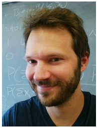
Learning an astronomical catalog of the visible universe through scalable Bayesian inference

Jeffrey Regier
with the DESI Collaboration

Department of Electrical Engineering and Computer Science
UC Berkeley

December 9, 2016

The DESI Collaboration



Jeff Regier

Ryan Giordano

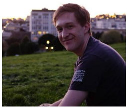
Steve Howard

Jon McAuliffe

Michael Jordan

Andy Miller

Ryan Adams



Jarrett Revels

Andreas Jensen

Kiran Pamnany

Debbie Bard

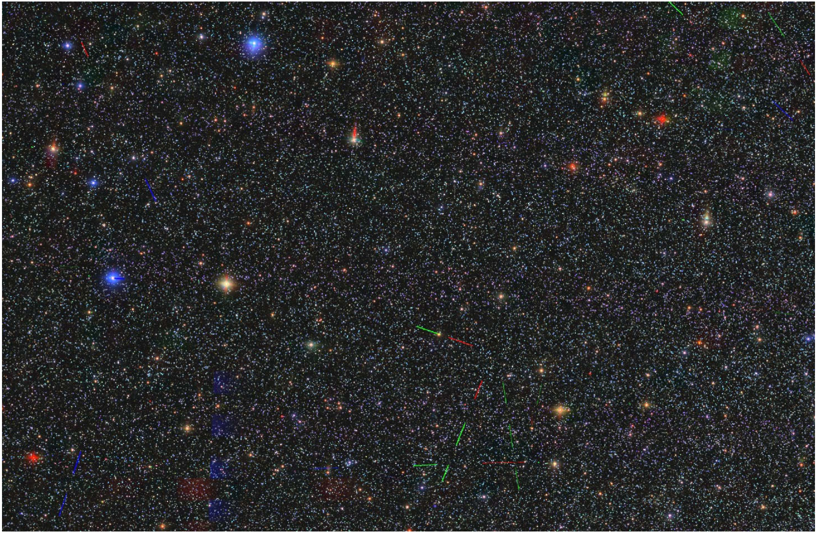
Rollin Thomas

Dustin Lang

David Schlegel

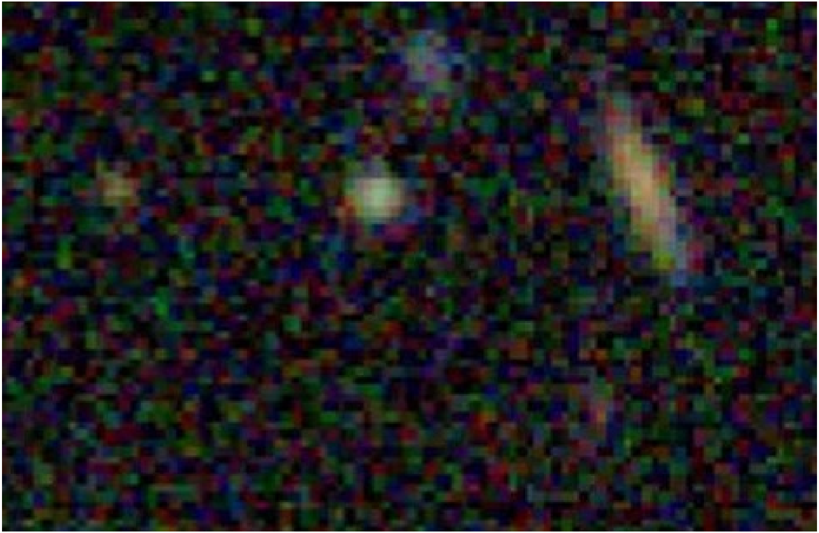
Prabhat

An astronomical image



An image from the Sloan Digital Sky Survey covering roughly one quarter square degree of the sky.

