Bayesian Methods in Cryo-EM

Marcus A. Brubaker
York University / Structura Biotechnology
Toronto, Canada

LASSONDE SCHOOL OF ENGINEERING | YORK UNIVERSITÉ UNIVERSITY

STRUCTURA BIOTECHNOLOGY INC.

Bayesian Methods in Cryo-EM

Bayesian methods already underpin many successful techniques

- Likelihood methods for refinement/3D classification
- 2D classification

May provide a framework to answer some outstanding problems

- Flexibility
- Validation
- CTF estimation
- Others?

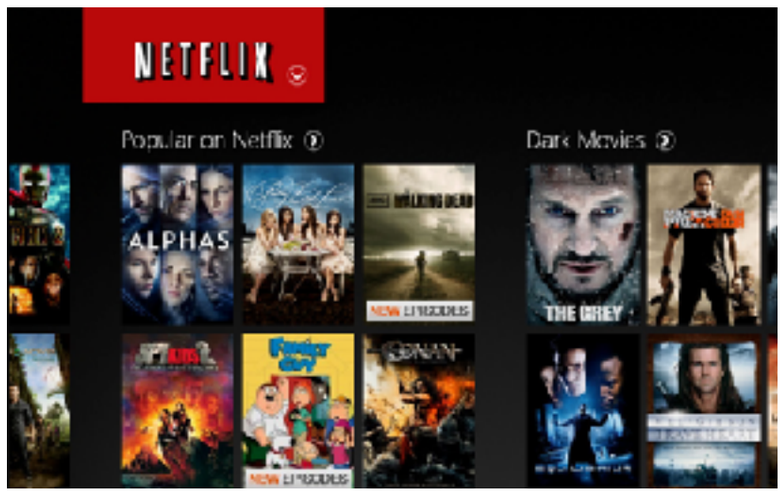
What are Bayesian Methods?

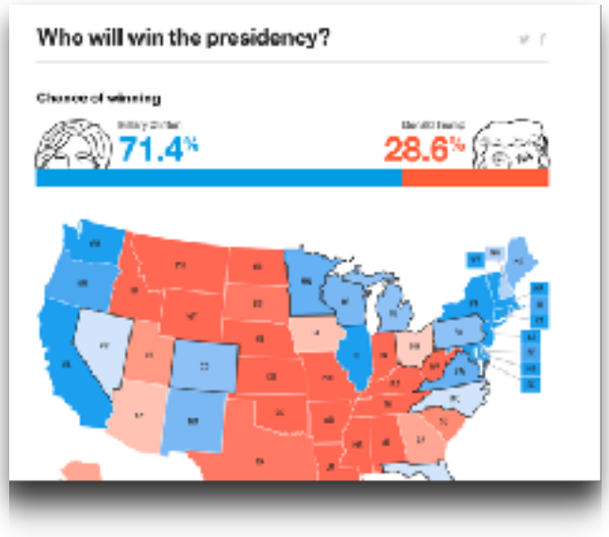
Probabilities are traditionally defined by counting the frequency of events over multiple trials.

- This is the **frequentist** view

The **Bayesian** view is that probabilities provide a numerical measure of belief in an outcome or event, even if they are unique.

- They can be applied to any problem which has uncertainty





Bayesian Probabilities

Do we have to use Bayesian probabilities to represent uncertainty?

- No, but according to Cox's Theorem you probably are anyway

In short: any representation of uncertainty which is consistent with boolean logic is equivalent to standard probability theory.

[Richard Cox]

AMERICAN JOURNAL of PHYSICS

A Journal Devoted to the Instructional and Cultural Aspects of Physical Science

VOLUME 14, NUMBER 1 JANUARY–FEBRUARY, 1946

Probability, Frequency and Reasonable Expectation

R T. Cox
The Johns Hopkins University, Baltimore 18, Maryland

What are Bayesian Methods?

Bayesian methods attempt to capture and maintain uncertainty.

Consists of two main steps:

- **Modelling:** capturing the available knowledge about a set of variables
- **Inference:** given a model and a set of data, computing the distribution of unknown variables of interest

Bayesian Modelling

In modelling use domain knowledge to define the distribution $p(\Theta|\mathcal{D})$

- Θ are parameters we want to know about
- $\mathcal{D}$ is the data that we have

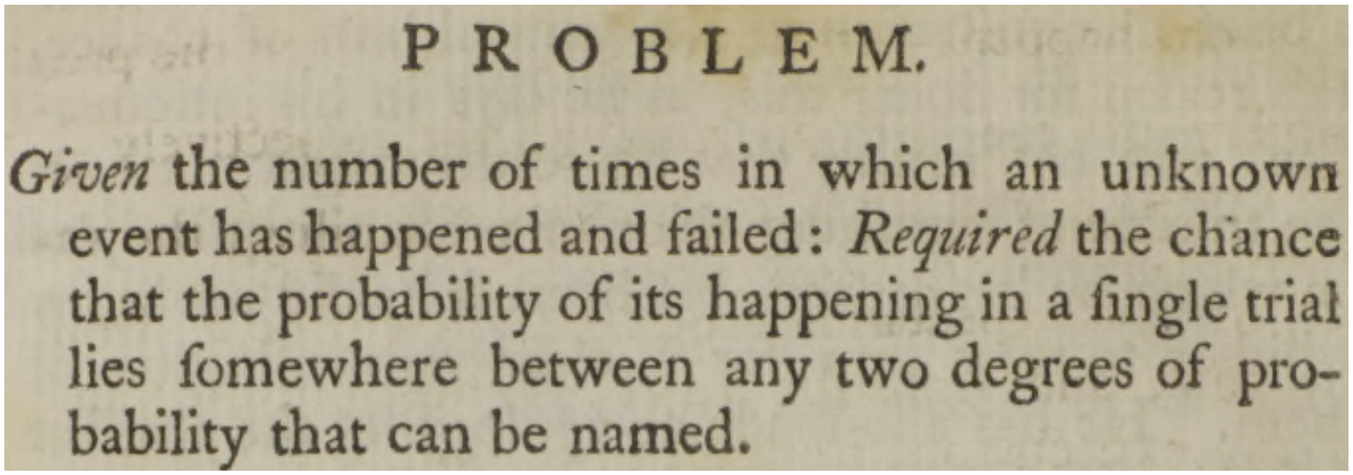
This is called the *posterior distribution*

- Encapsulates all knowledge about Θ given the prior knowledge used to construct the posterior and the data $\mathcal{D}$

Bayesian Modelling

How do we define the posterior?

