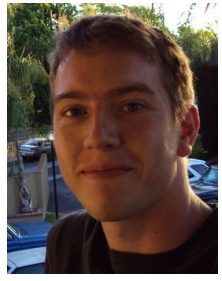
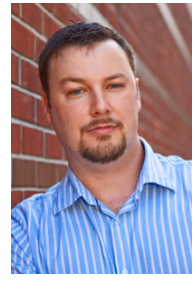
Early Stopping as Nonparametric Variational Inference

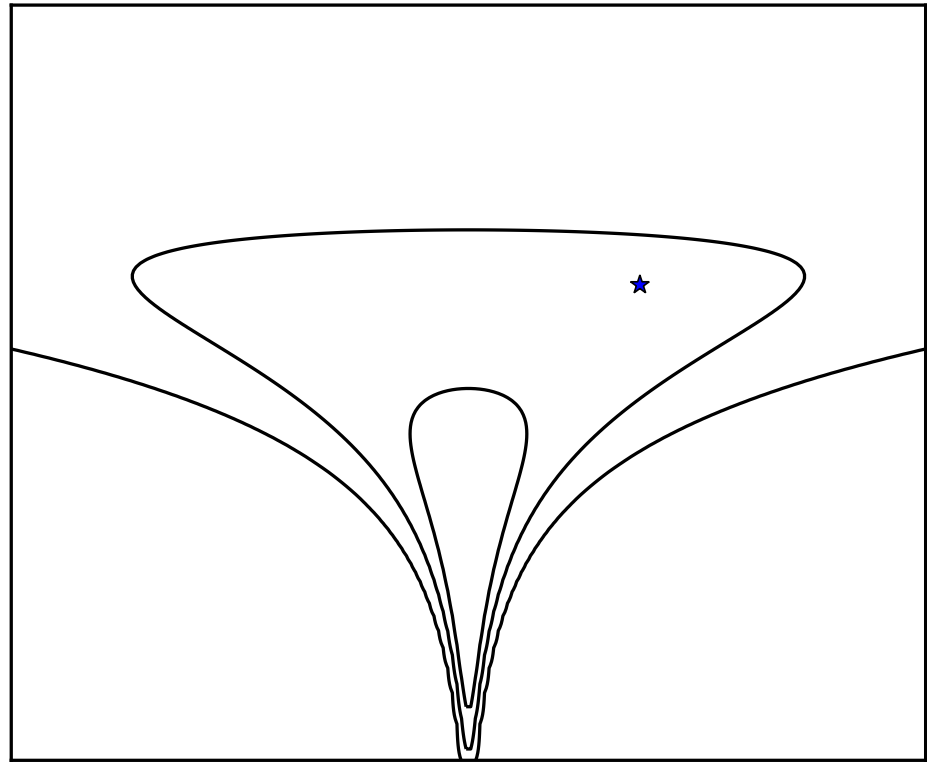
Dougal Maclaurin, David Duvenaud, Ryan Adams

Harvard University

June 15, 2015

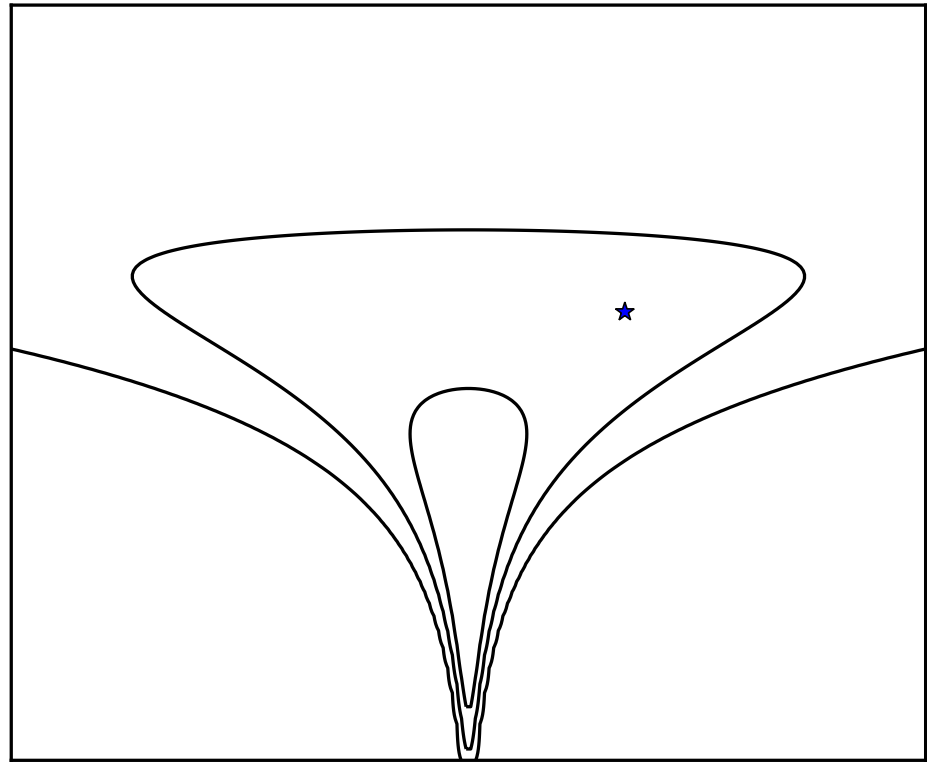
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



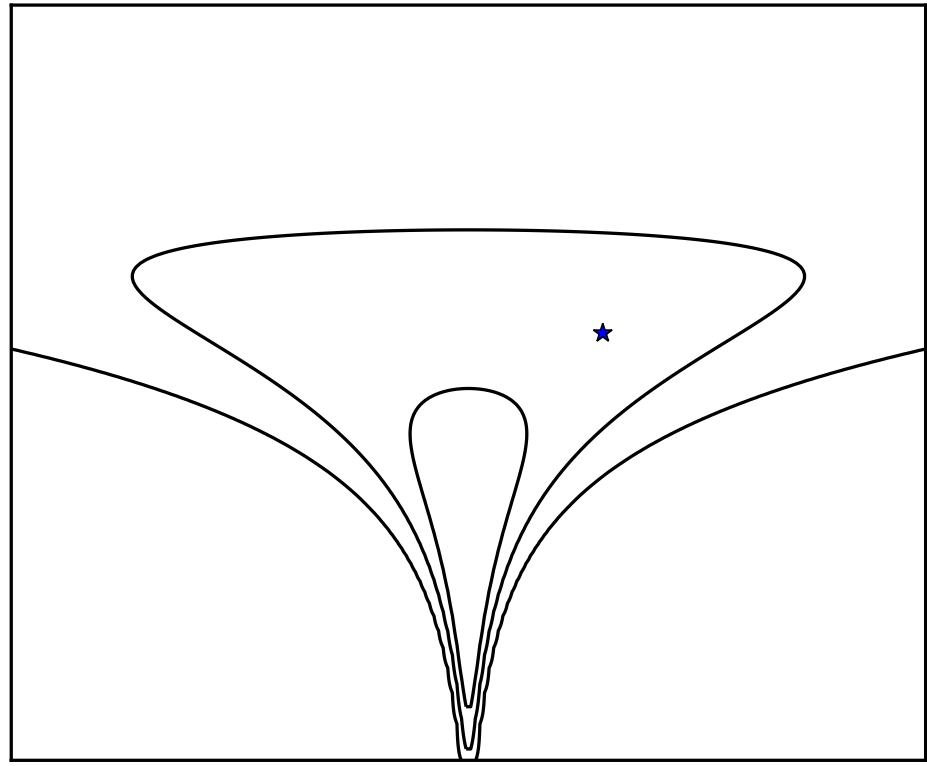
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



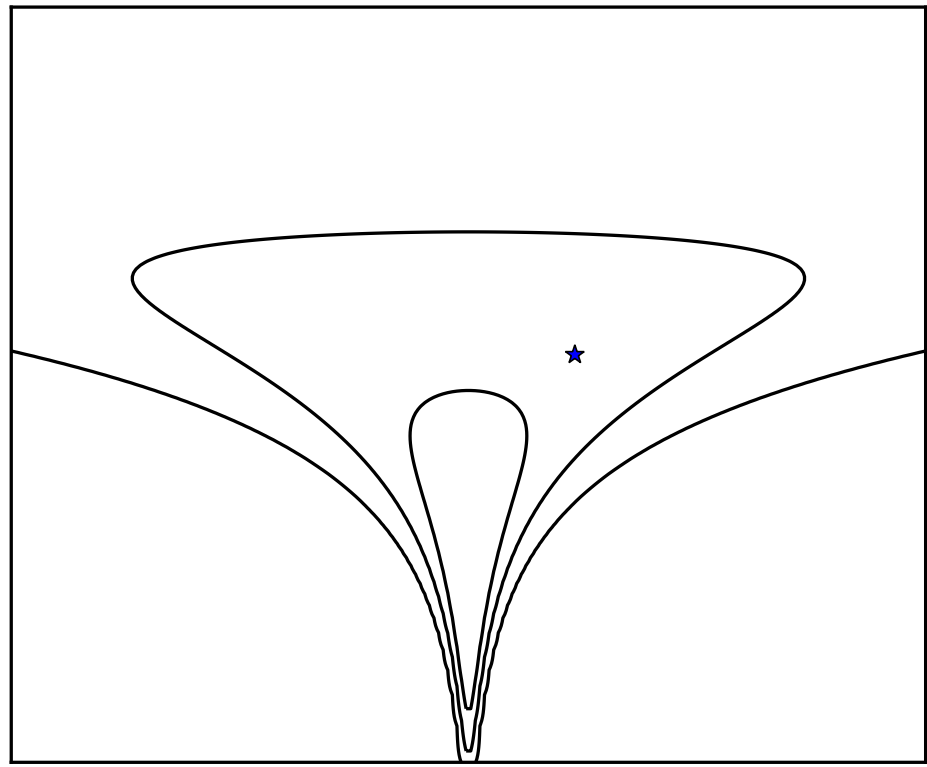
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



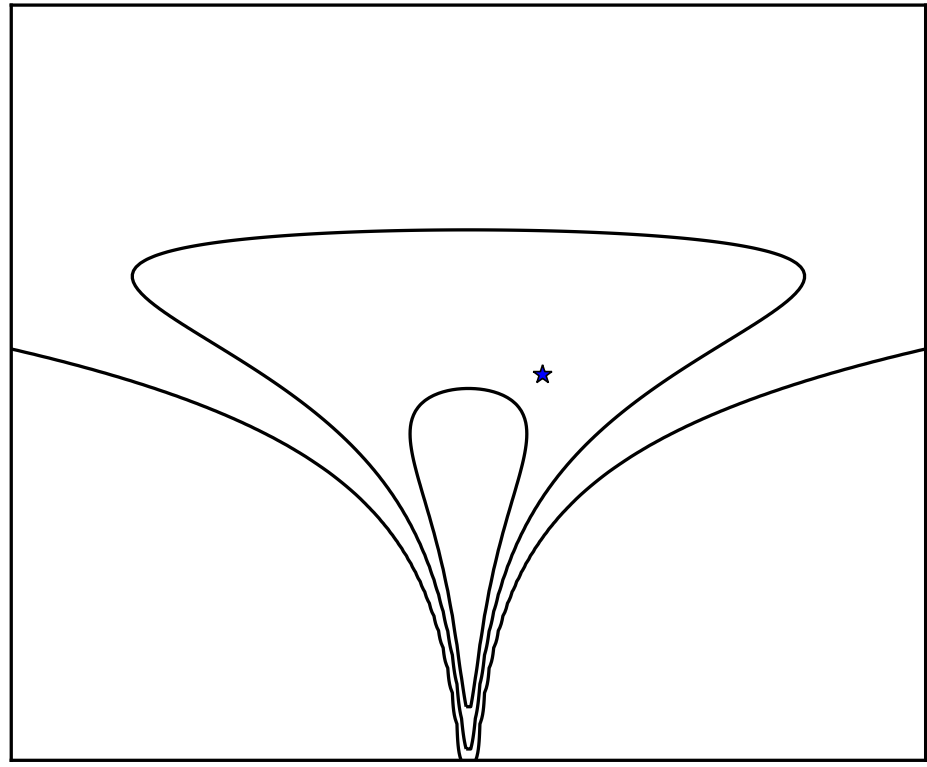
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



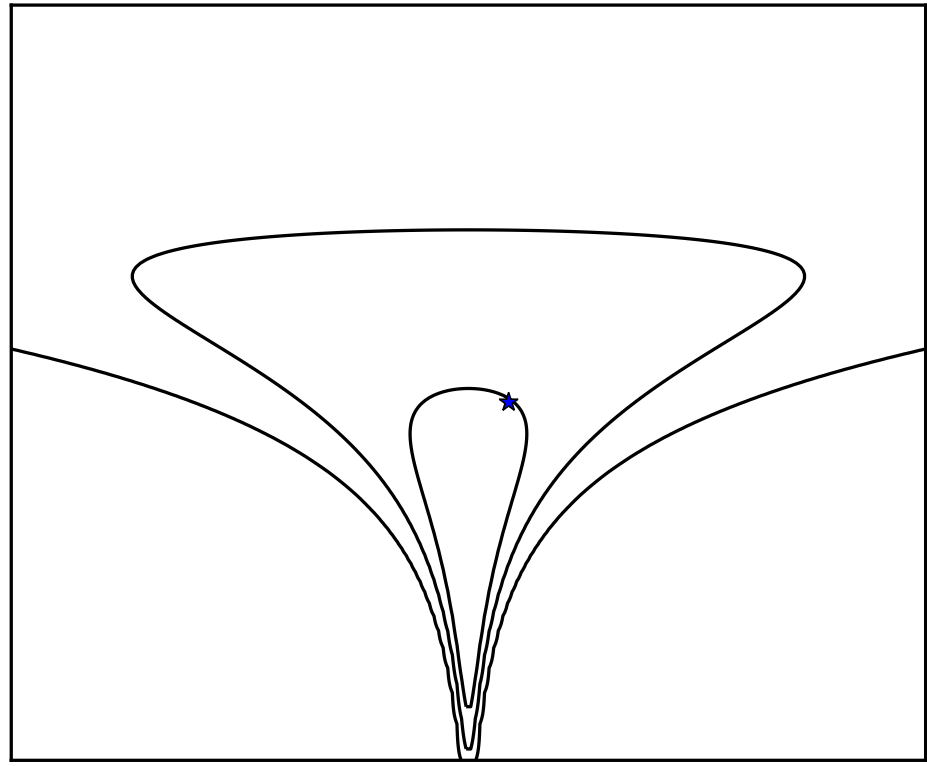
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