Faint light sources

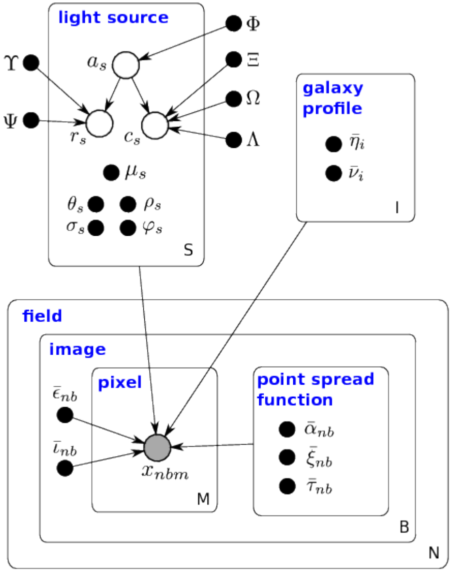


Most light sources are near the detection limit.

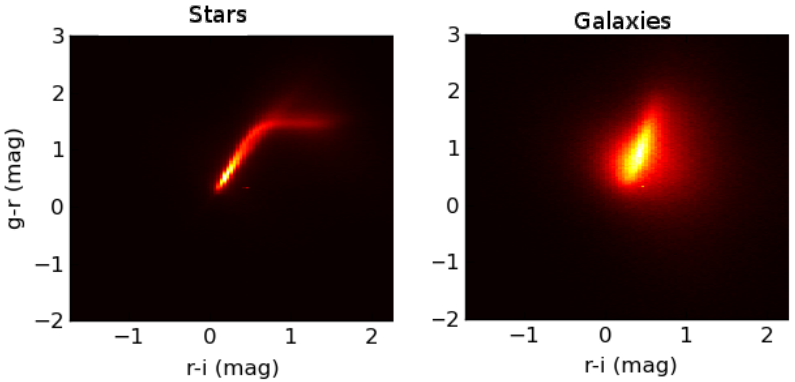
Outline

1. our graphical model for astronomical images (Celeste)
2. scaling approximate posterior inference to catalog the visible universe
3. model extensions

The Celeste graphical model



Scientific color priors



Galaxies: light-density model

The light density for galaxy s is modeled as mixture of two extremal galaxy prototypes:

$$h_s(w) = \theta_s h_{s1}(w) + (1 - \theta_s) h_{s0}(w).$$

Each prototype ($i = 0$ or $i = 1$) is a mixture of bivariate normal distributions:

$$h_{si}(w) = \sum_{j=1}^{J} \bar{\eta}_{ij} \phi(w; \mu_s, \bar{\nu}_{ij} Q_s).$$

Shared covariance matrix Q_s accounts for the scale σ_s, rotation φ_s, and axis ratio ρ_s.

An elliptical galaxy, $\theta_s = 0$

A spiral galaxy, $\theta_s = 1$

Idealized sky view

The brightness for sky position w is

$$G_b(w) = \sum_{s=1}^{S} \ell_{sb} g_s(w)$$

where

$$g_s\left(w\right) = \begin{cases} \mathbf{1}\left\{\mu_s = w\right\}, \text{ if } a_s = 0 \text{ (``star'')} \\ h_s(w), \text{ if } a_s = 1 \text{ (``galaxy'')}. \end{cases}$$

Astronomical images

Images differ from the idealized sky view due to

1. pixelation and point spread

$$f_{nbm}(w) = \sum_{k=1}^{K} \bar{\alpha}_{nbk} \phi\left(w_m; w + \bar{\xi}_{nbk}, \bar{\tau}_{nbk}\right)$$

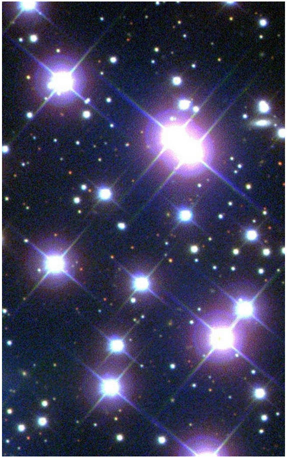
$$G_{nbm} = G_b * f_{nbm}$$

2. background radiation and calibration

$$F_{nbm} = \iota_{nb}\left[\epsilon_{nb} + G_{nbm}\right]$$

3. finite exposure duration

$$x_{nbm} | \left(a_s, r_s, c_s\right)_{s=1}^{S} \sim \mathrm{Poisson}\left(F_{nbm}\right)$$