Rev Thomas Bayes wrote a paper answering this question:



PROBLEM.

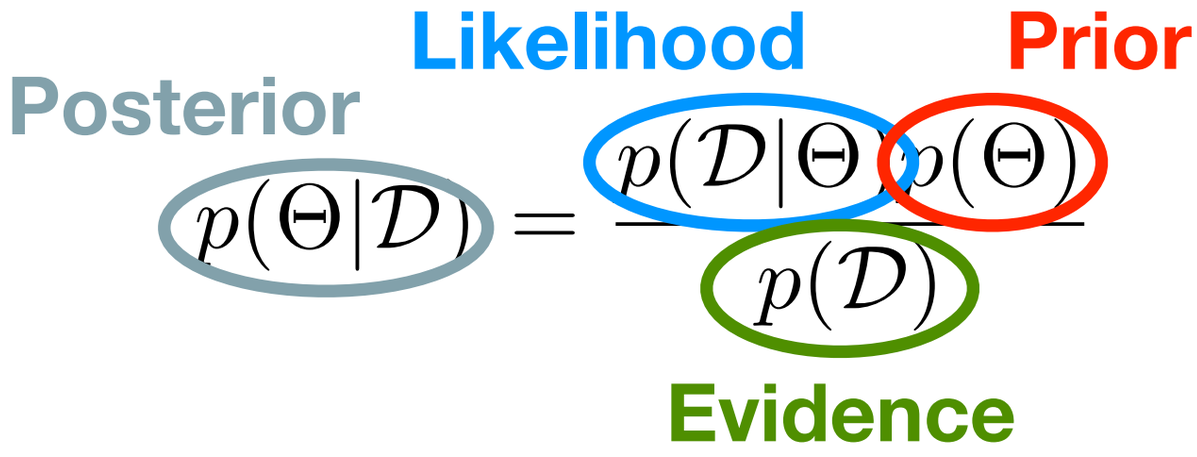
Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

[*Philosophical Transactions of the Royal Society*, vol 53 (1763)]

[Rev. Thomas Bayes]

This led to the first description of *Bayes' Rule*

Bayes' Rule



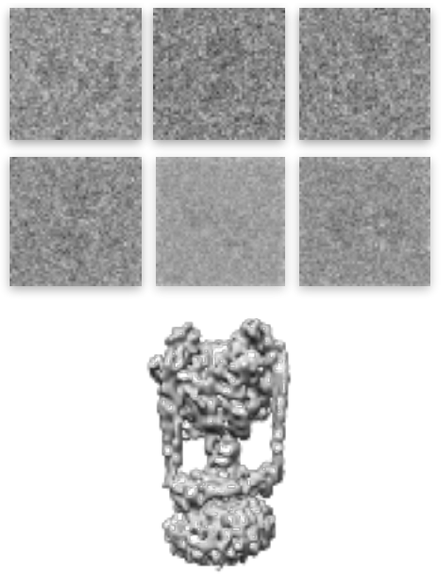
The posterior consists of

- the **likelihood** $p(\mathcal{D}|\Theta)$
- the **prior** $p(\Theta)$

The **evidence** is determined by the likelihood and the prior

Bayesian Modelling for Structure Estimation

Consider the problem of estimating a structure from a particle stack.



- $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_N\}$: stack of particle images
- $\Theta = \mathcal{V}$: 3D structure

A common prior is a Gaussian equivalent to Wiener filter

$$p(\Theta) = \mathcal{N}(\mathcal{V}|0, \Sigma)$$

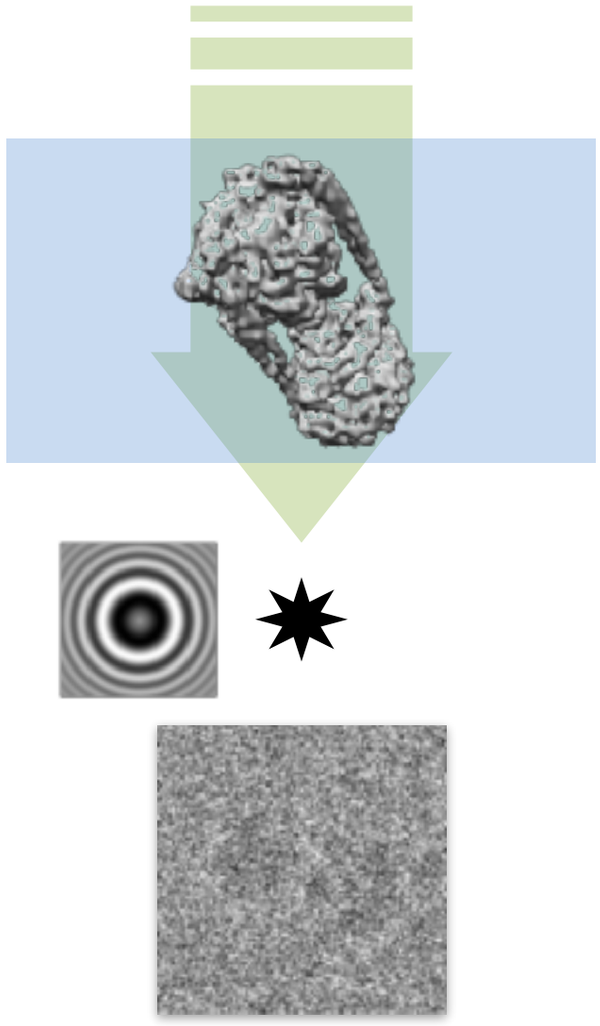
- Many other choices possible

What about the likelihood?

$$p(\mathcal{D}|\Theta) = \prod_{i=1}^{N} p(\mathcal{I}_i|\mathcal{V})$$

Particle Image Likelihood in Cryo-EM

An image $\mathcal{I}$ of a 3D density $\mathcal{V}$ in a pose given by 3D rotation $\mathbf{R}$ and 2D offset $\mathbf{t}$



Integral Projection

Noise

$$\mathcal{I} = \mathbf{C}\,\mathbf{P}_{\mathbf{R},\mathbf{t}}\,\mathcal{V} + \epsilon$$