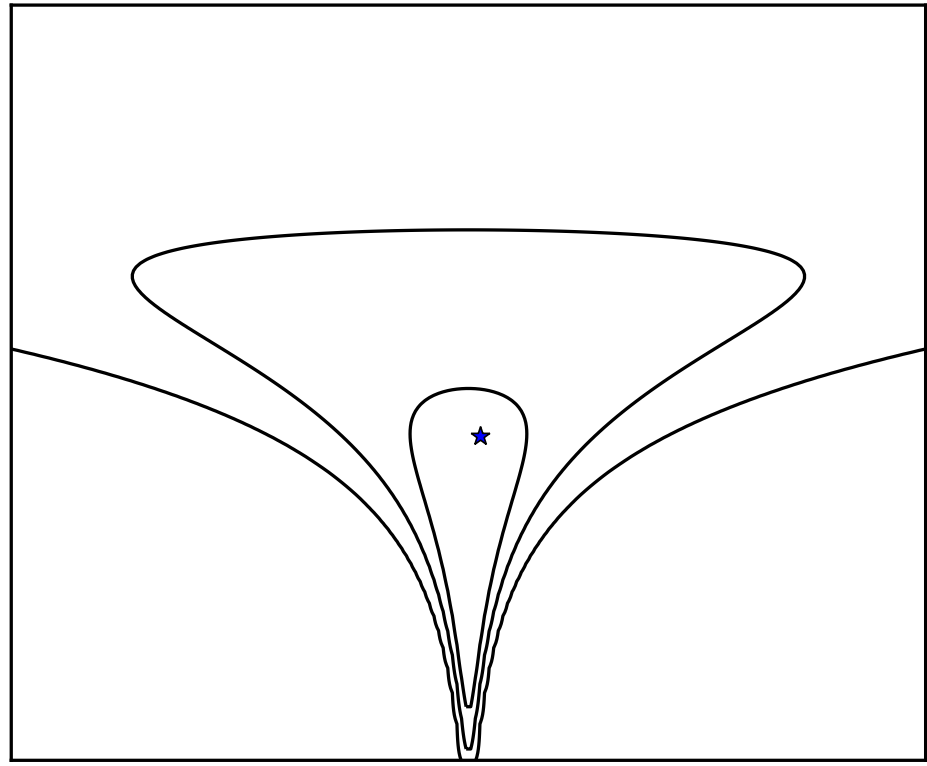
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



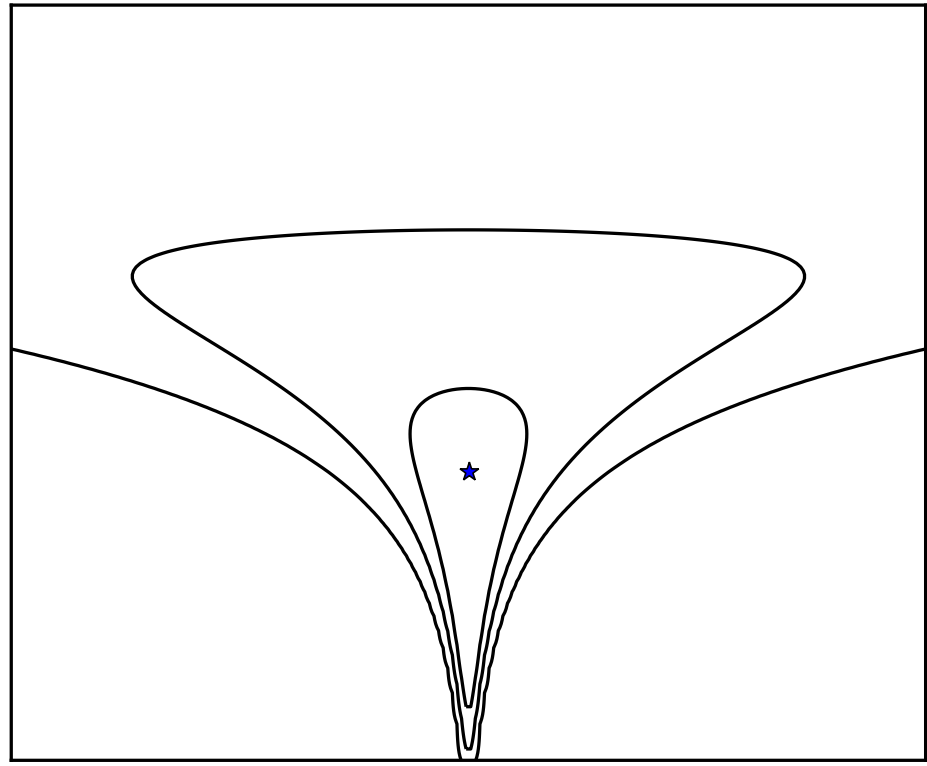
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



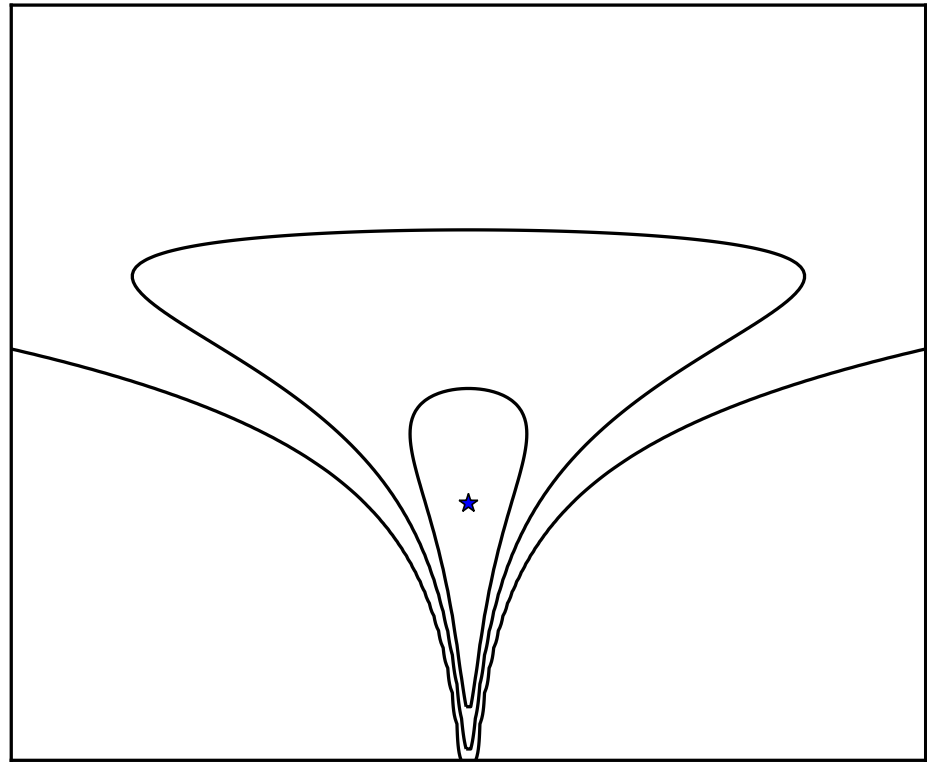
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



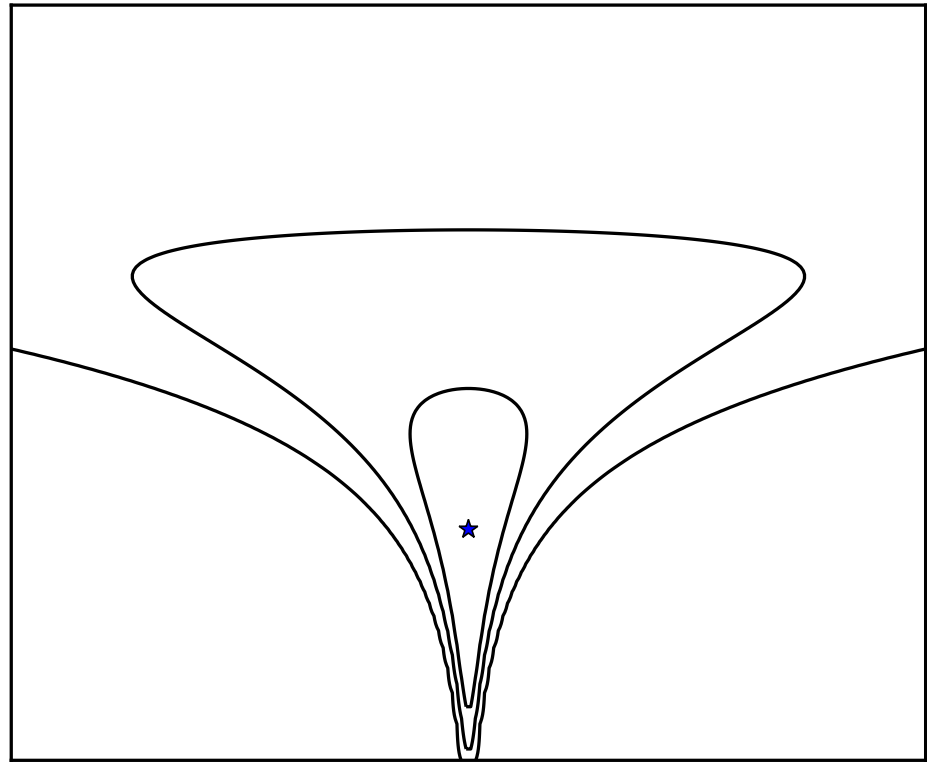
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



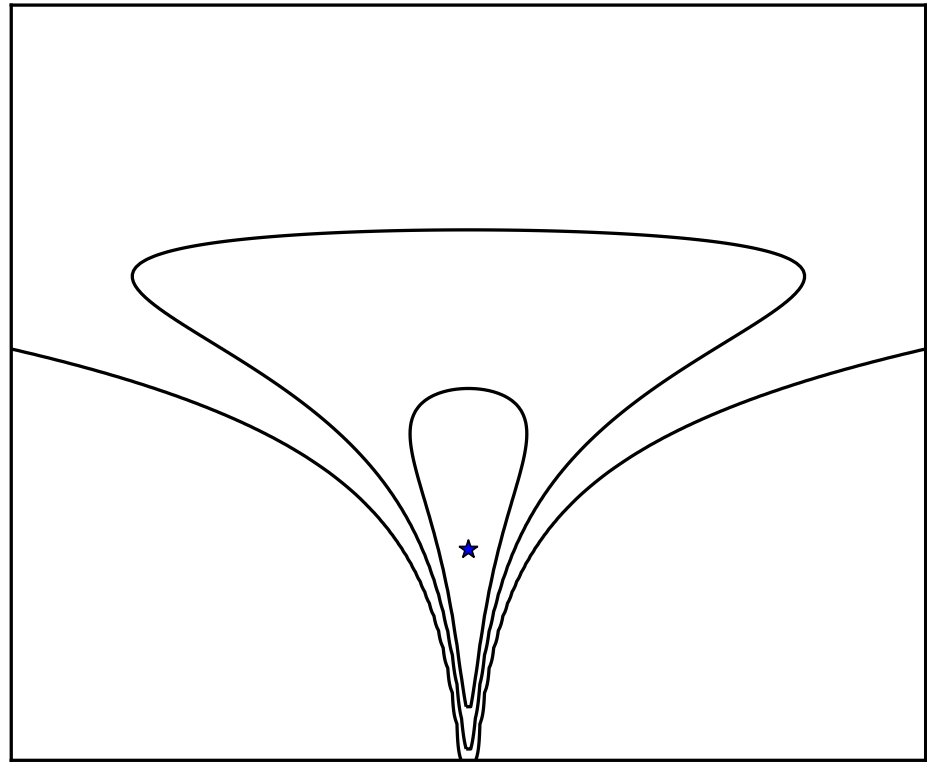
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



