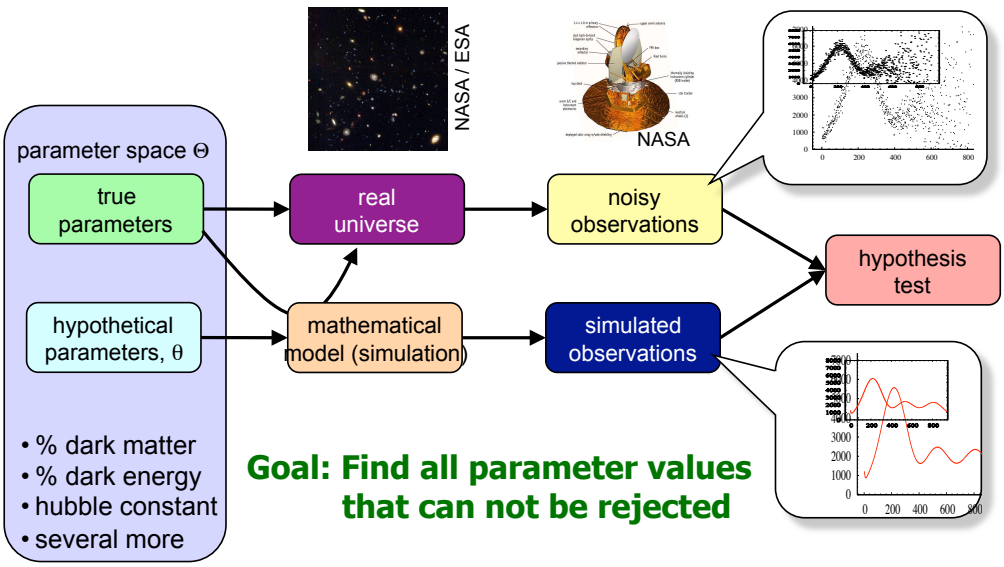
Active Learning for Fitting Simulation Models to Observational Data

Jeff Schneider

Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

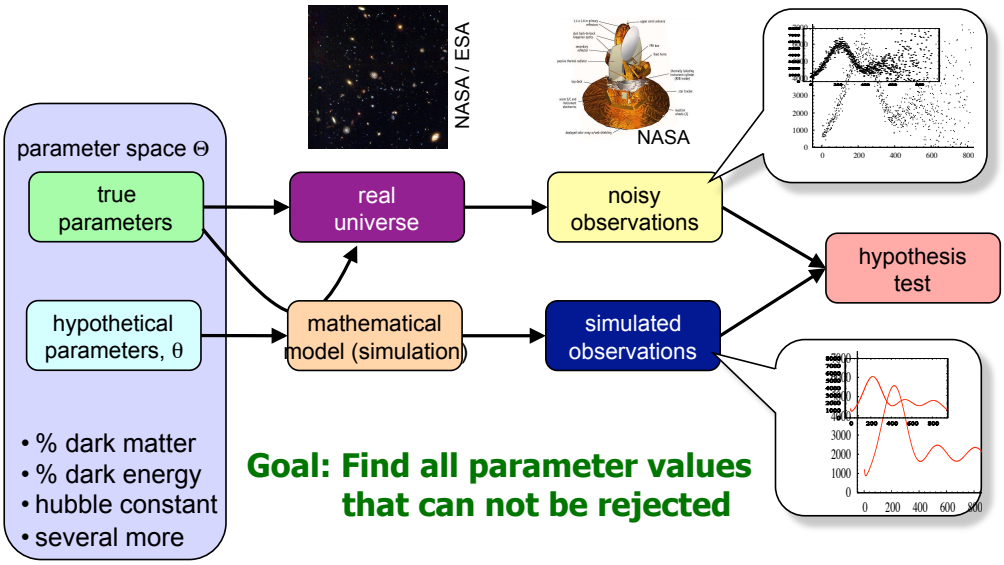
Example: Active Learning for Cosmology Model Fitting



Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Example: Active Learning for Cosmology Model Fitting