Contrast Transfer Function

3D Density

Additive Gaussian Noise

$$p(\mathcal{I} \mid \mathbf{R}, \mathbf{t}, \mathcal{V}) = \mathcal{N}(\mathcal{I} \mid \mathbf{C}\,\mathbf{P}_{\mathbf{R},\mathbf{t}}\,\mathcal{V}, \sigma^2\mathbf{I})$$

Particle Image Likelihood in Cryo-EM

Particle pose is unknown

$$p(\mathcal{I}\,|\,\mathcal{V}) \;=\; \int_{\mathbb{R}^2}\int_{\mathcal{SO}(3)} p(\mathcal{I},\mathbf{R},\mathbf{t}|\mathcal{V}_k)d\mathbf{R}d\mathbf{t} \qquad \text{Marginalization}$$

$$= \int_{\mathbb{R}^2}\int_{\mathcal{SO}(3)} p(\mathcal{I}|\mathbf{R},\mathbf{t},\mathcal{V})p(\mathbf{R})p(\mathbf{t})d\mathbf{R}d\mathbf{t}$$

What if there are multiple structures?

[Sigworth, *J. Struct. Bio.* (1998)]

Particle Likelihood with Structural Heterogeneity

If there are K different independent structures and each image is equally likely to be of any of the structures

$$\Theta = \{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$$

$$p(\mathcal{I}|\mathcal{V}_1, \ldots, \mathcal{V}_K) = \frac{1}{K}\sum_{k=1}^{K} p(\mathcal{I}|\mathcal{V}_k)$$

$$= \frac{1}{K}\sum_{k=1}^{K} \int_{\mathbb{R}^2} \int_{\mathcal{SO}(3)} p(\mathcal{I}|\mathbf{R}, \mathbf{t}, \mathcal{V}_k) p(\mathbf{R}) p(\mathbf{t}) d\mathbf{R} d\mathbf{t}$$

Particle Image Likelihood in Cryo-EM

Computing the marginal likelihood

$$p(\mathcal{I} \mid \mathcal{V}) = \int_{\mathbb{R}^2} \int_{\mathcal{SO}(3)} p(\mathcal{I}|\mathbf{R}, \mathbf{t}, \mathcal{V}) p(\mathbf{R}) p(\mathbf{t}) d\mathbf{R} d\mathbf{t}$$

$$\approx \sum_j w_j p(\mathcal{I}|\mathbf{R}_j, \mathbf{t}_j, \mathcal{V})$$

Requires Numerical Approximation

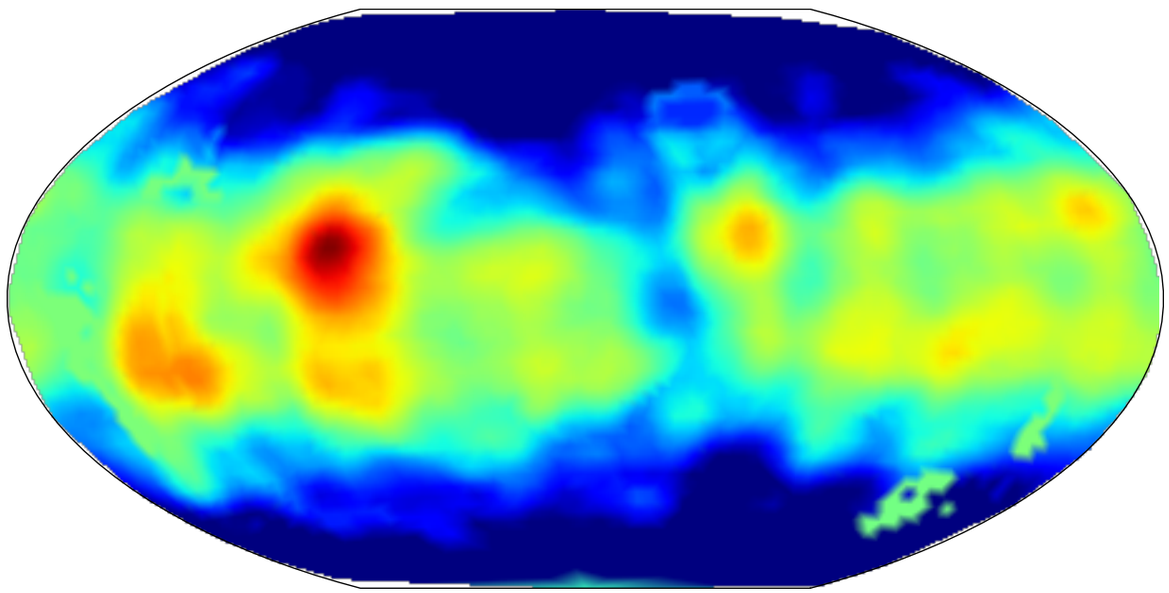
Many different approximations:

- Importance sampling [Brubaker et al. IEEE CVPR (2015); IEEE PAMI (2017)]
- Numerical quadrature [e.g., Scheres et al, *J. Mol.* Bio. (2012); RELION, Xmipp, etc]
- Point approximations [e.g., cryoSPARC; Projection Matching Algorithms]