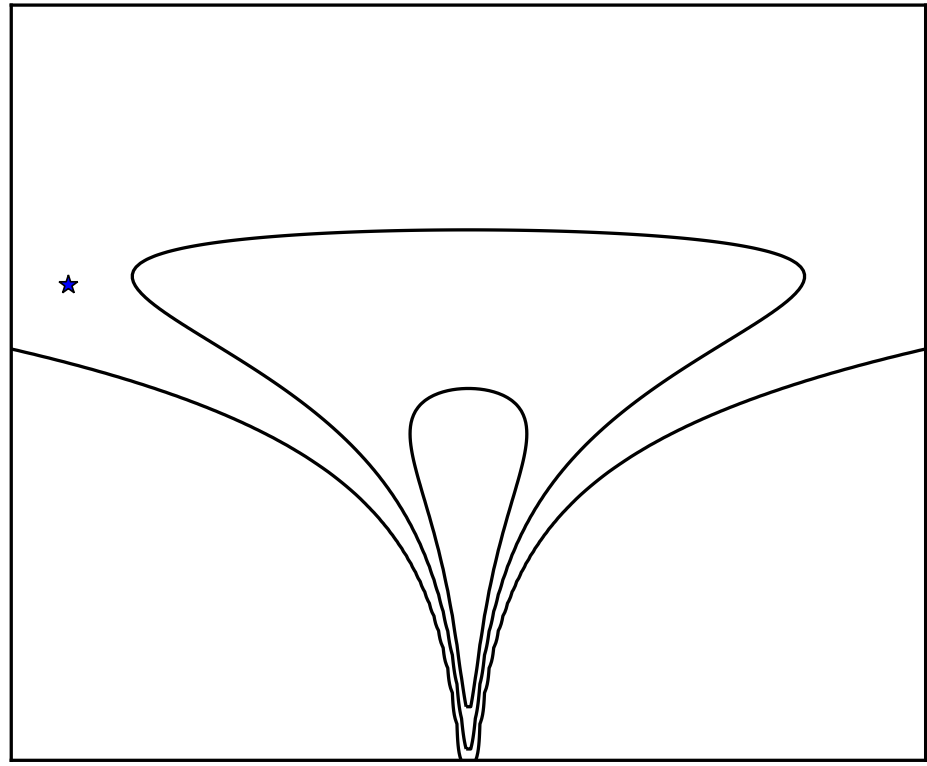
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



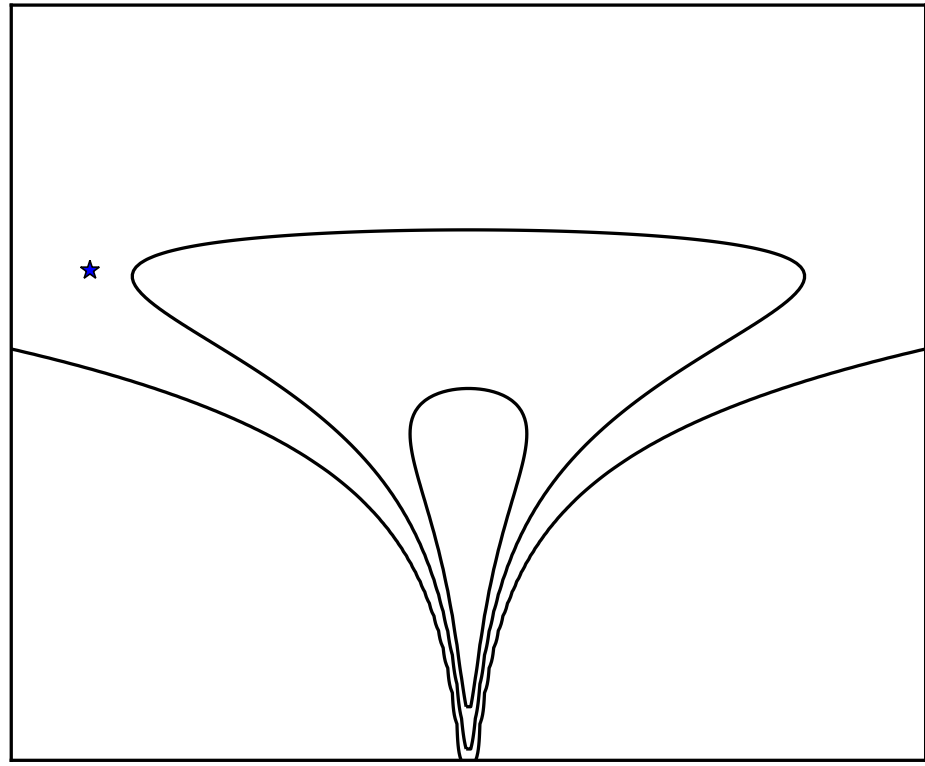
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



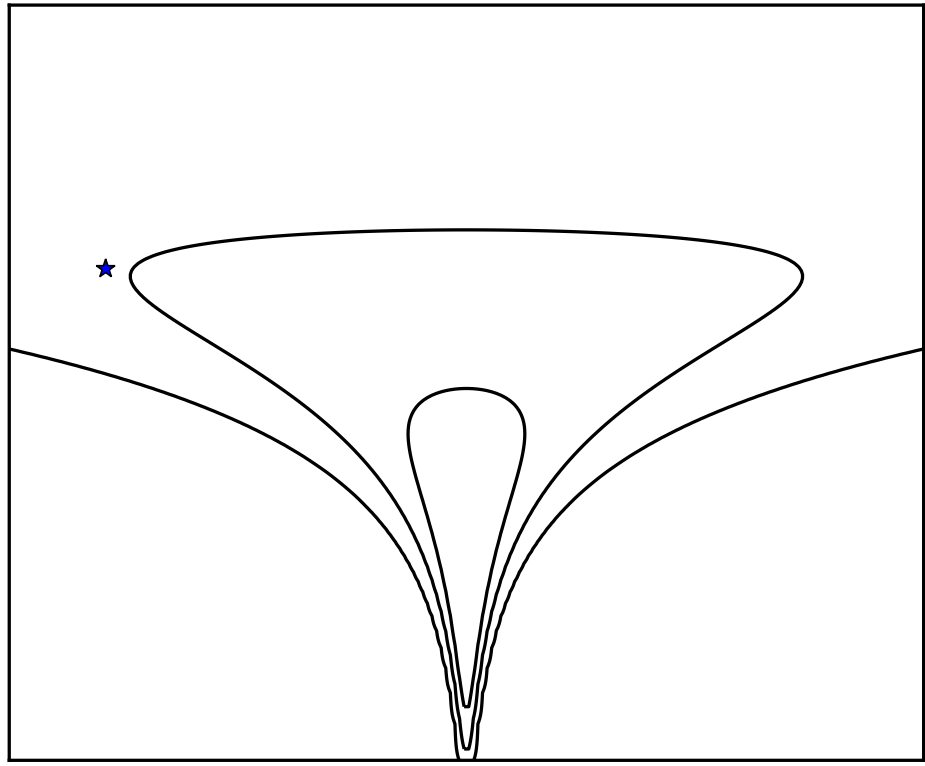
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



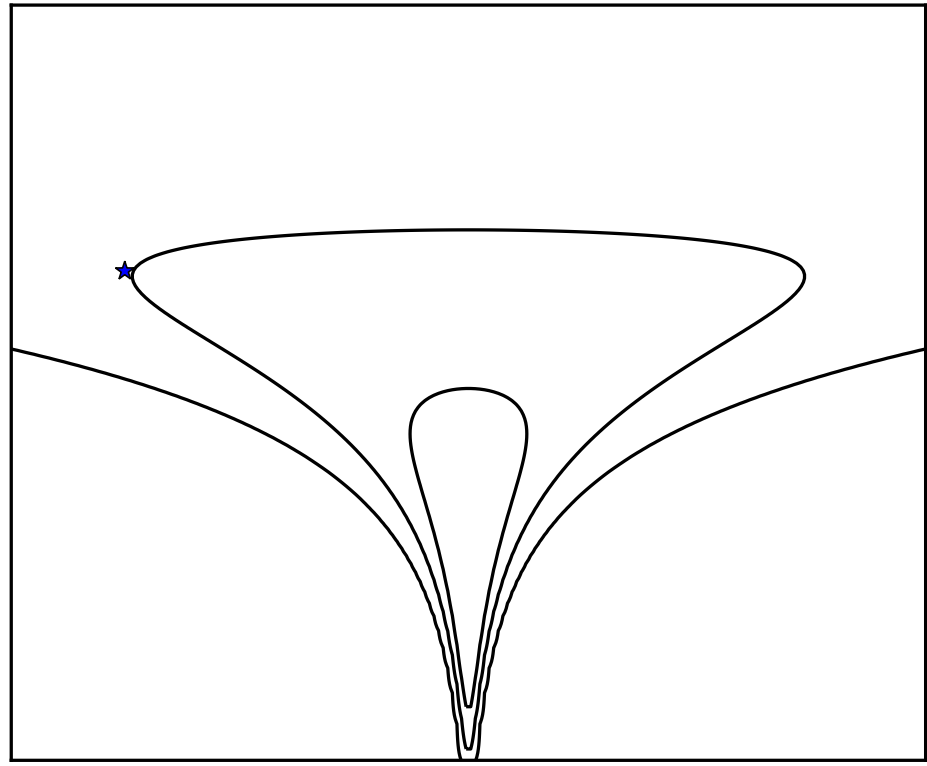
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



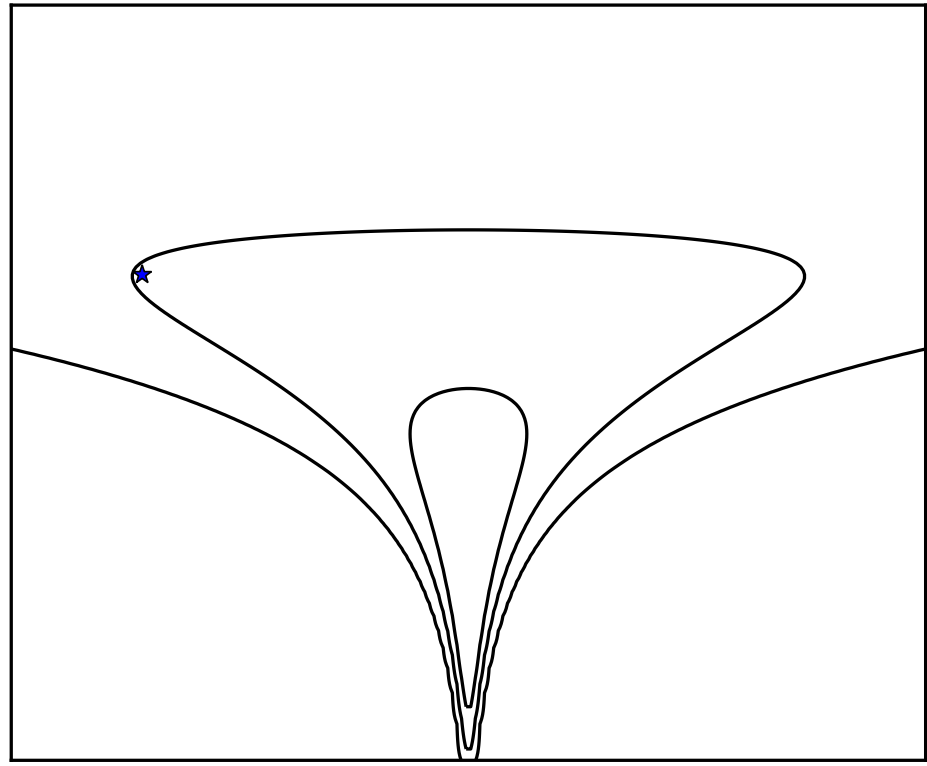
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



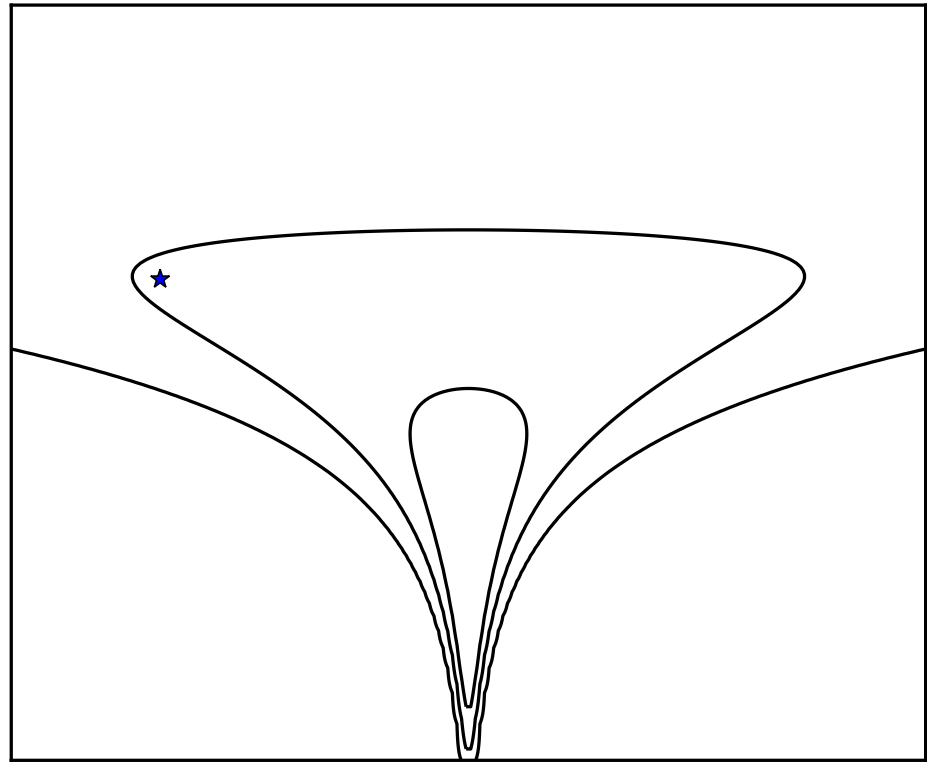
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