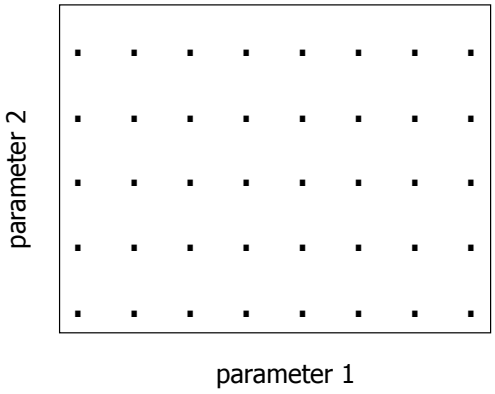


How should we search this high dimensional space?

Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Use a Grid?



- A popular method
- Guarantees good coverage
- Doesn't scale well with number of parameters (exponential)
- Does not focus effort on most important areas

Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

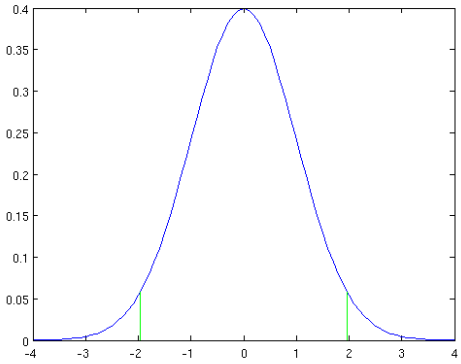
MCMC for Credible Region Construction

Metropolis-Hastings Algorithm

- choose θ^0
- generate proposal from $Q(\theta'|\theta^i)$
- accept proposal with probability $P(\theta')Q(\theta^i|\theta')/P(\theta^i)Q(\theta'|\theta^i)$
- if accepted $\theta^{i+1} = \theta'$
- else $\theta^{i+1} = \theta^i$
- repeat to step 2

- choosing $P(\theta) = f(\theta|x)f(\theta)$ yields samples from the posterior distribution
- Q is often chosen to be a normal distribution

suppose this is the true posterior:



Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

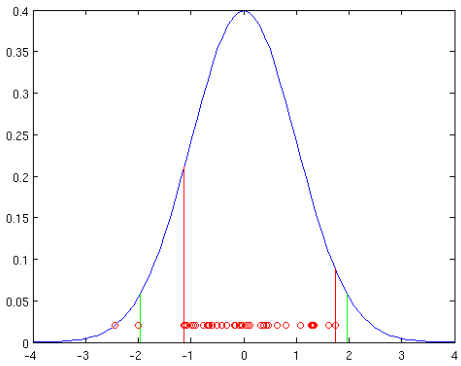
MCMC for Credible Region Construction

Metropolis-Hastings Algorithm

- choose θ^0
- generate proposal from $Q(\theta'|\theta^i)$
- accept proposal with probability $P(\theta')Q(\theta^i|\theta')/P(\theta^i)Q(\theta'|\theta^i)$
- if accepted $\theta^{i+1} = \theta'$
  else $\theta^{i+1} = \theta^i$
5. repeat to step 2

- choosing $P(\theta) = f(\theta|x)f(\theta)$ yields samples from the posterior distribution
- Q is often chosen to be a normal distribution

MCMC wastes samples in places of no interest and thus converges to the true region slowly:



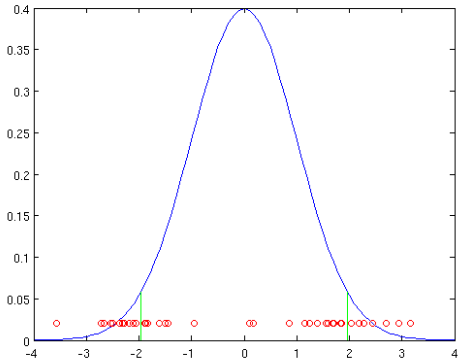
Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Active Learning for Confidence Region Construction?

We hope for sampling more like this:

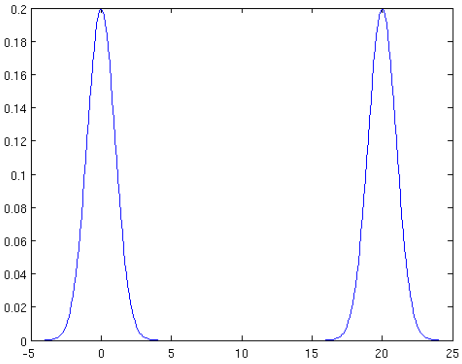
Is it possible?



Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

A Second Problem for MCMC



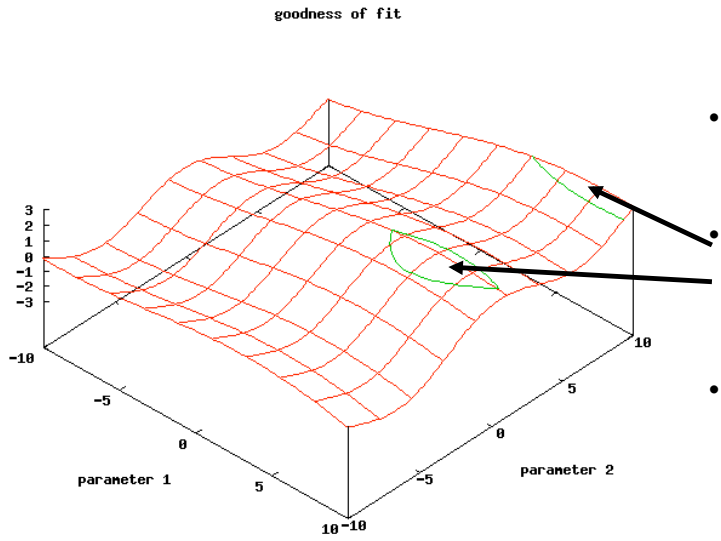
Especially in high dimensions it becomes nearly impossible to reach the second peak, and if you try, you fail to sample the first one well

MCMC is sampling, but a search algorithm is needed

Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

A Goodness of Fit Surface



- goodness of fit is the hypothesis test statistic
- goal: identify the regions not rejected by the hypothesis test
- use active learning to choose the parameters to test

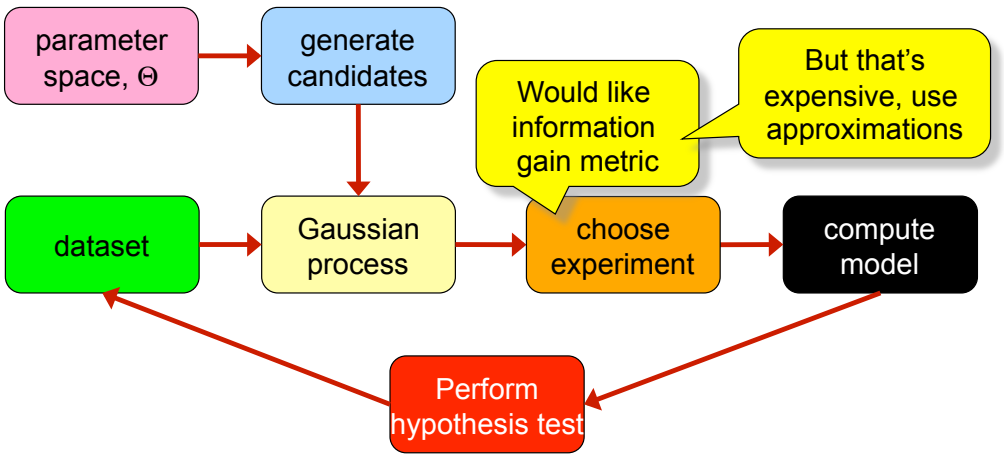
Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Computing Function Level-Sets