Intractable posterior

Let $\Theta = (a_s, r_s, c_s)_{s=1}^{S}$. The posterior on Θ is intractable because of coupling between the sources:

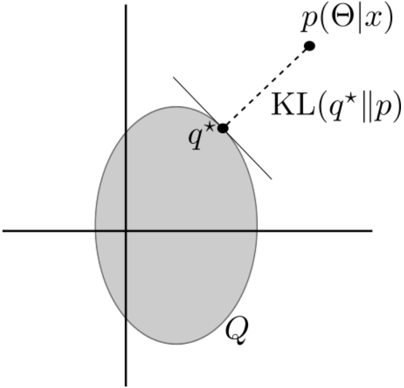
$$p(\Theta|x) = \frac{p(x|\Theta)p(\Theta)}{p(x)}$$

and

$$\begin{aligned} p(x) &= \int p(x|\Theta)p(\Theta)\, d\Theta \\ &= \int \prod_{n=1}^{N} \prod_{b=1}^{B} \prod_{m=1}^{M} p(x_{nbm}|\Theta)p(\Theta)\, d\Theta. \end{aligned}$$

Variational inference

Variational inference approximates the exact posterior p with a simpler distribution $q^\star \in Q$.



Variational inference for Celeste

- An approximating distribution that factorizes across light sources (a "structured mean-field" assumption) makes most expectations tractable:

$$q(\Theta) = \prod_{s=1}^{S} q(\Theta_s).$$

- The delta method for moments approximates the remaining expectations.
- Existing catalogs provide good initial settings for the variational parameters.
- Light sources are unlikely to contribute photons to distant pixels.
- The model contains an auxiliary variable indicating the mixture component that generated each source's colors.
- Newton's method converges in tens of iterations.

Validation from Stripe 82

	Photo	Celeste
position	0.37	**0.24**
missed gals	23 / 421	**8 / 421**
missed stars	**10 / 421**	38 / 421
color u-g	1.25	**0.70**
color g-r	0.37	**0.21**
color r-i	0.25	**0.17**
color i-z	0.31	**0.15**
brightness	**0.20**	0.37
profile	0.26	0.31
axis ratio	0.19	**0.13**
scale	1.64	1.76
angle	17.04	**12.64**

Average error. Lower is better. Highlight scores are more than 2 standard deviations better.

Scaling inference to the visibile universe

The setting

Big data

- Sloan Digital Sky Survey: 55 TB of images; hundreds of millions of stars and galaxies
- Large Synoptic Survey (2019): 15 TB of images nightly

Fancy hardware

- Cori supercomputer, Phase 1: 1,630 nodes, each with 32 cores.
- Cori supercomputer, Phase 2: 9,300 nodes, each with 272 hardware threads. 30 teraflops.
- 1.5 PB array of SSDs

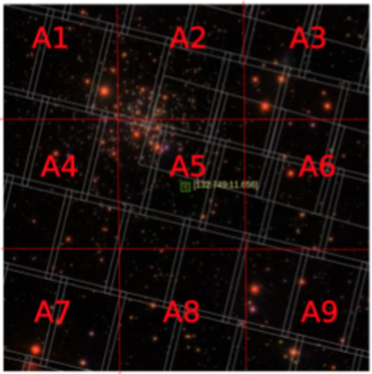
Julia programming language

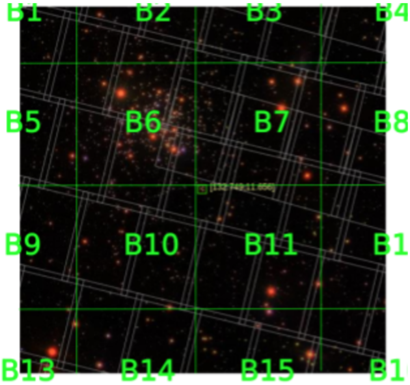
- high-level syntax
- as fast as C++ (when necessary)
- a single language for "hotspots" and the rest
- multi-threading (experimental), not just multi-processing

Fast serial optimization algorithm

- analytic expectations (and one in delta method for moments)
- block coordinate ascent
- Newton steps rather than L-BFGS or a first-order method
- manually coded gradients and Hessians
- 3x overhead from computing exact Hessians