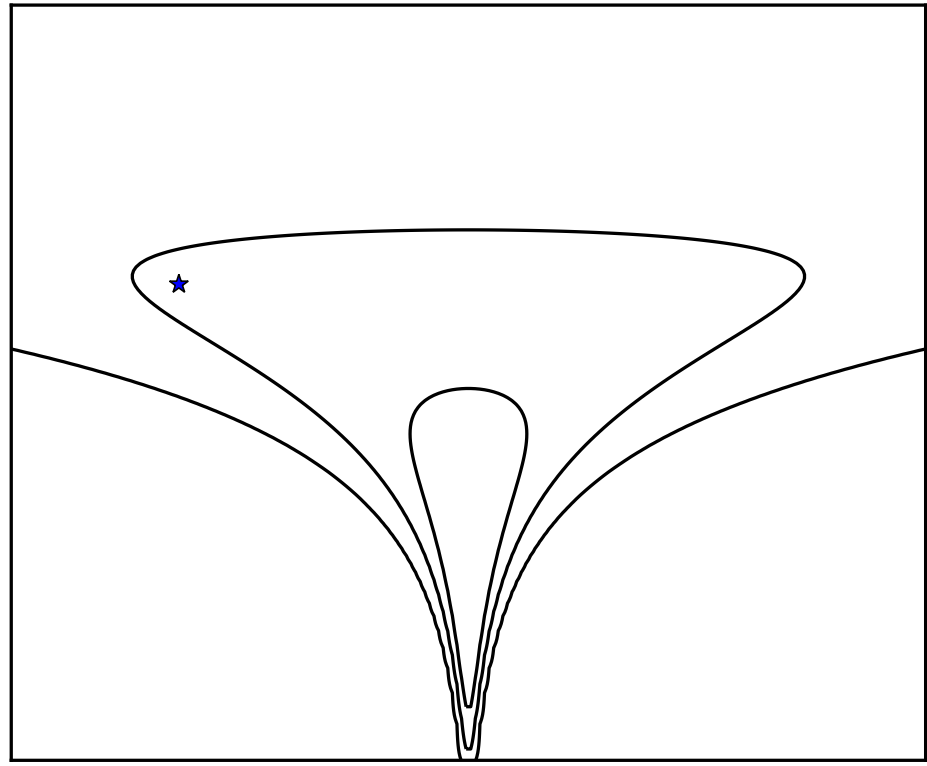
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



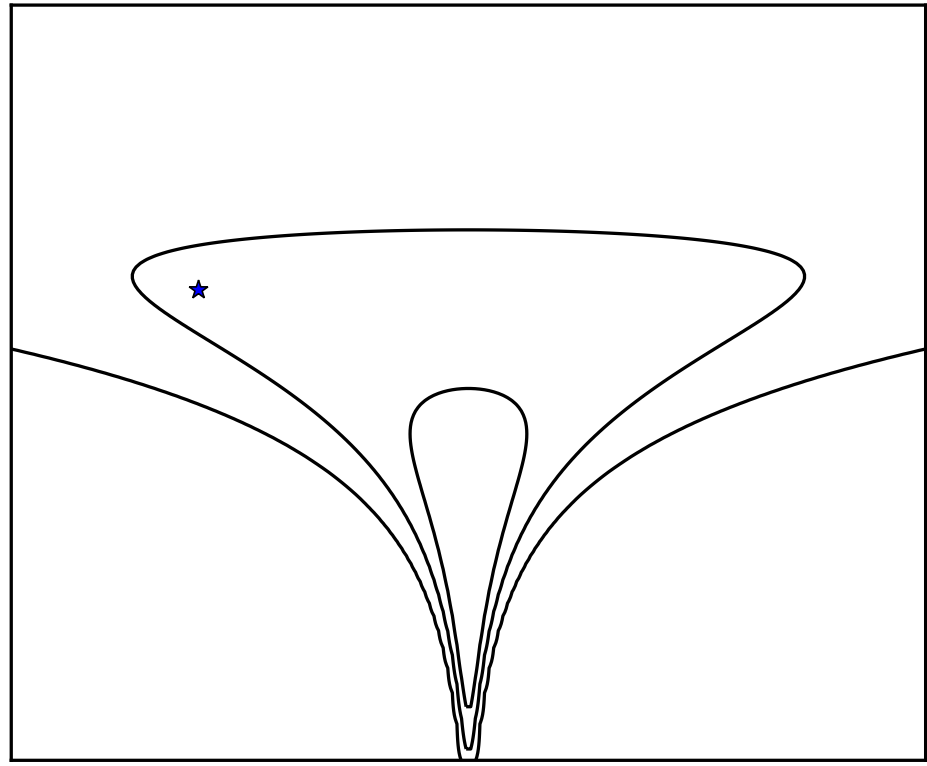
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



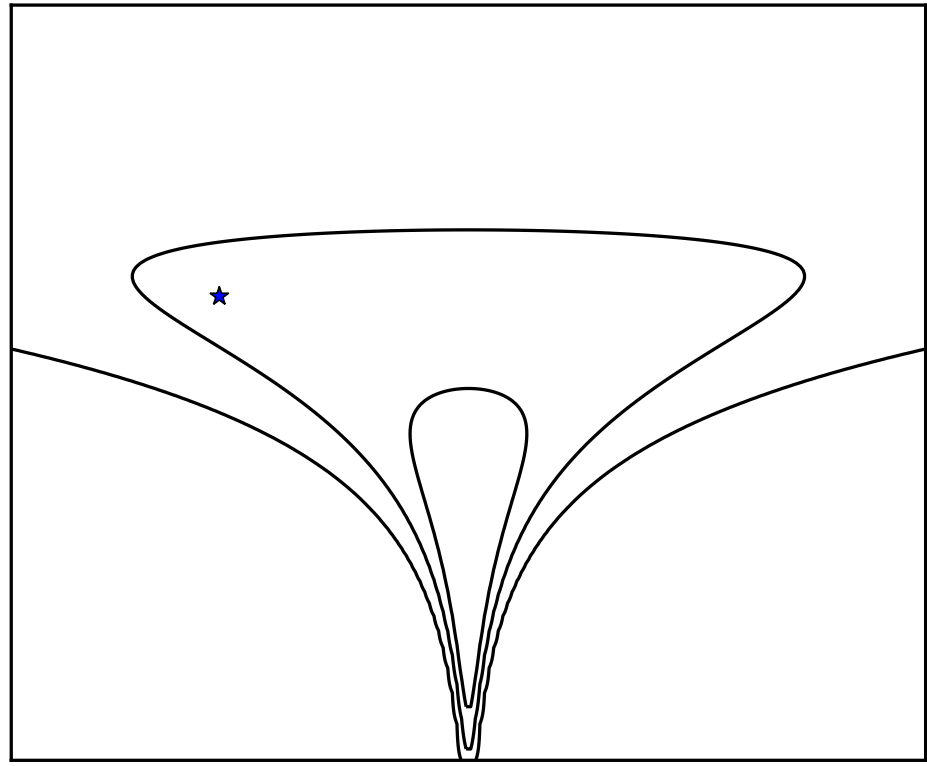
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



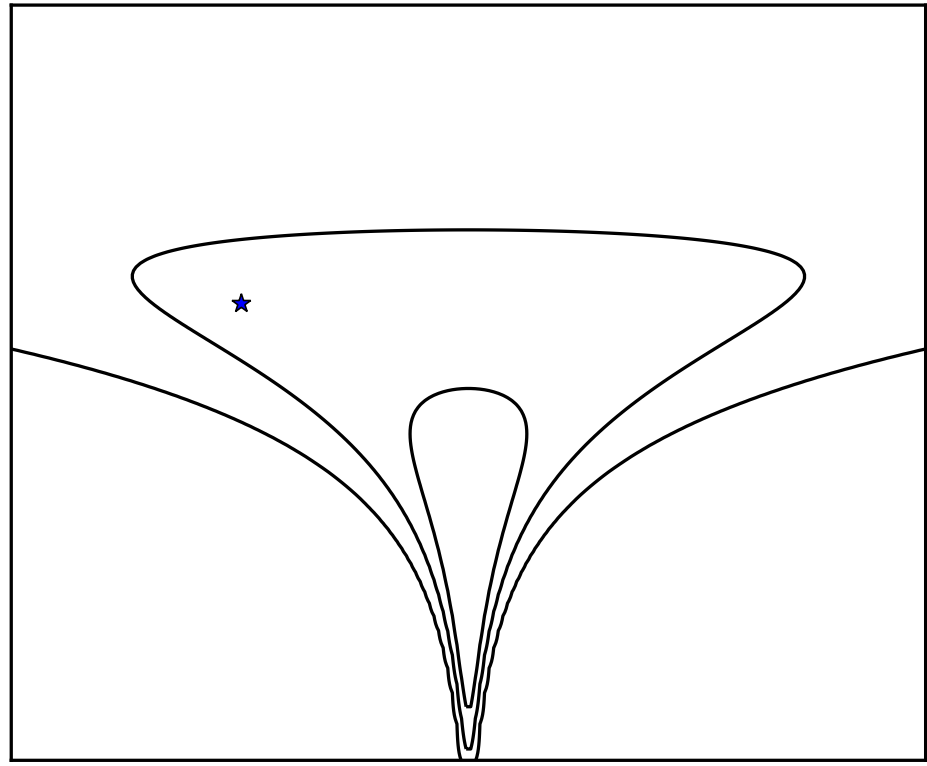
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



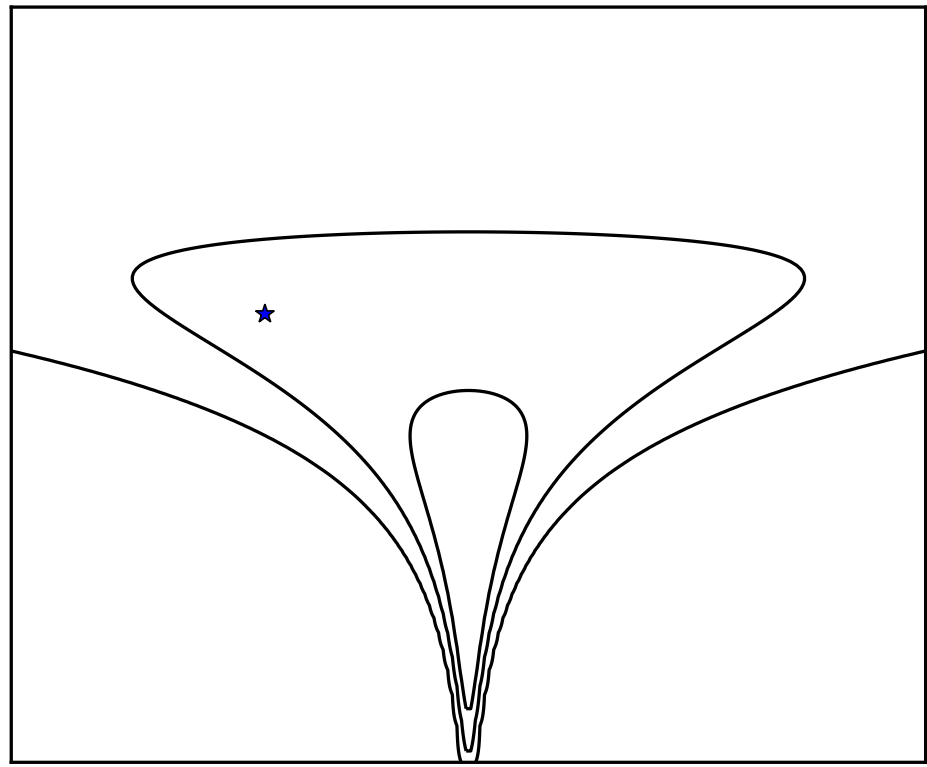
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



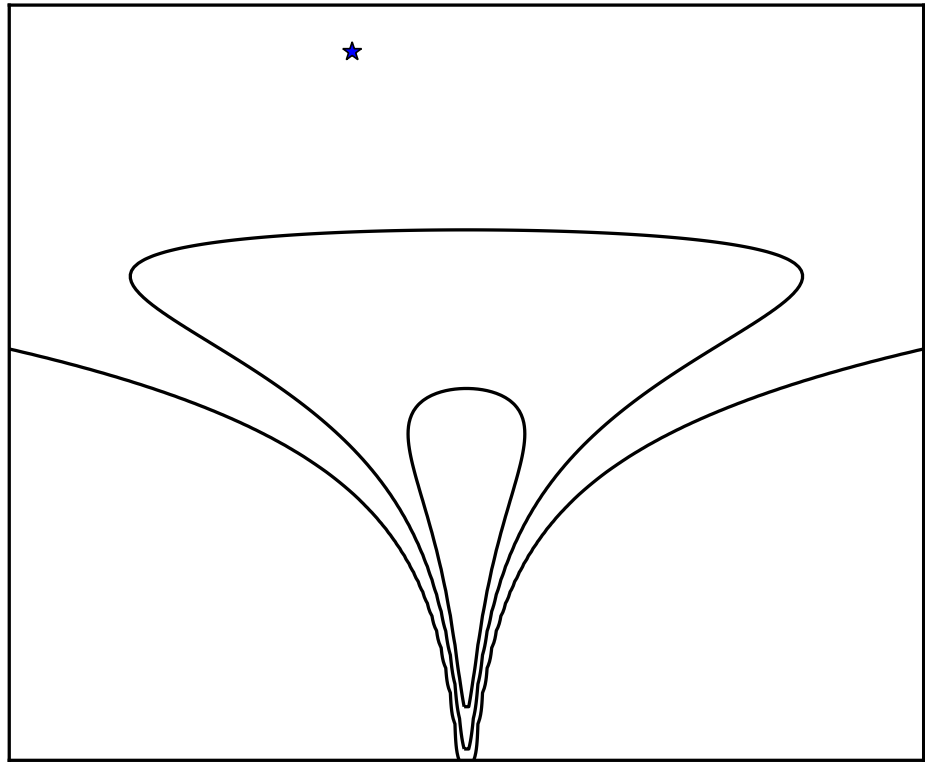
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



