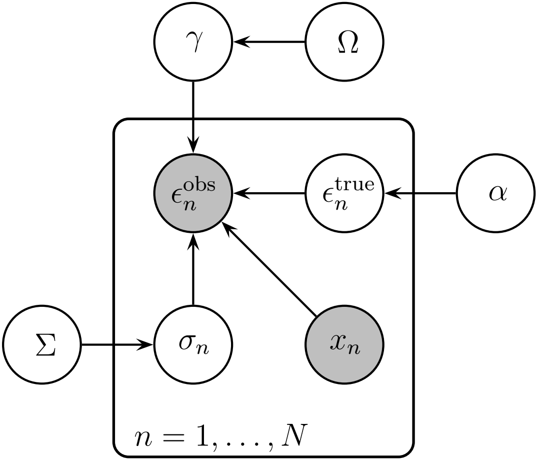
there are no linear problems

- Even if your noise is Gaussian, you never know the noise variance at high precision.
- In most real situations, the data are produced by a *mixture of processes*.
- There are always multiple modes to the likelihood function, and broad support in parameter space.

Bayesian inference *isn't* map–reduce

- $p(\theta \mid D) = \frac{1}{Z} p(\theta) \prod_n p(d_n \mid \theta)$
- map:
  - compute functions $p(d_n \mid \theta)$
- reduce:
  - product functions together (starting with the prior)
- but think about how you pass forward those functions
  - θ has 10^6 or more parameters
  - functions are multi-modal
  - support is broader than Gaussian
  - and non-parametrics are deadly
- But that's not all...

marginalization is hard—and unavoidable



Bayesian state-of-the-art

- there *are* huge non-parametric Bayesian inferences with massive marginalizations out there
- How were they done?
  - carefully chosen priors that make the inferences and marginalizations analytic or tractable
  - *we can't do this*
  - why not? Because for us the priors *actually are* our prior beliefs. Our prior beliefs are not conjugate to anything!
- "Bayesian" is becoming a bad word

my approach

- brute force
  - (plus some help from applied math and computer vision)

My day job

- Lang & Hogg (forthcoming): a 10^9-parameter model of the 10^{13} *SDSS* pixels (*The Tractor*)
- Brewer *et al.* (forthcoming): Bayesian non-parametrics but with priors that represent our actual prior knowledge
- Foreman-Mackey *et al.* (arXiv:1202.3665): *emcee*, the MCMC Hammer: flexible, parallelized, adaptive sampler
- Bovy, Murray, Hogg (arXiv:0903.5308): a dynamical inference fully marginalizing out a non-parametric distribution function
- Bovy *et al.* (arXiv:1105.3975): a 60,000-parameter model of 700,000 flux measurements, followed by predictions for 160,000,000 point sources
- Bovy, Hogg, Roweis (arXiv:0905.2979): *extreme deconvolution*: hierarchical inference in the presence of missing data and heterogeneous noise

punchlines

- The map–reduce framework (or something like it) does important tasks in $\log N$ time; it is the "only" framework for big data operations at the present day.
  - good news: We can do maximum-likelihood problems in map–reduce!
- bad news: The next generation of astronomy projects must go beyond maximum-likelihood methods to deliver.
  - *Gaia*, *LSST*, *Euclid*, *etc.*
- We don't know how to do this "at scale".
  - call to arms
  - (and get rich too!)