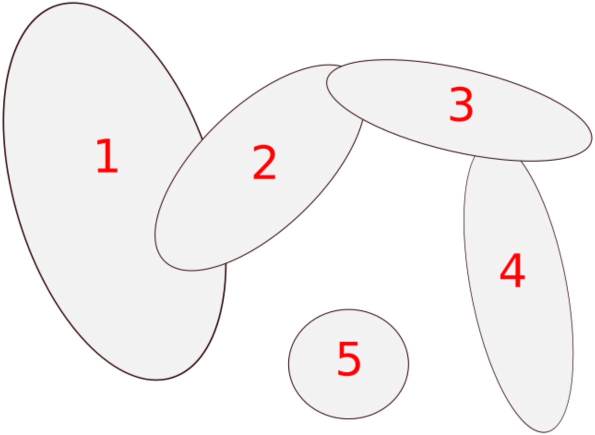
Parallelism among nodes





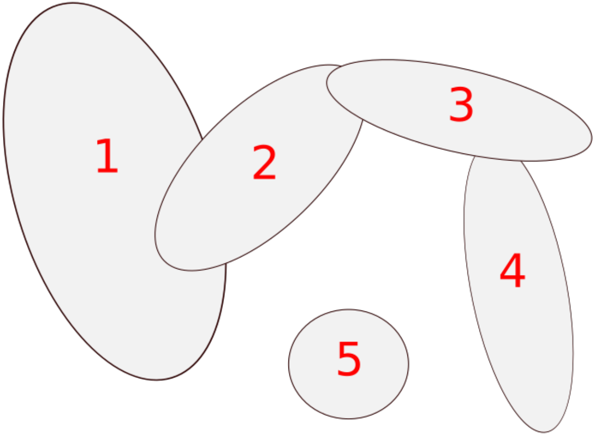
A region of the sky, shown twice, divided into 25 overlapping boxes: A1,...,A9 and B1,...,B16. Each box corresponds to a task: to optimize all the light sources within its boundaries.

Parallelism among threads



Light sources that do not overlap may be updated concurrently.

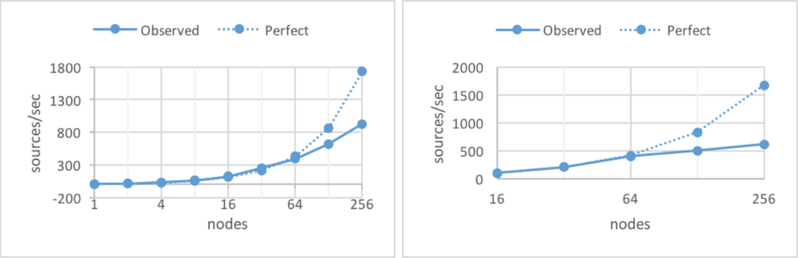
Parallelism among threads



Light sources that do not overlap may be updated concurrently.

[1] Pan, Xinghao, et al. "CYCLADES: Conflict-free Asynchronous Machine Learning." NIPS 2016.

Weak and strong scaling



(a) Weak scaling (b) Strong scaling

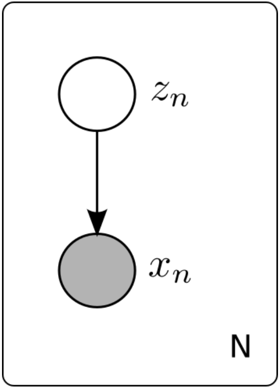
Celeste light sources/second. We observe perfect scaling up to 64 nodes. The we are limited by interconnect bandwidth.

Hero run: catalog of the Sloan Digital Sky Survey

- 512 nodes $\times$ 32 cores/node $=$ 16,384 cores
- 16,384 cores $\times$ 16 hours $=$ 250,000 core hours
- input: 55 TB of astronomical images
- output: catalog of 250 million light sources

A deep generative model for galaxies

A deep generative model for galaxies

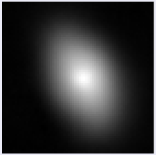


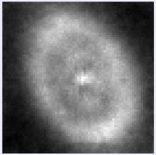
$$z_n \sim \mathcal{N}(0, \mathrm{I})$$

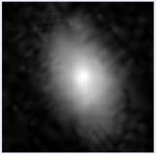
$$x_n | z_n \sim \mathcal{N}(f_\mu(z), f_\sigma(z))$$