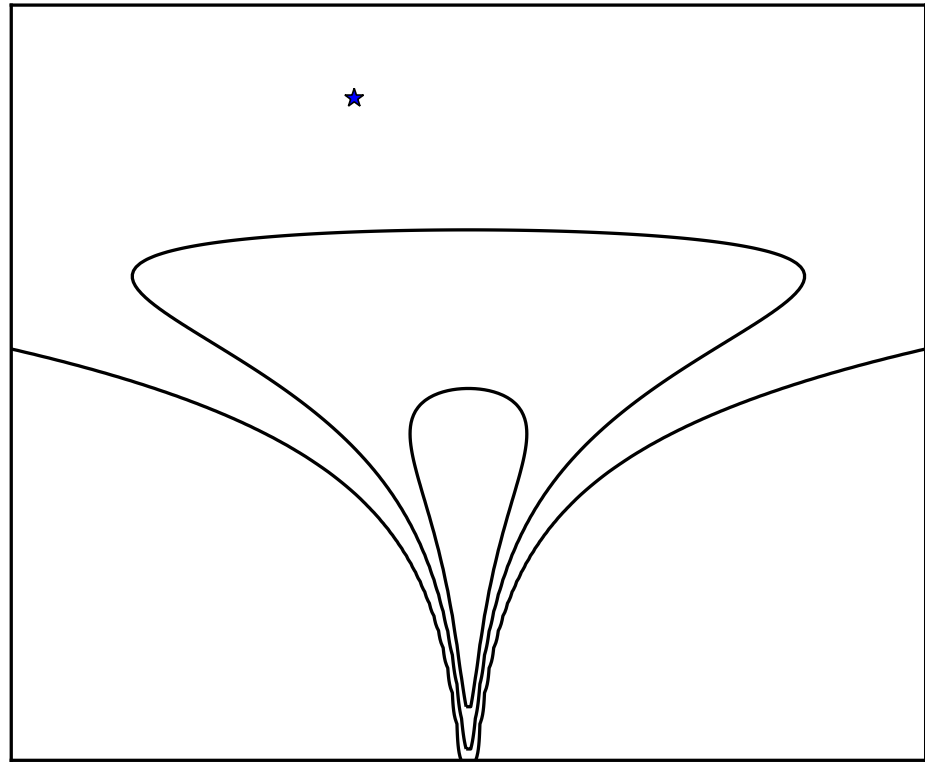
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



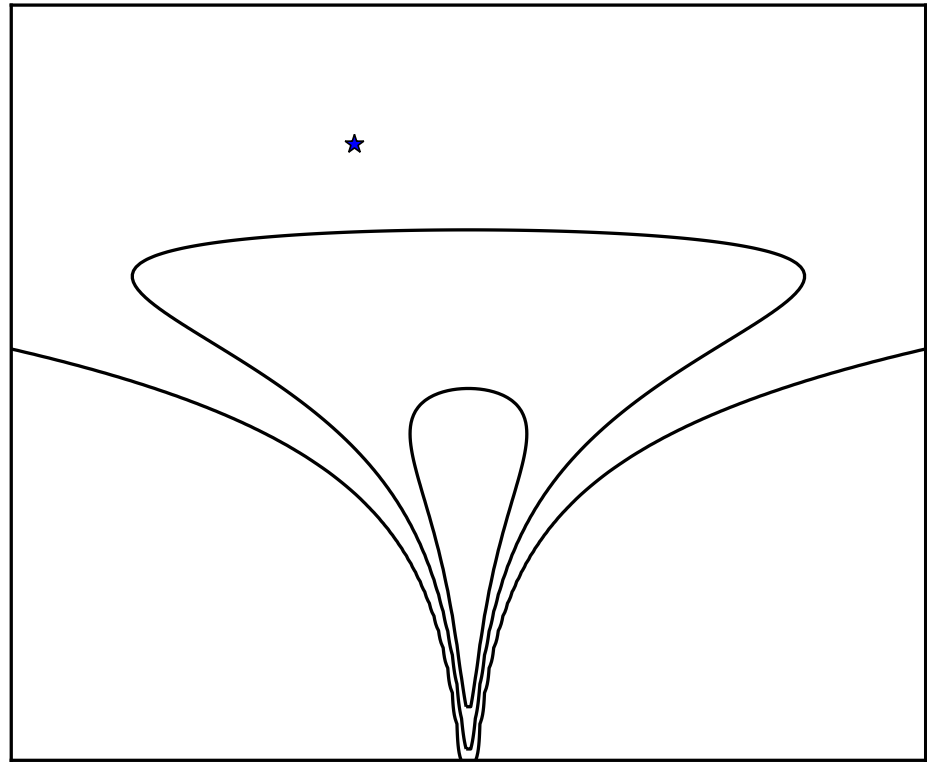
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



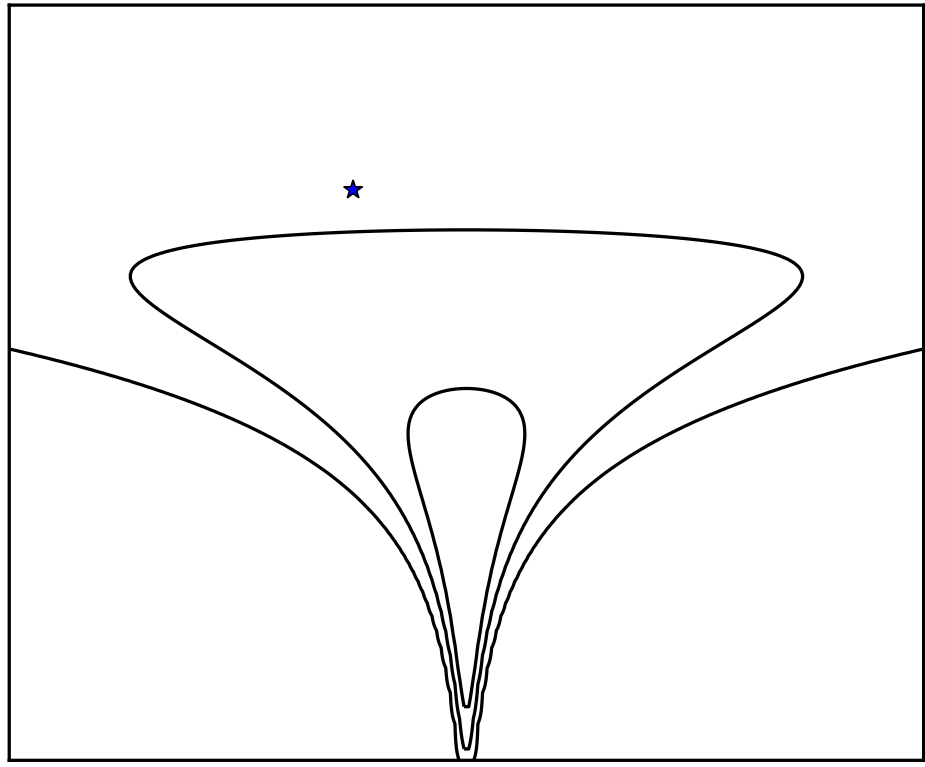
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



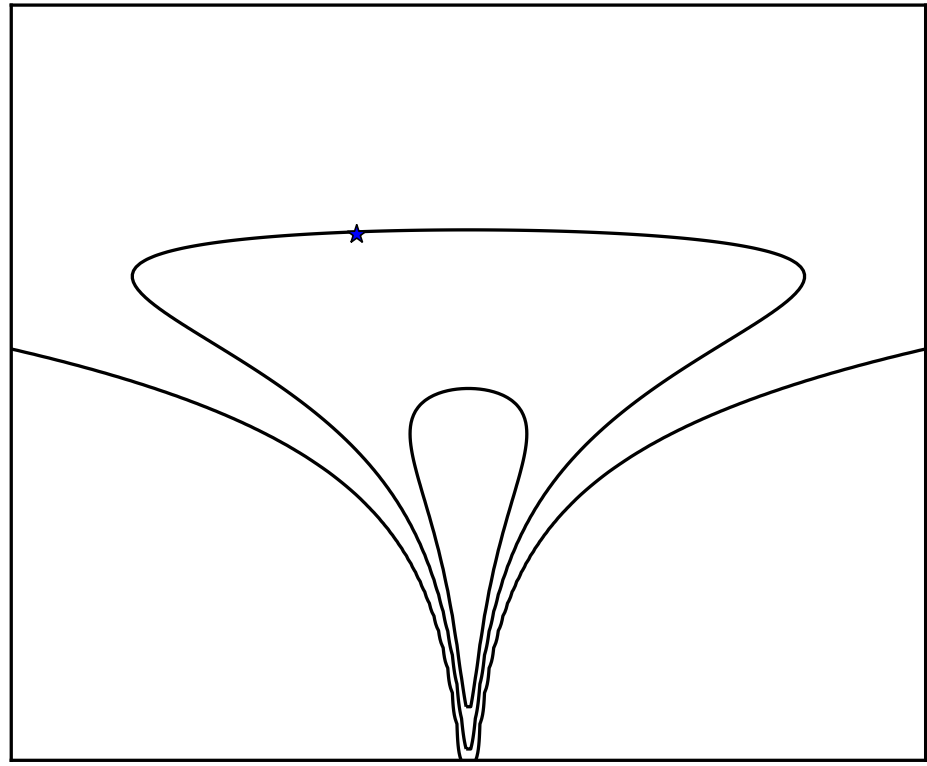
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



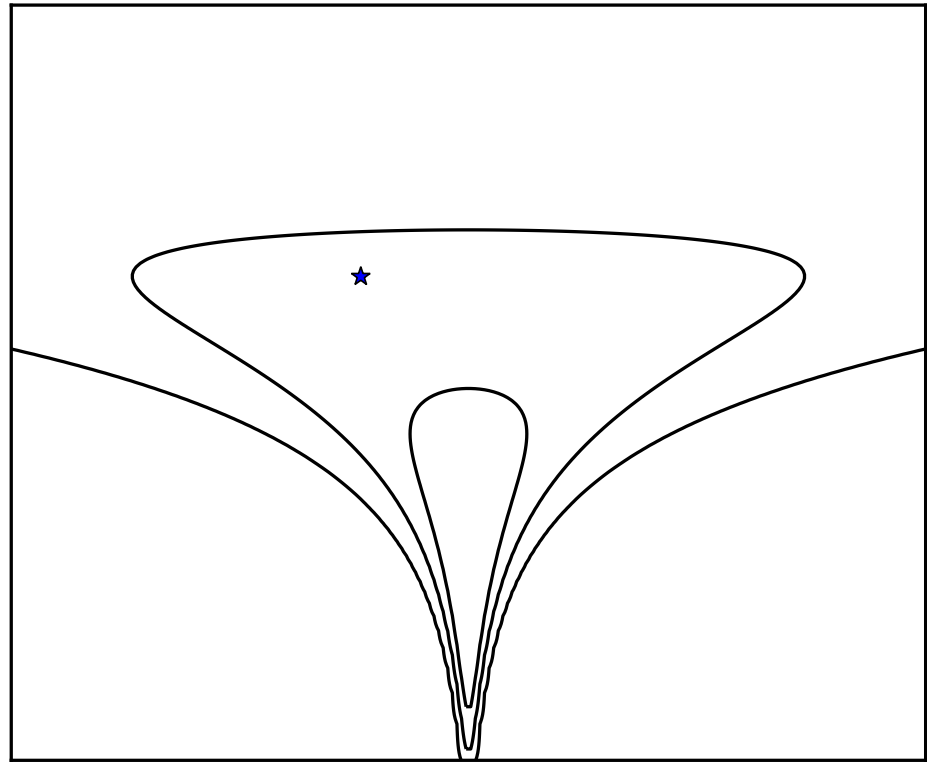
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



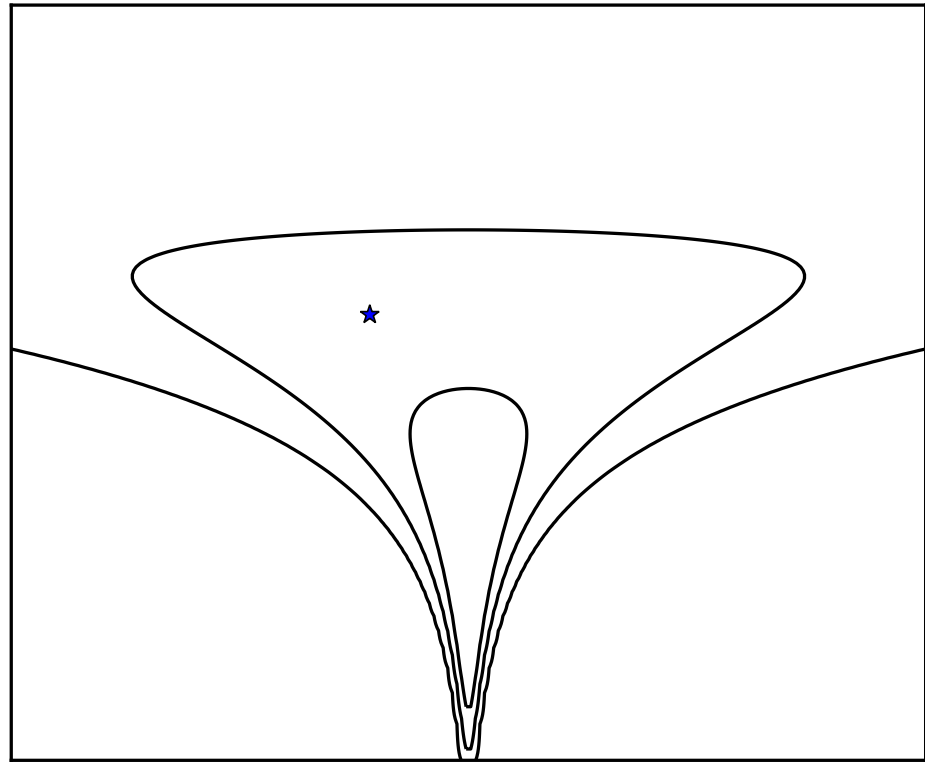
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