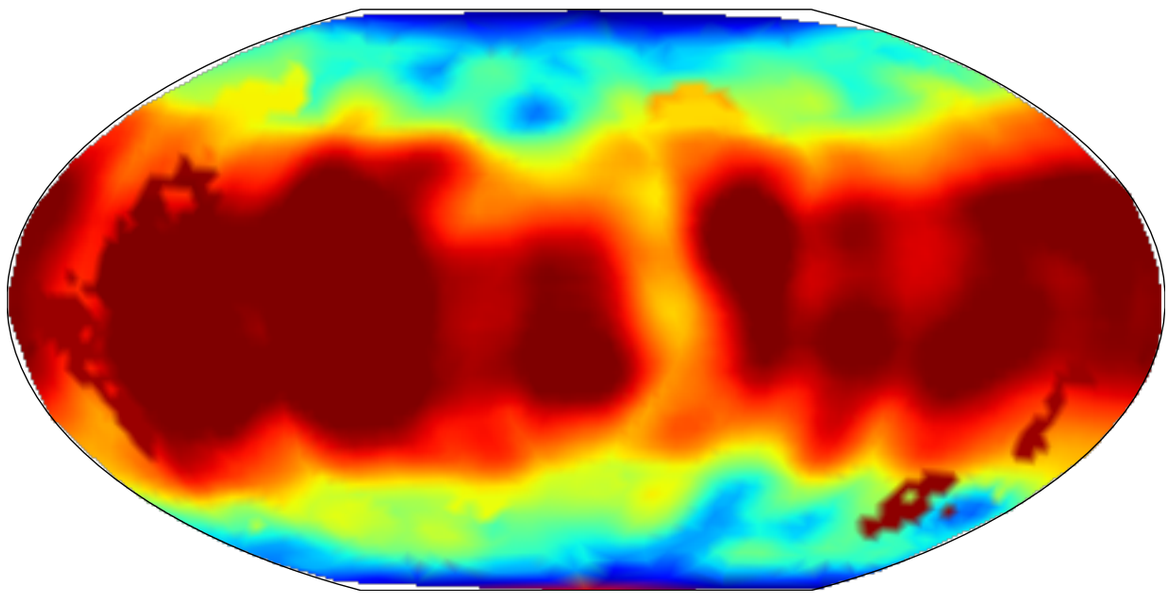
Approximate Marginalization

Integration over viewing direction

Structure at 10Å



Structure at 35Å



High Probability — Low Probability

Particle Image Likelihood in Cryo-EM

Instead of marginalization can estimate poses

- Include poses in variables to estimate

$$\Theta = \{\mathcal{V}, \mathbf{R}_1, \mathbf{t}_1, \ldots, , \mathbf{R}_N, \mathbf{t}_N\}$$

- Likelihood becomes

$$p(\mathcal{D}|\Theta) = \prod_{i=1}^{N} p(\mathcal{I}_i|\mathbf{R}_i, \mathbf{t}_i, \mathcal{V})$$

- This is equivalent to projection matching approaches/point approximations

- Marginalizing over poses makes inference better behaved (Rao-Blackwell Theorem)

Bayesian Inference

The posterior $p(\Theta|\mathcal{D})$ is then used to make inferences

- What value of the parameters is most likely?

$$\arg\max_{\Theta} p(\Theta|\mathcal{D})$$

- What is the average (or expected) value of the parameters?

$$E[\Theta] = \int \Theta p(\Theta|\mathcal{D}) d\Theta$$

- How likely are the parameters to lie in a given range?

$$p(\Theta_0 \leq \Theta \leq \Theta_1|\mathcal{D}) = \int_{\Theta_0}^{\Theta_1} p(\Theta|\mathcal{D}) d\Theta$$

- How much uncertainty in a parameter? Are multiple parameter values are plausible? Many others…

- Inference is rarely analytically tractable

Bayesian Inference

Two major approaches to inference

Sampling $\Theta_j \sim p(\Theta|\mathcal{D})$

- If posterior uncertainty is needed

$$E[f(\Theta)] = \int f(\Theta)p(\Theta|\mathcal{D})d\Theta \approx \frac{1}{M}\sum_{j=1}^{M} f(\Theta_j)$$

- Almost always requires approximations and very expensive

Optimization for Bayesian Inference

Optimization often only practical choice for large problems

$$\begin{aligned}\arg\max_{\Theta} p(\Theta|\mathcal{D}) &= \arg\min_{\Theta} -\log p(\Theta)p(\mathcal{D}|\Theta) \\ &= \arg\min_{\Theta} O(\Theta)\end{aligned}$$

Sometimes referred to as the “Poor Mans Bayesian Inference”

Many different kinds of optimization algorithms

- Derivative free (brute-force search, simplex, …)
- Variational methods (expectation maximization, …)
- Gradient based (gradient descent, BFGS, …)

Gradient-based Optimization

Recall from calculus: negative gradient is the direction of fastest decrease

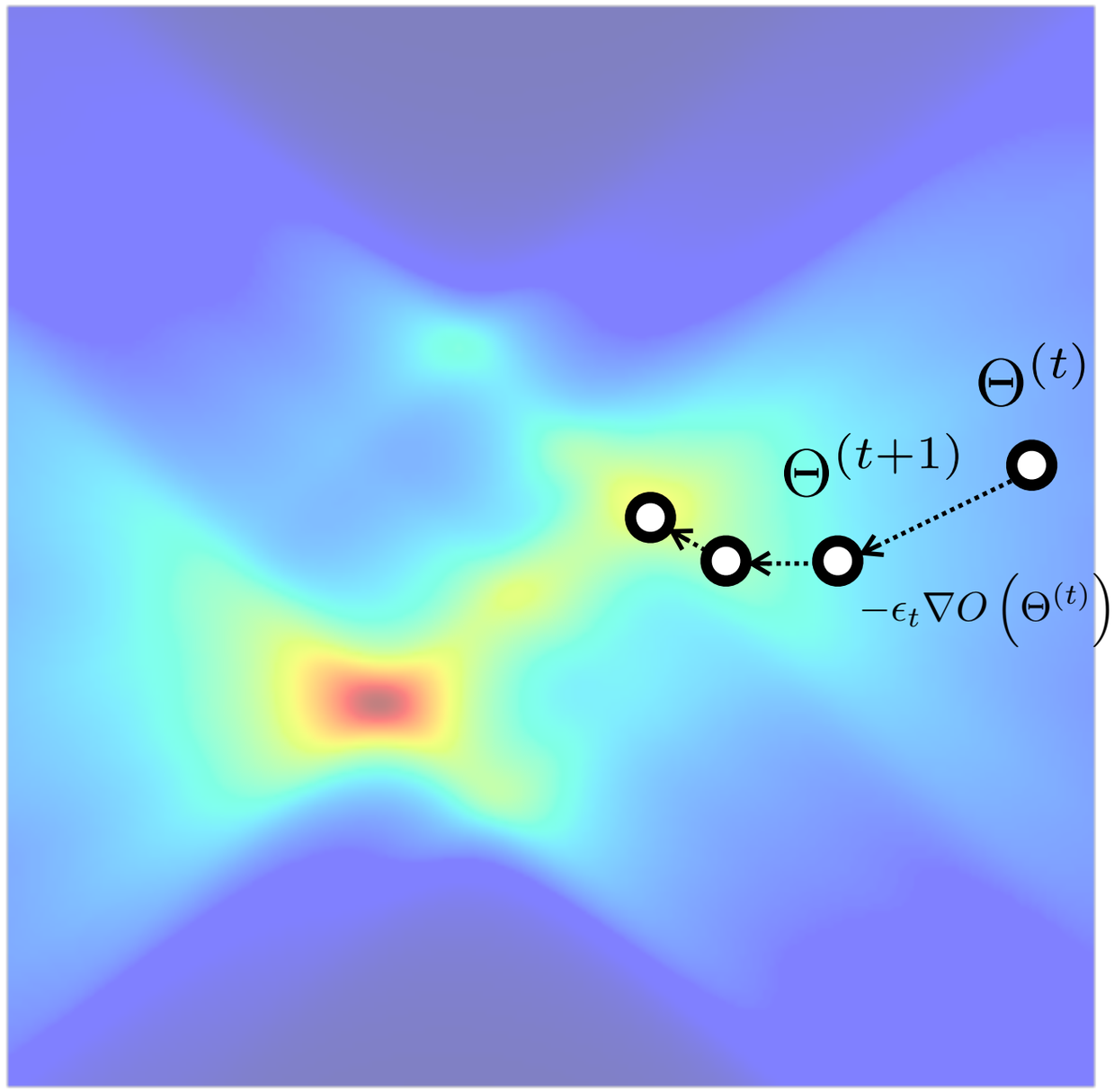
- All gradient-based algorithms iterate an equation like:

$$\Theta^{(t+1)} = \Theta^{(t)} - \epsilon_t \nabla O\left(\Theta^{(t)}\right)$$

Gradient of Objective Function

Variations include:

- CG [e.g., CTFFIND, J. Struct. Bio. (2003)]
- LBFGS [e.g., alignparts, *J. Struct. Bio.* (2014)]
- Many others [Nocedal and Wright (2006)]



Gradient-based Optimization

Problems with gradient-based optimization for structure estimation

- Large datasets means expensive to compute gradient
- Sensitive to initial value $\Theta^{(0)}$

Can we do better?

- Recall the objective function

$$\arg\min_{\Theta} O(\Theta) = \arg\min_{\mathcal{V}} O(\mathcal{V})$$

$$O(\mathcal{V}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\mathcal{V})$$

$$f_i(\mathcal{V}) = -\log p(\mathcal{V}) - N \log p(\mathcal{I}_i|\mathcal{V})$$

Gradient-based Optimization for CryoEM

Lets look at the objective more closely

$$O(\mathcal{V}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathcal{V})$$

Average Error Over Images

Optimization problems like this have been studied under various names

- M-estimators, risk minimization, non-linear least-squares, …

One algorithm has recently been particularly successful

- Stochastic Gradient Descent (SGD)
- Very successful in training neural nets and elsewhere

Stochastic Gradient Descent

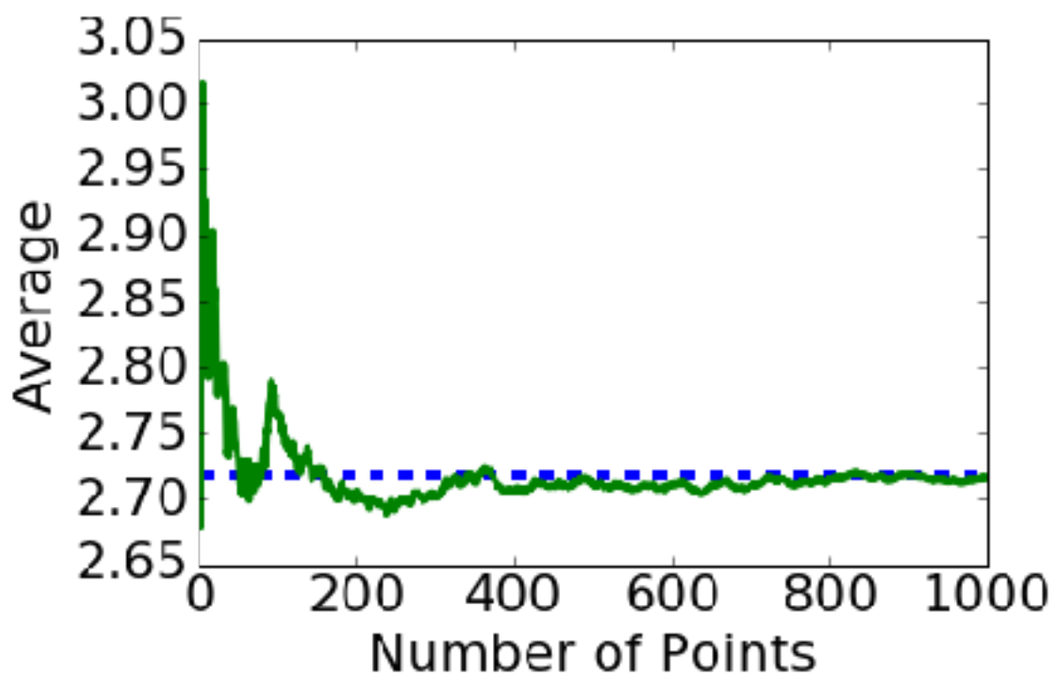
Consider computing the average of a large list of numbers

- `2.845, 3.157, 2.033, 3.483, 3.549, 3.031, 2.120, 3.211, 2.453, 3.155, 2.855, …`

Computing the exact answer is expensive

What if an approximate answer is sufficient?