Example

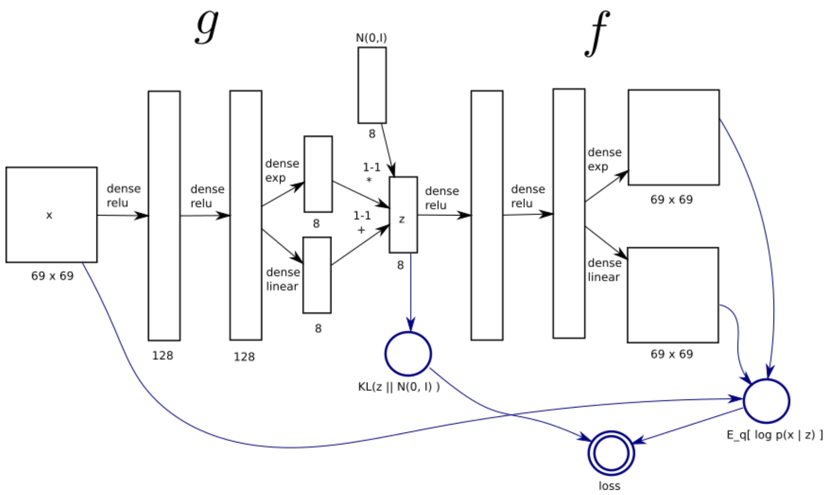
$$z_n = [0.1, -0.5, 0.2, 0.1]^\intercal$$

$f_\mu(z_n) =$ 

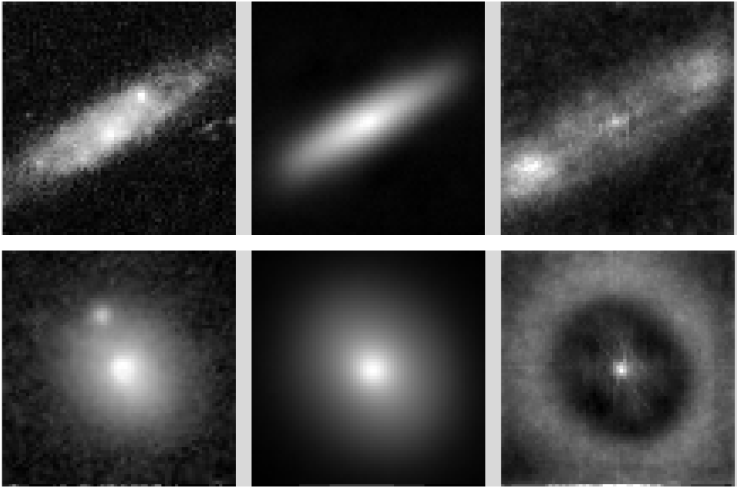
$f_\sigma(z_n) =$ 

$x_n =$ 

Autoencoder architecture



Sample fits



Each row corresponds to a different example from a test set. The left column shows the input x. The center column shows the output $f_\mu(z)$ for a z sampled from $\mathcal{N}(g_\mu(x), g_\sigma(x))$. The right column shows the output $f_\sigma(z)$ for the same z.

Second-order stochastic variational inference

Second-order Stochastic Variational Inference

Require: ω is the initial vector of variational parameters; $\delta \in (\delta_{\min}, \delta_{\max})$ is the initial trust-region radius; $\gamma > 1$ is the trust region expansion factor; and $\eta_1 \in (0, 1)$ and $\eta_2 > 0$ are constants.