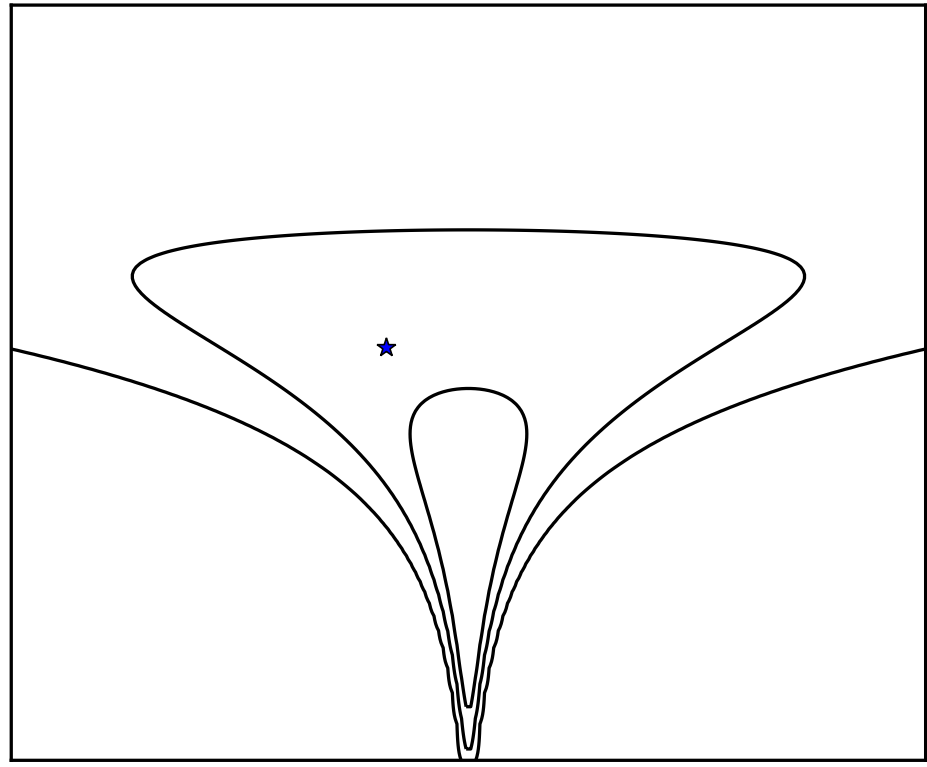
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



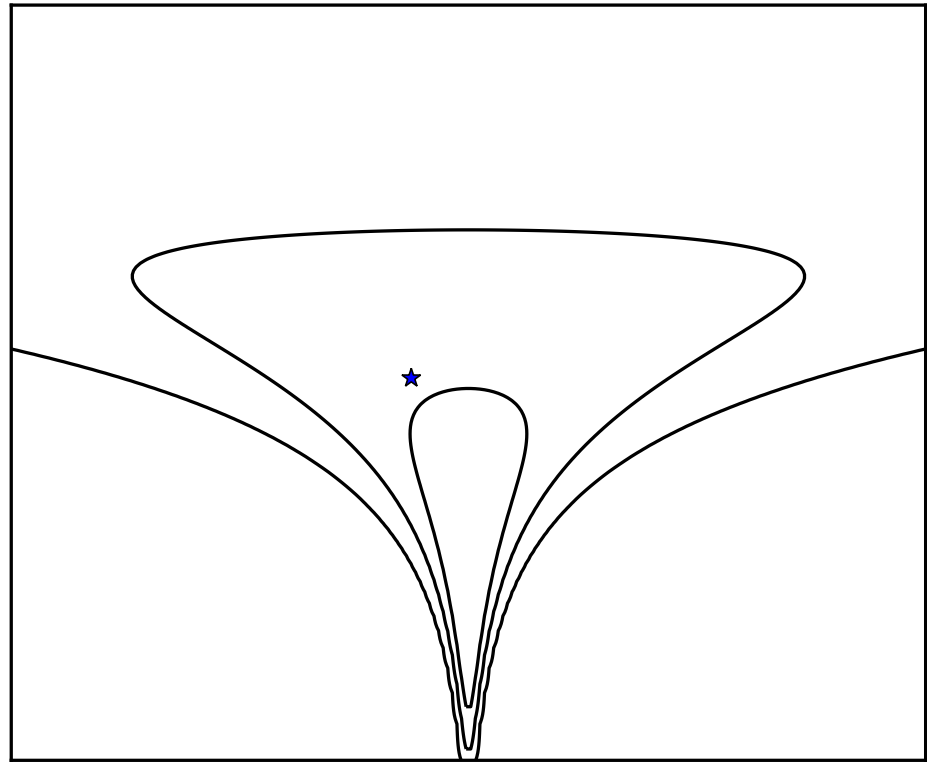
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



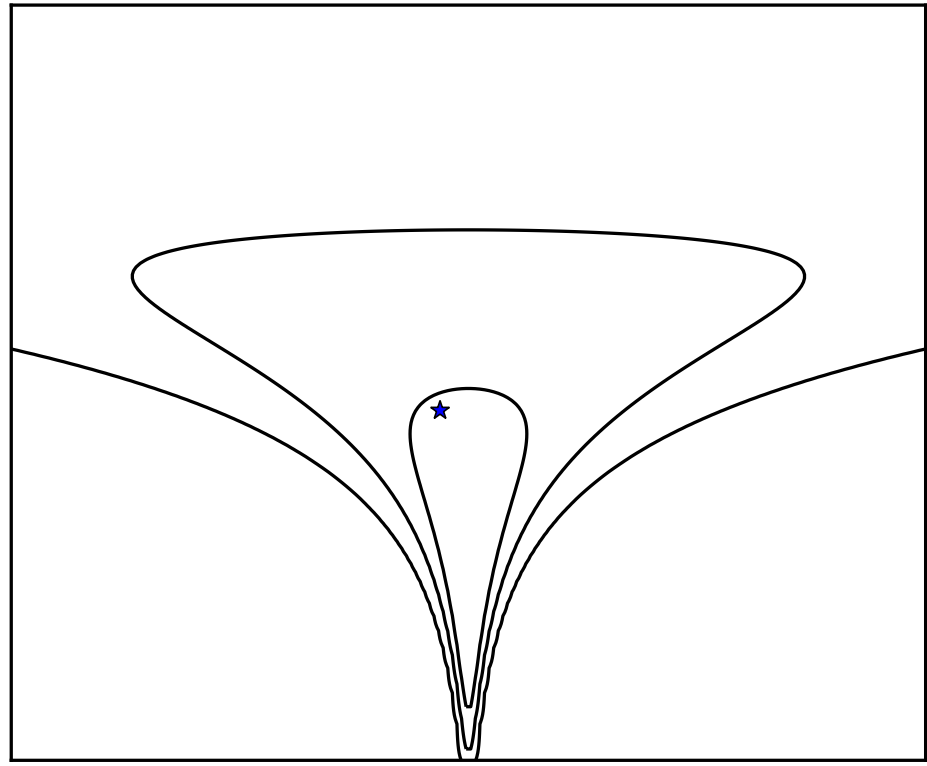
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



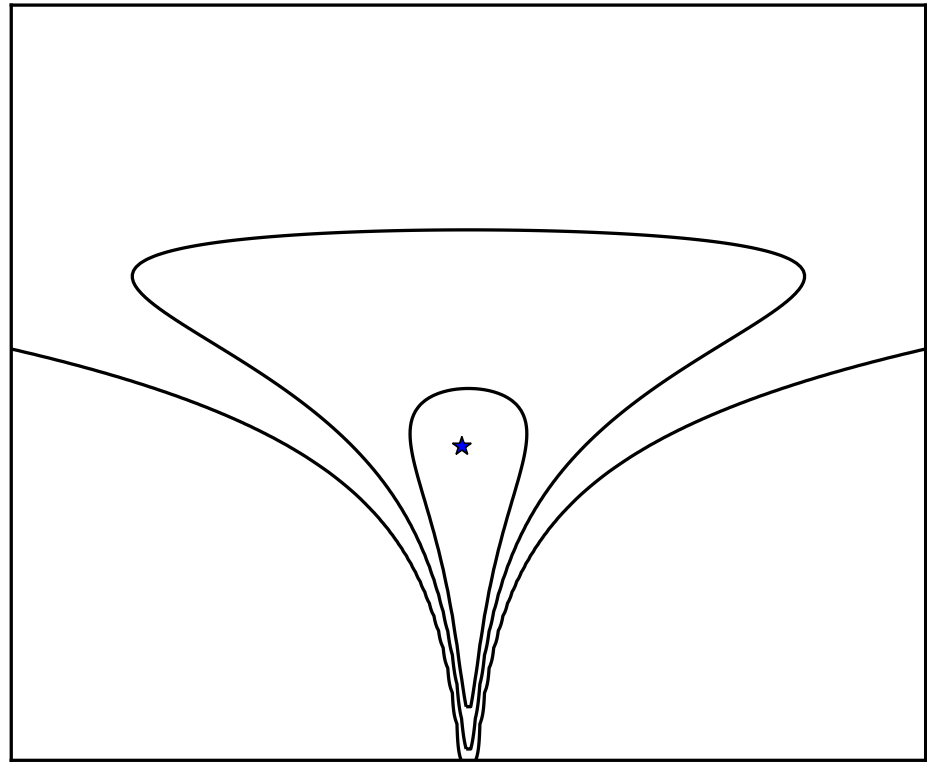
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



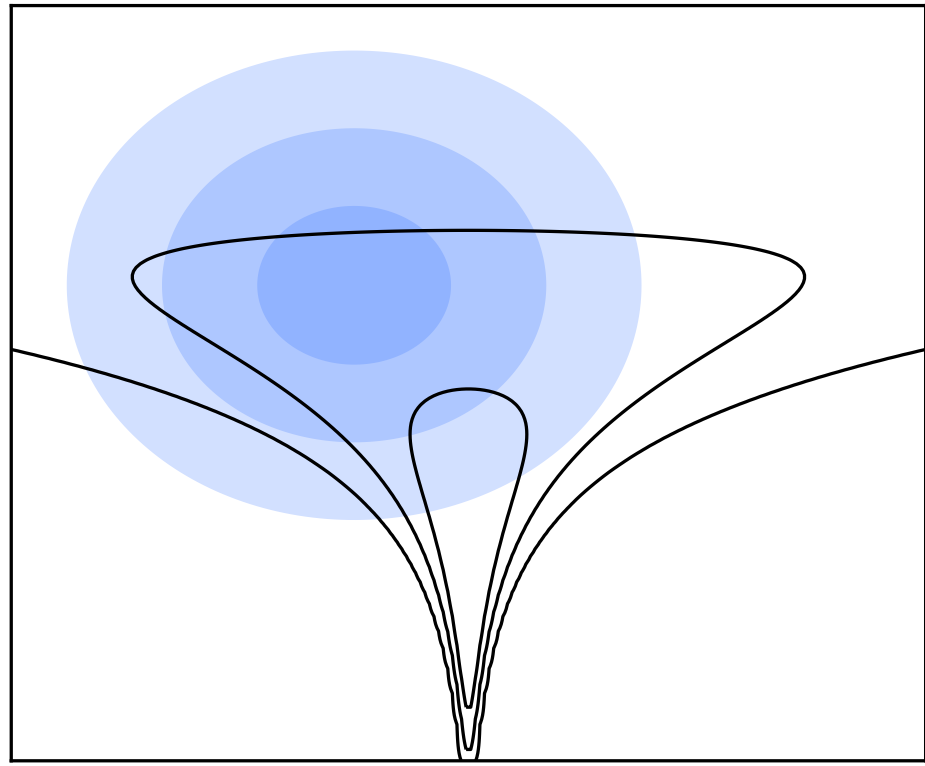
Optimization as Inference

- Optimization paths start from random init, and converges to modes...