- Average a random subset



SGD applies this intuition to approximate the objective function

Stochastic Gradient Descent

SGD approximates the objective using a random subset of terms

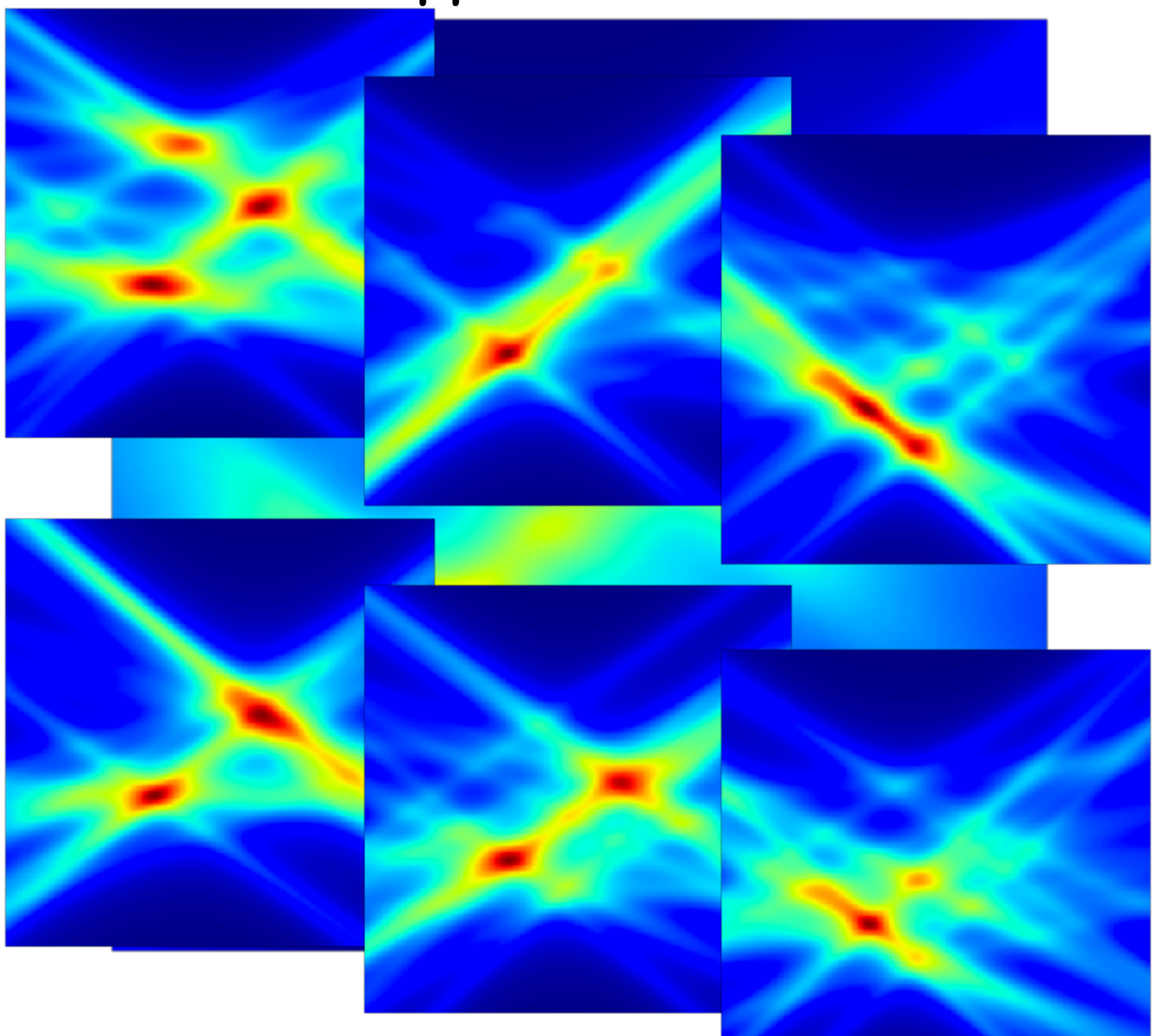
$$O(\mathcal{V}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathcal{V})$$

$$\approx \frac{1}{|\mathfrak{J}|}\sum_{i\in\mathfrak{J}} f_i(\mathcal{V})$$

Full Objective

Random Subset

Approximations



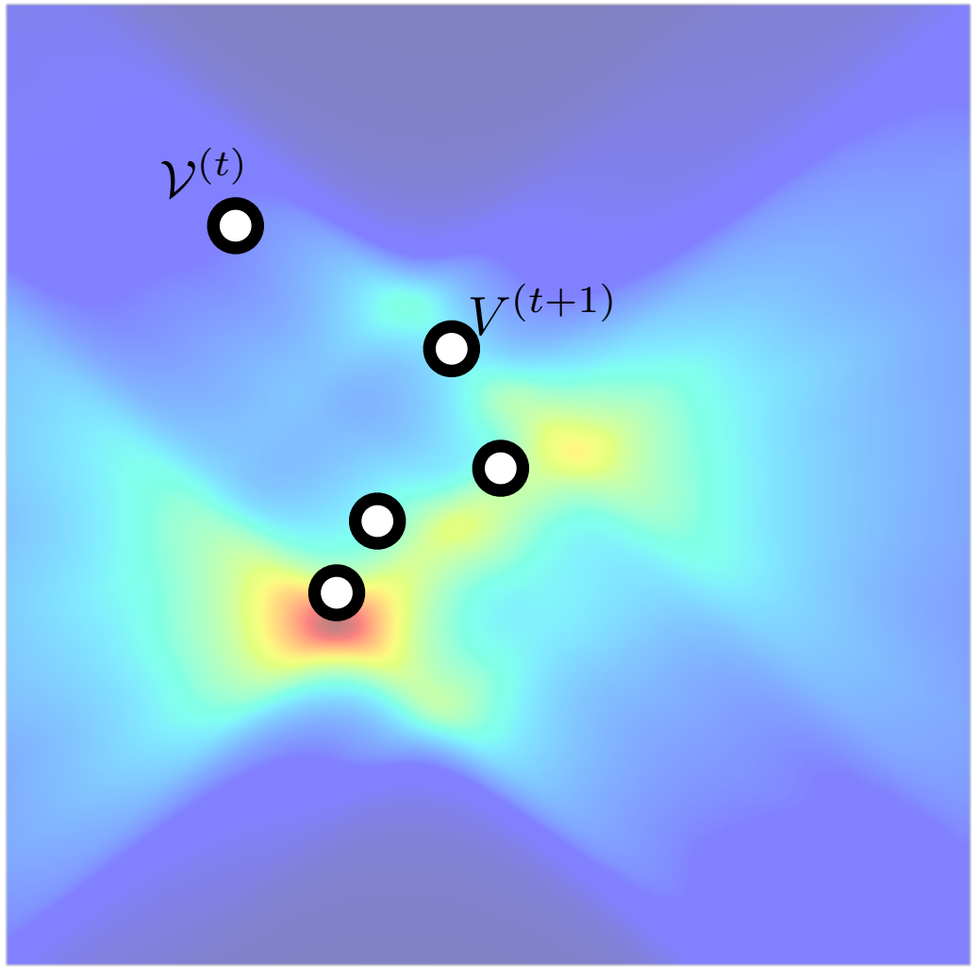
Stochastic Gradient Descent

The approximate gradient is then an average over the random subset $\mathfrak{J}$