Approximate mapping from parameters to "goodness of fit" with Gaussian Process

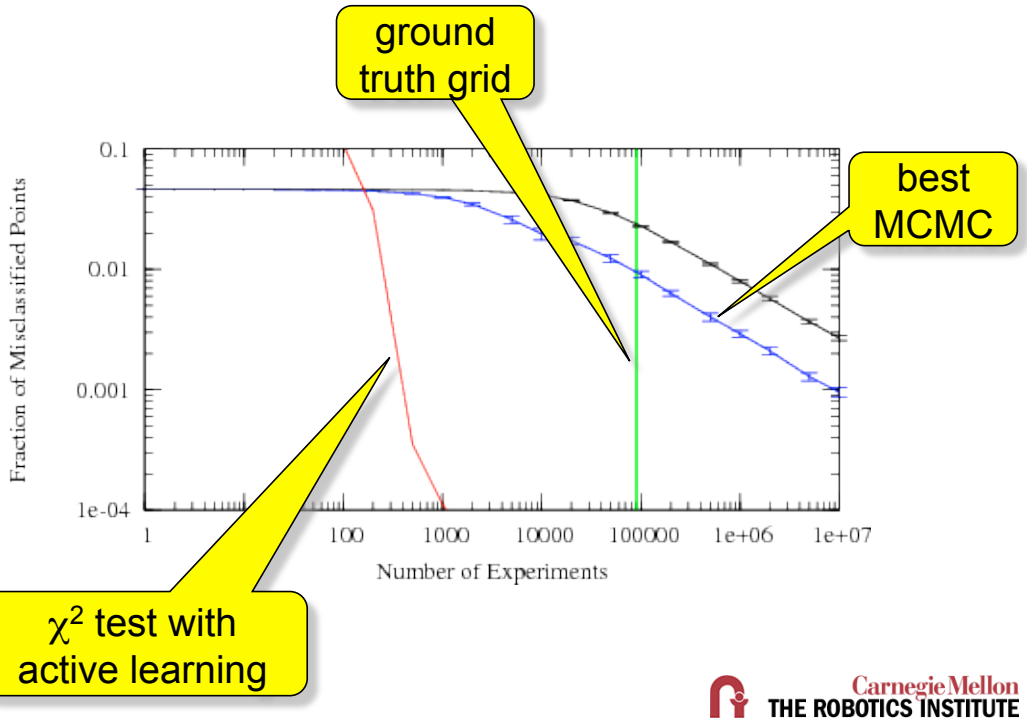
Sample Points to refine Gaussian Process



Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Performance Comparison



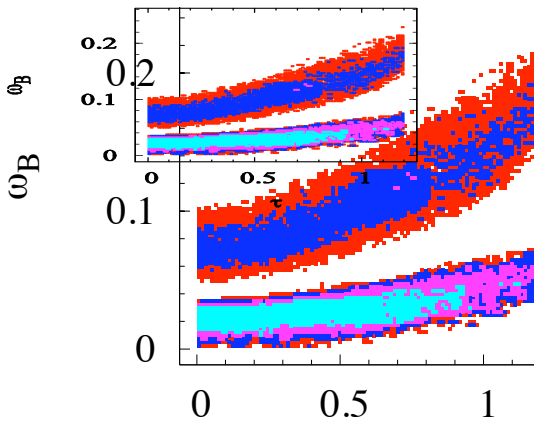
Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Cosmological Results

After over 1 million experiments …

A second plausible region of parameter space was identified



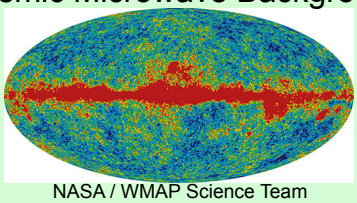
τ[Bryan et. al., Astrophysical Journal, 2007]

Auton Lab

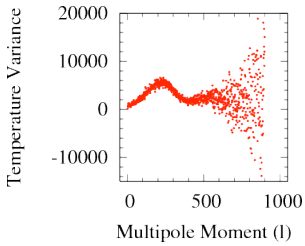
Carnegie Mellon THE ROBOTICS INSTITUTE

But other tests rule it out …

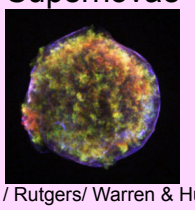
Cosmic Microwave Background



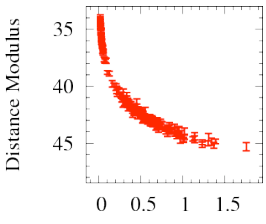
NASA / WMAP Science Team



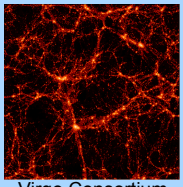
Supernovae



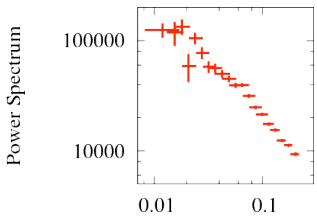
NASA / CXC / Rutgers/ Warren & Hughes et al.