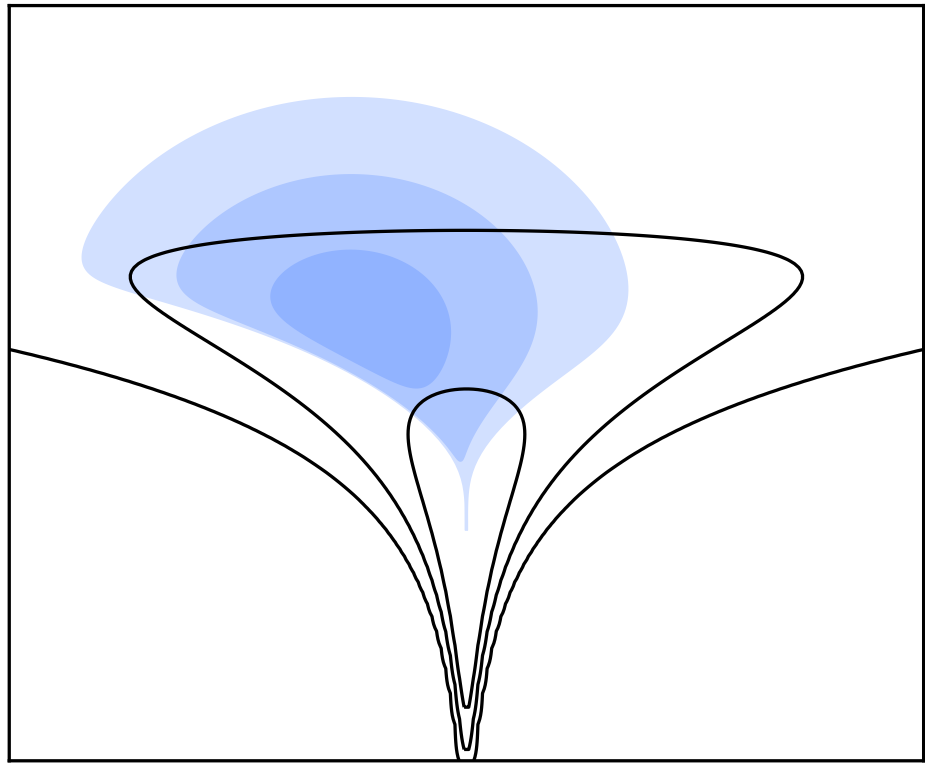
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



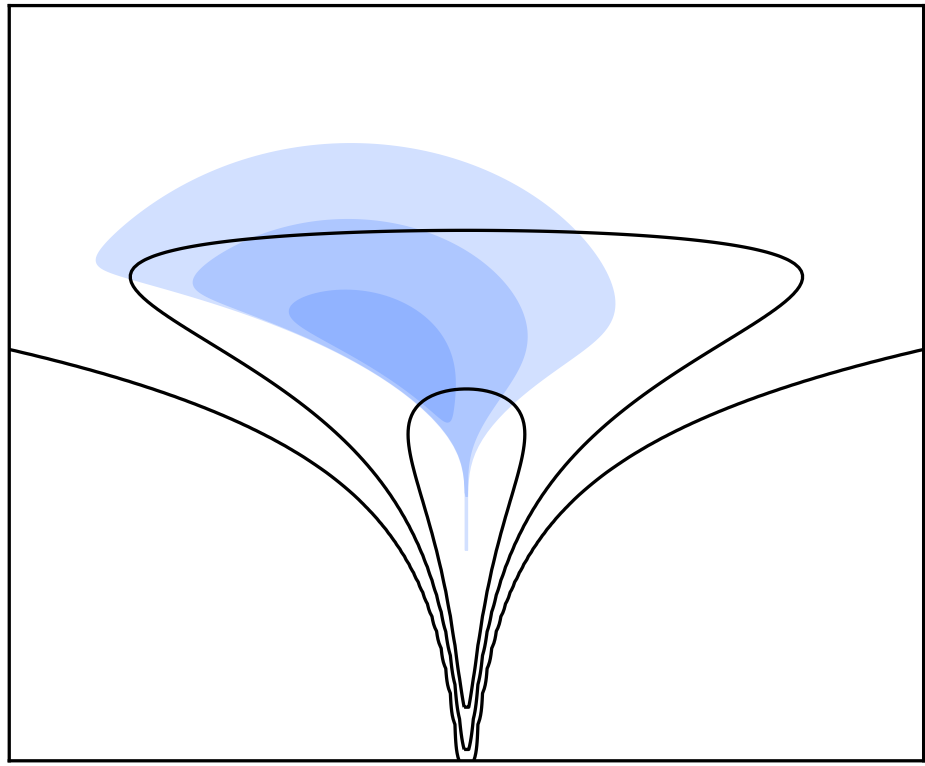
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



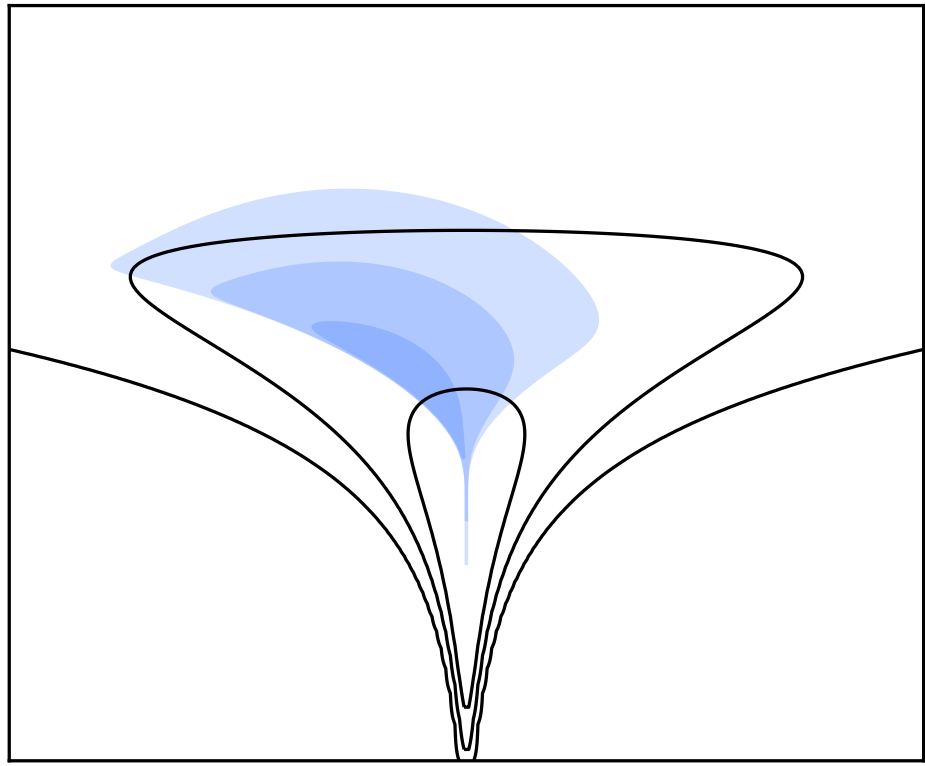
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



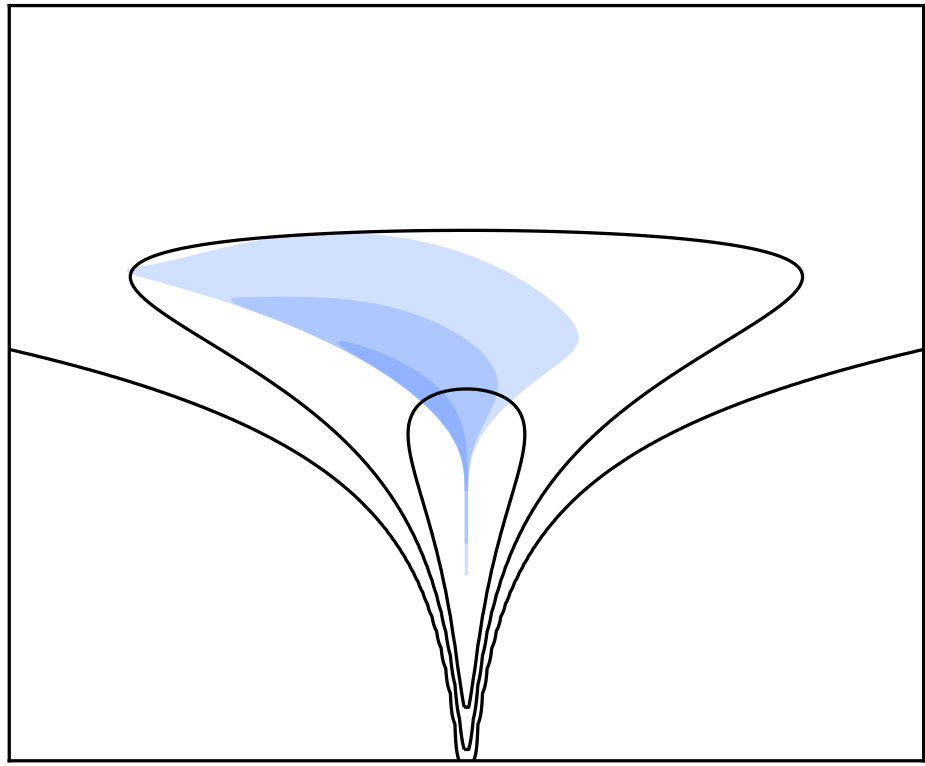
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



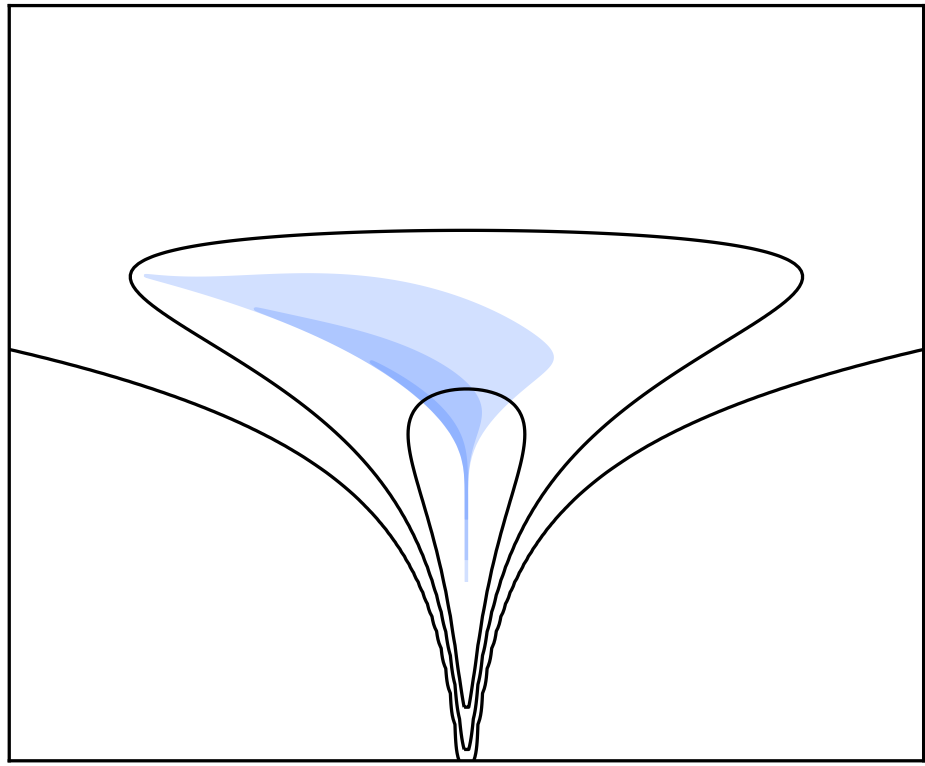
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



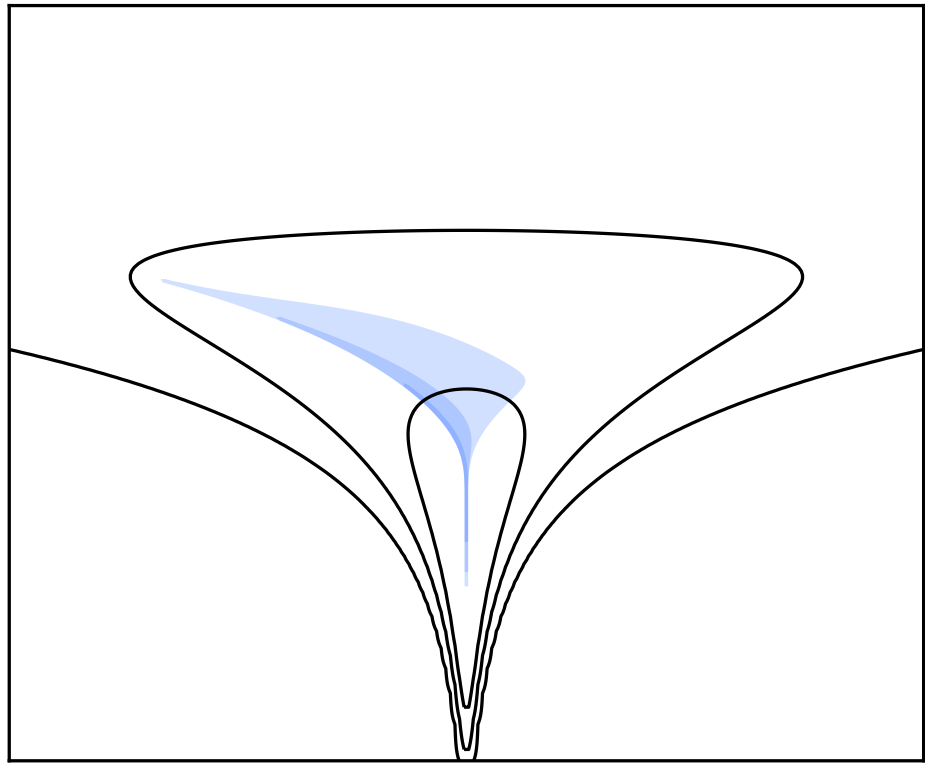
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