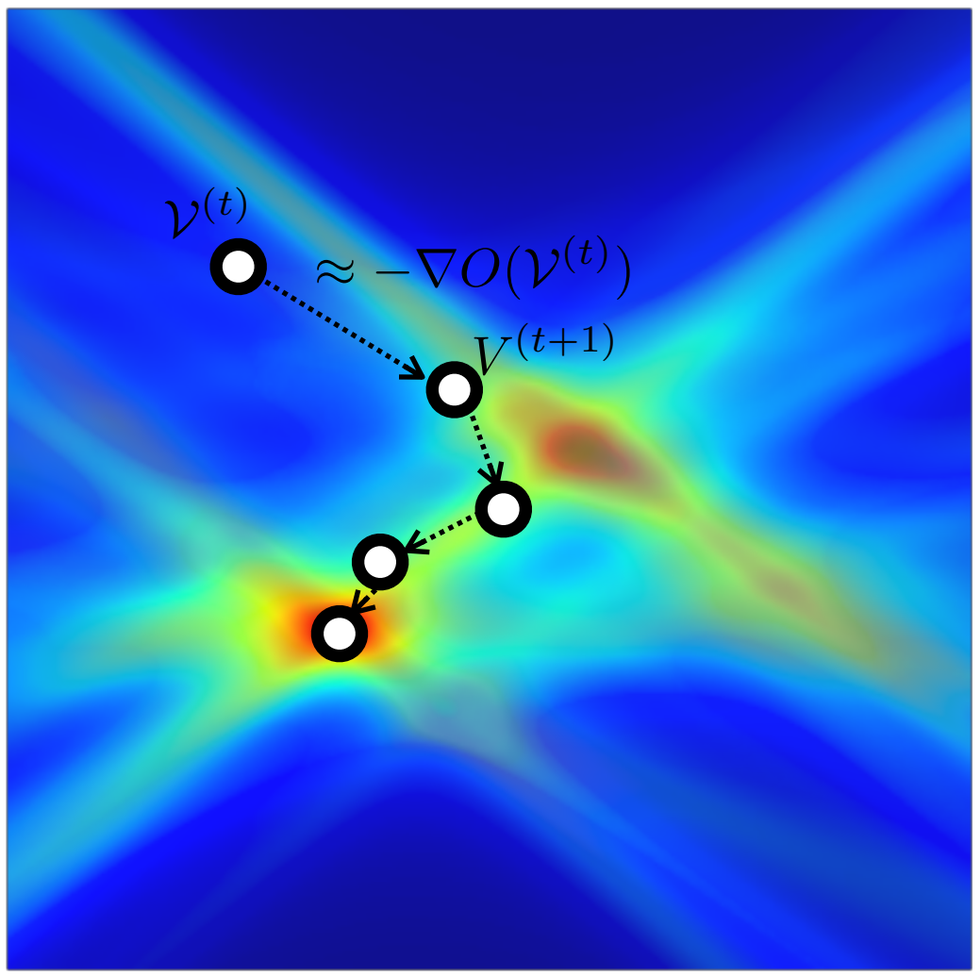
$$\nabla O(\mathcal{V}) \approx \frac{1}{|\mathfrak{J}|} \sum_{i \in \mathfrak{J}} \nabla f_i(\mathcal{V})$$