for $i \leftarrow 1$ to M **do**
    Sample $e_1, \ldots, e_N$ iid from base distribution ϵ.
    $g \leftarrow \nabla_\nu \hat{\mathcal{L}}(\nu; e_1, \ldots, e_N)|_\omega$
    $H \leftarrow \nabla^2_\nu \hat{\mathcal{L}}(\nu; e_1, \ldots, e_N)|_\omega$
    $\omega' \leftarrow \arg\max_\nu \{g^\intercal \nu + \nu^\intercal H \nu : \|\nu\| \leq \delta\}$ ▷ non-convex quadratic optimization
    $\beta \leftarrow g^\intercal \omega' + \omega'^\intercal H \omega'$ ▷ the expected improvement
    Sample $e'_1, \ldots, e'_N$ iid from base distribution ϵ.
    $\alpha \leftarrow \hat{\mathcal{L}}(\omega'; e'_1, \ldots, e'_N) - \hat{\mathcal{L}}(\omega; e'_1, \ldots, e'_N)$ ▷ the observed improvement
    if $\alpha/\beta > \eta_1$ and $\|g\| \geq \eta_2 \delta$ **then**
        $\omega \leftarrow \omega'$
        $\delta \leftarrow \max(\gamma\delta, \delta_{max})$
    else
        $\delta \leftarrow \delta/\gamma$
    if $\delta < \delta_{\min}$ or $i = M$ **then return** ω

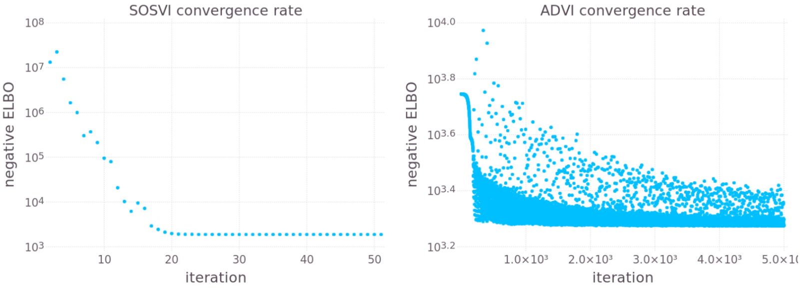
Second-order Stochastic Variational Inference

Require: ω is the initial vector of variational parameters; $\delta \in (\delta_{\min}, \delta_{\max})$ is the initial trust-region radius; $\gamma > 1$ is the trust region expansion factor; and $\eta_1 \in (0,1)$ and $\eta_2 > 0$ are constants.

for $i \leftarrow 1$ to M **do**
    Sample $e_1, \ldots, e_N$ iid from base distribution ϵ.
    $g \leftarrow \nabla_\nu \hat{\mathcal{L}}(\nu; e_1, \ldots, e_N)|_\omega$
    $H \leftarrow \nabla^2_\nu \hat{\mathcal{L}}(\nu; e_1, \ldots, e_N)|_\omega$
    $\omega' \leftarrow \arg\max_\nu \{g^\intercal \nu + \nu^\intercal H \nu : \|\nu\| \le \delta\}$ ▷ non-convex quadratic optimization
    $\beta \leftarrow g^\intercal \omega' + \omega'^\intercal H \omega'$ ▷ the expected improvement
    Sample $e'_1, \ldots, e'_N$ iid from base distribution ϵ.
    $\alpha \leftarrow \hat{\mathcal{L}}(\omega'; e'_1, \ldots, e'_N) - \hat{\mathcal{L}}(\omega; e'_1, \ldots, e'_N)$ ▷ the observed improvement
    if $\alpha/\beta > \eta_1$ and $\|g\| \ge \eta_2 \delta$ **then**
        $\omega \leftarrow \omega'$
        $\delta \leftarrow \max(\gamma\delta, \delta_{max})$
    else
        $\delta \leftarrow \delta/\gamma$
    if $\delta < \delta_{\min}$ or $i = M$ **then return** ω

68X fewer iterations than ADVI/SGD.

Case study: empirical convergence rates



Open questions

- How do we more accurately approximate the posterior distribution?
  - normalizing flows without an encoder network
  - hybrid VI/MCMC
  - linear response variational Bayes (LRVB)
- How do we model a spatially varying point spread function (PSF)?
  - Variational autoencoders fit independent PSFs well.
  - But it isn't easy to account for dependence among the PSFs of nearby astronomical objects.
- How can we easily account for the details we don't yet account for?
  - cosmic rays, airplanes, and satelights
  - camera saturation (censoring) and imaging artifacts
  - terrestrial vs extraterestrial background radiation
  - transient events: supernovae, exoplanets, and near-Earth asteroids
  - additional imaging datasets, spectrographic datasets, calibration datasets

Thank you!