Large Scale Structure



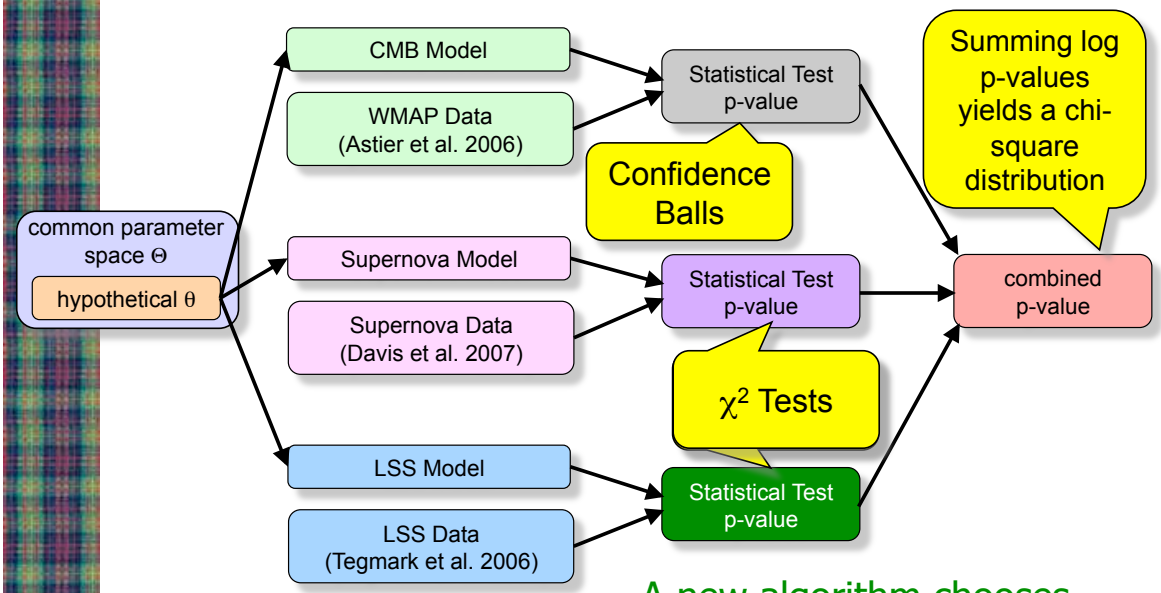
Virgo Consortium



Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Fisher's Method for Combining p-values



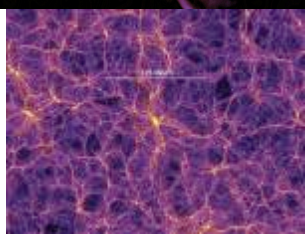
A new algorithm chooses both the parameters and which model to test with

[Bryan and Schneider, ICML, 2008]

Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Where Next?



- Simulation efforts have moved to large scale structure and galaxies
- Often require many more cycles and thus parallel execution
- The quality and computational expense are traded off by choosing the number of particles
- Current trend: "my simulation is bigger than yours"
- An alternative: run lots of smaller simulations to find better matches to observational data

Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Active Learning for Massive Scale Simulations

- An active learning algorithm must
  - choose parameters to test
  - choose how much to invest in the test
  - choose batches of tests
  - tradeoff throughput and latency (i.e. how many cores to put on each)
- How do we extend methods of "computation allocation" to other astrophysical problems?
  - computation limited data mining
  - telescope control for transients

Auton Lab

Carnegie Mellon THE ROBOTICS INSTITUTE

Why do we build simulations?

- To understand the dynamics of systems where we can not "watch" their evolution and thus learn the dynamics directly from data

Is it possible to learn dynamics from data that is not in trajectories?

see [Huang and Schneider, ICML 2009]