Random Subset

Exact Objective



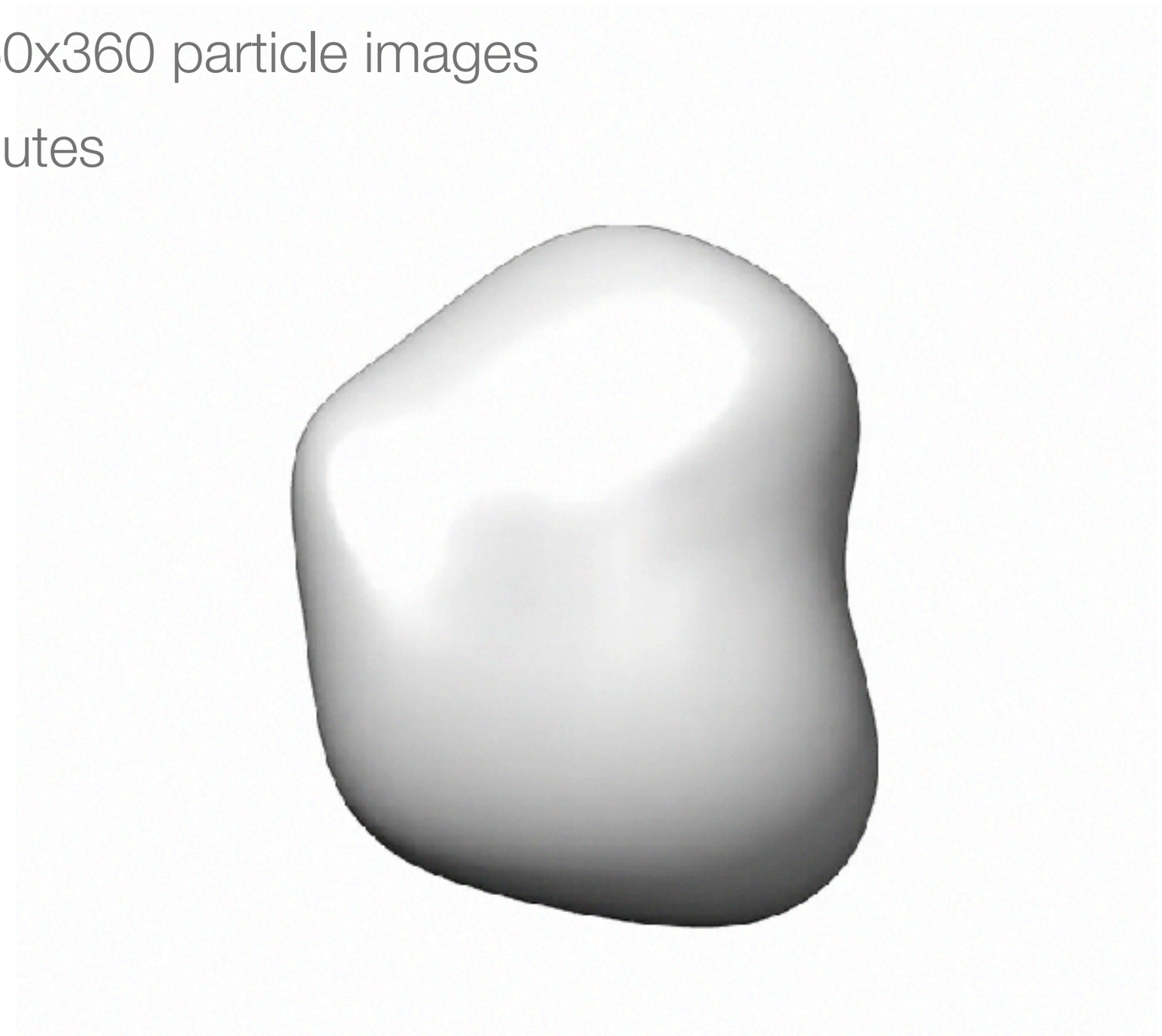
Approximation



Ab Initio Structure Determination with SGD

80S Ribosome [Wong et al 2014, EMPIAR-10028]

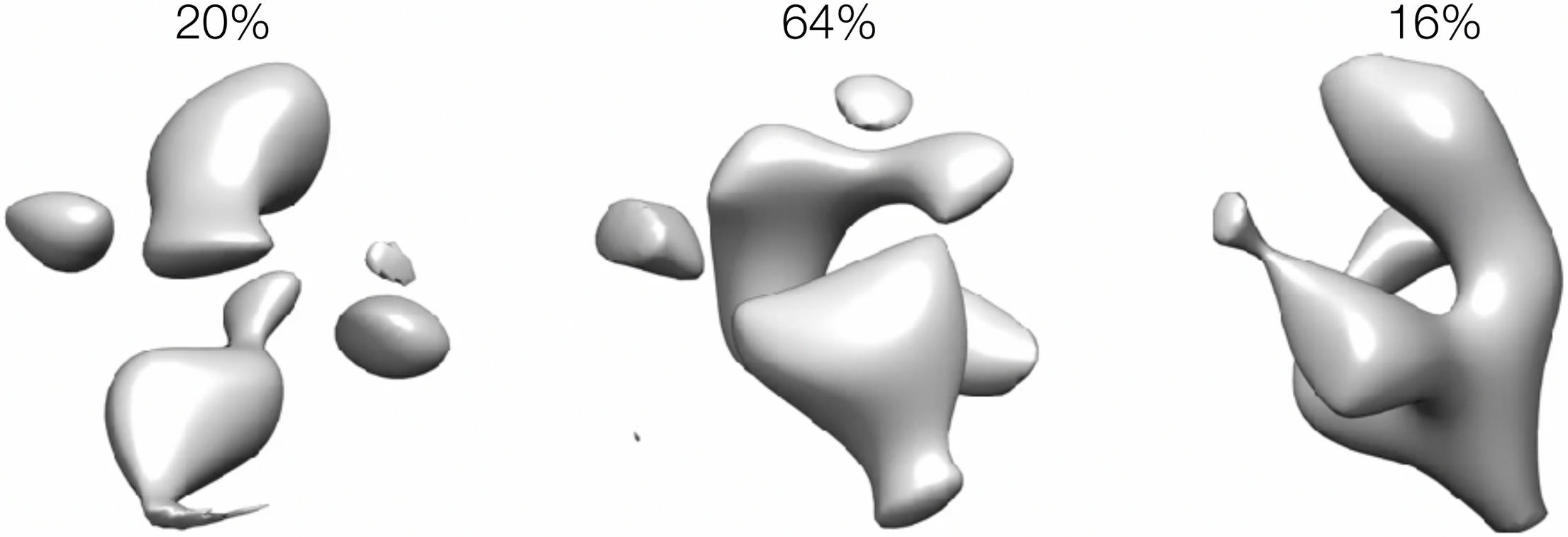
- 105k 360x360 particle images
- ~35 minutes



Ab Initio 3D Classification with SGD

T. thermophilus V/A-type ATPase [Schep et al 2016]

- 120k 256x256 particles from an F20/K2,
- ~3 hours



Stochastic Gradient Descent

Computational cost determined by number of samples, not dataset size

- Surprisingly small numbers of samples can work
- Only need a direction to move which is “good enough”

Applicable to any differentiable error function

- Projection matching, likelihood models, 3D classification, …

In theory converges to a local minima

- In practice, often converges to good (global?) minima
- Not theoretically understood but widely observed
- Ideally suited to ab initio structure estimation

Conclusions

Bayesian Methods provide a framework for problems with uncertainty

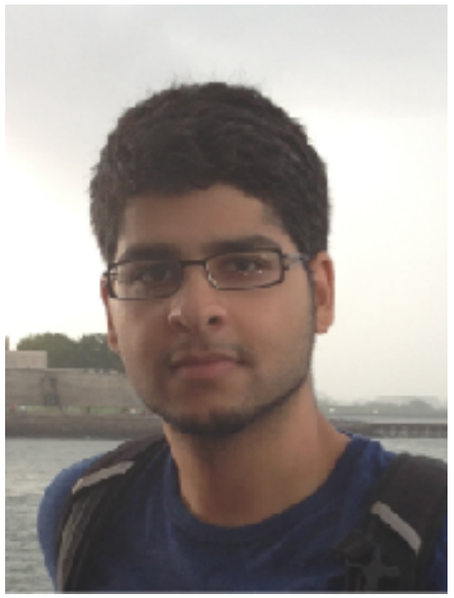
- Allows us to incorporate domain specific knowledge in a principled manner in the form of the likelihood model and priors
- Limitations of our image processing algorithms can be understood as limitations or poor assumptions built into our models (e.g., discrete vs continuous heterogeneity)

Defining better models is usually easy

- Inference and good approximations are the hard part
- No need to reinvent the wheel, many of our problems are well trodden ground (e.g., optimization)

Thanks! Questions?

Looking for interns, graduate students, postdocs, etc!



Ali Punjani
University of Toronto

David J Fleet
University of Toronto

John Rubinstein
Sick Kids Hospital / University of Toronto

CIFAR
CANADIAN INSTITUTE for ADVANCED RESEARCH

NSERC CRSNG