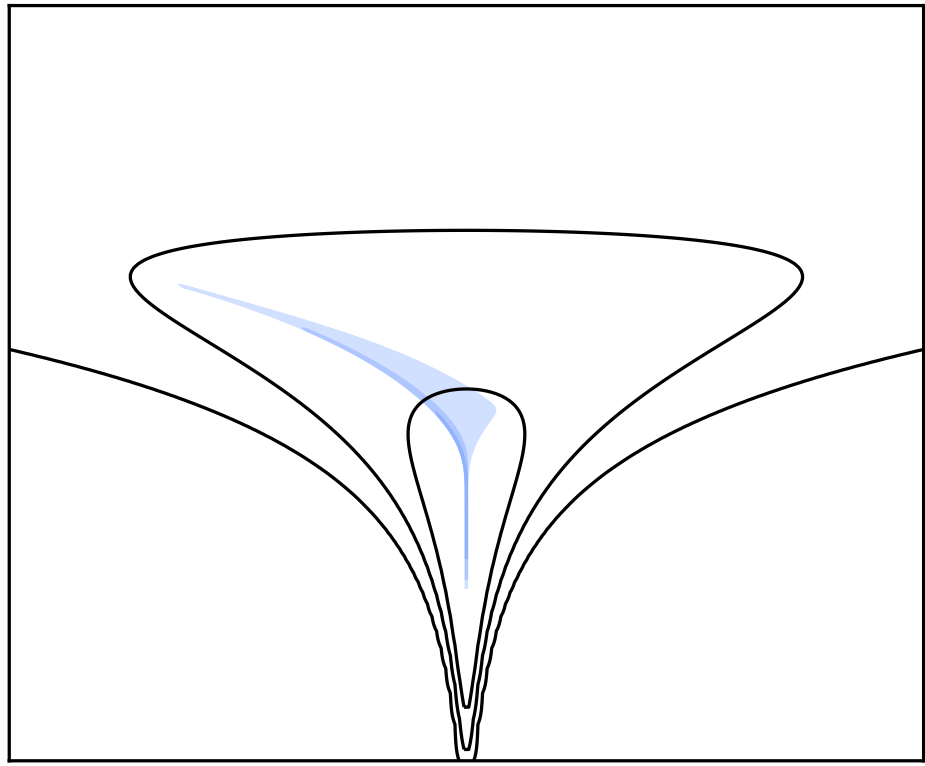
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



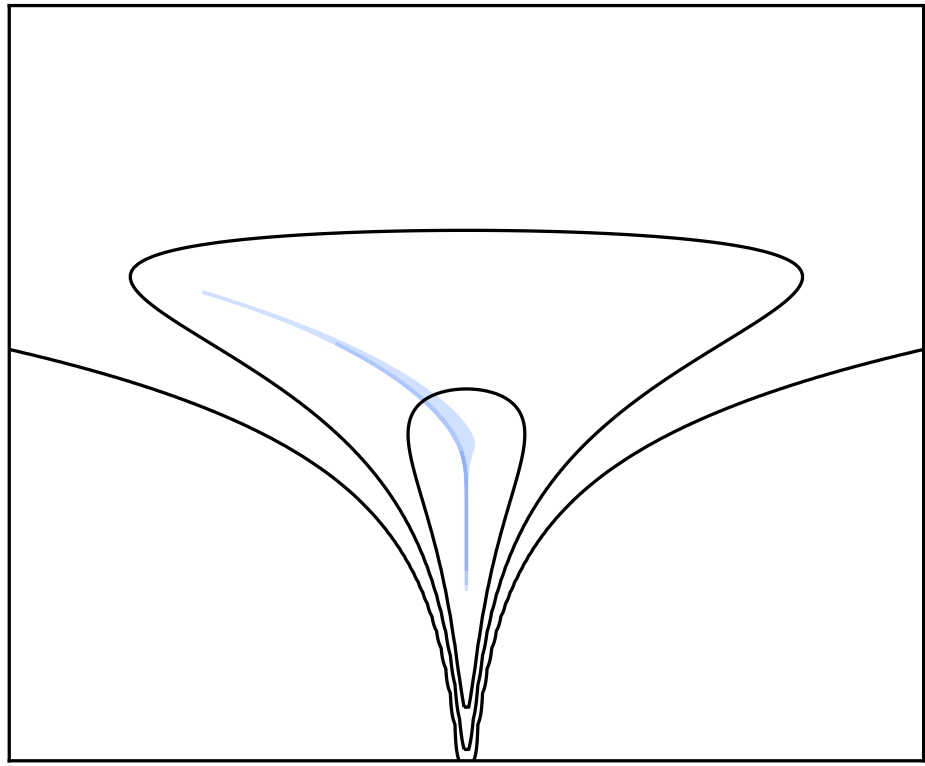
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



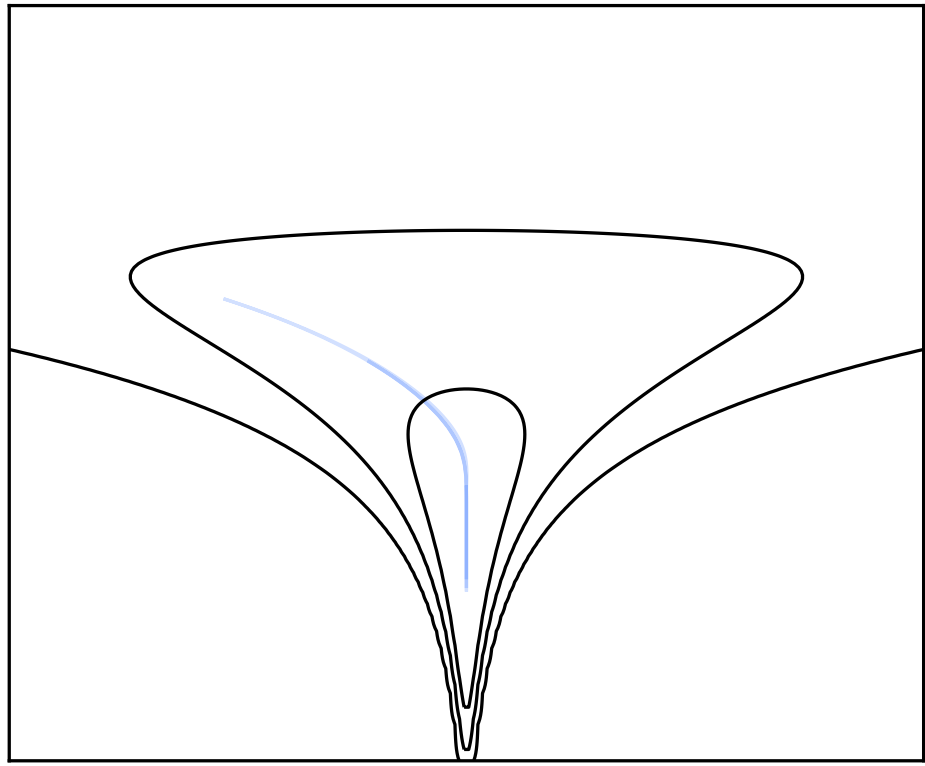
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



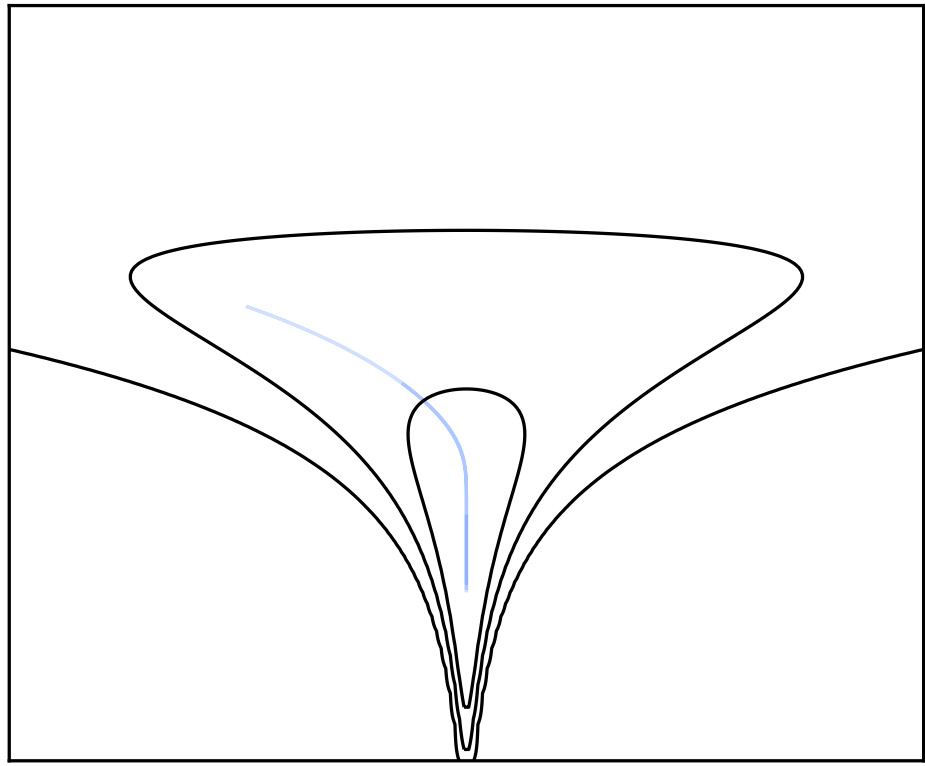
Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling



Main Idea

- What about the implicit distribution of parameters after optimizing for t steps?
- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Choosing best intermediate dist $=$ early stopping
- Taking multiple samples from dist $=$ ensembling

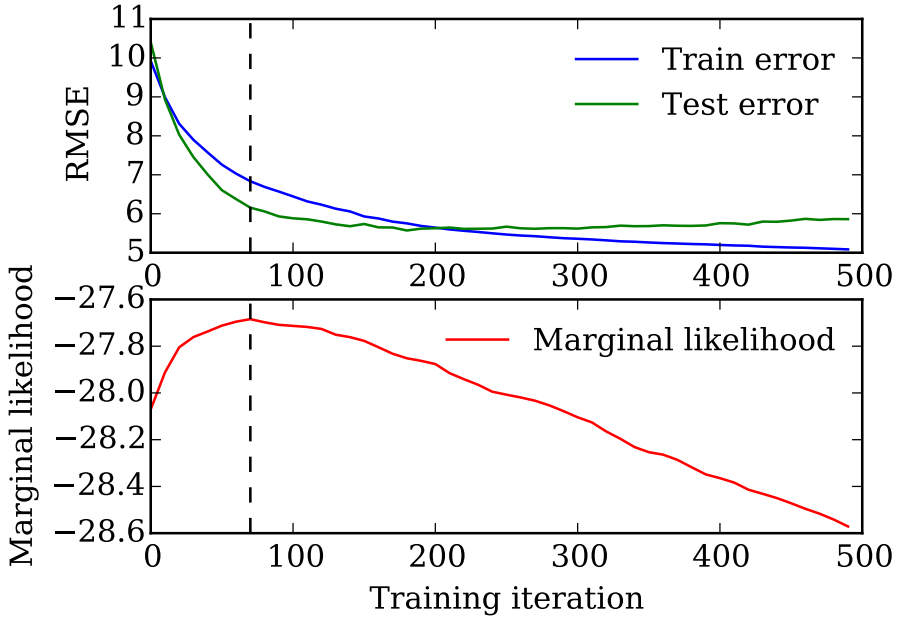


Cross Validation vs Marginal Likelihood

- What if we could evaluate marginal likelihood of implicit distribution?
- Currently, hyperparameters chosen by cross-validation
- Could choose model and learning hypers to maximize marginal likelihood
- No need for validation set

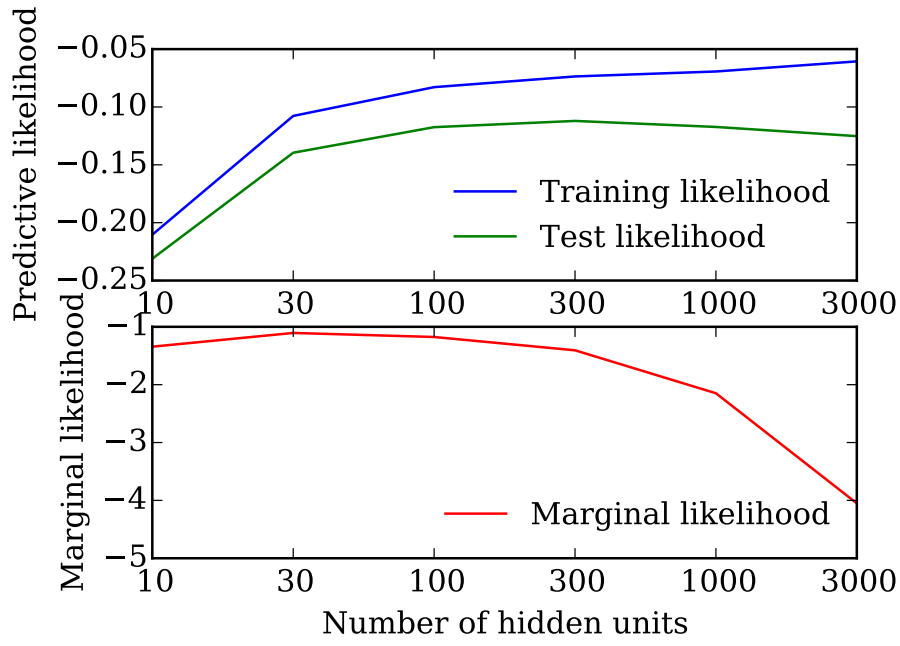
Experiments

- Early stopping
- Top: Training and test-set error on the Boston housing dataset.
- Bottom: Stochastic gradient descent marginal likelihood estimates.



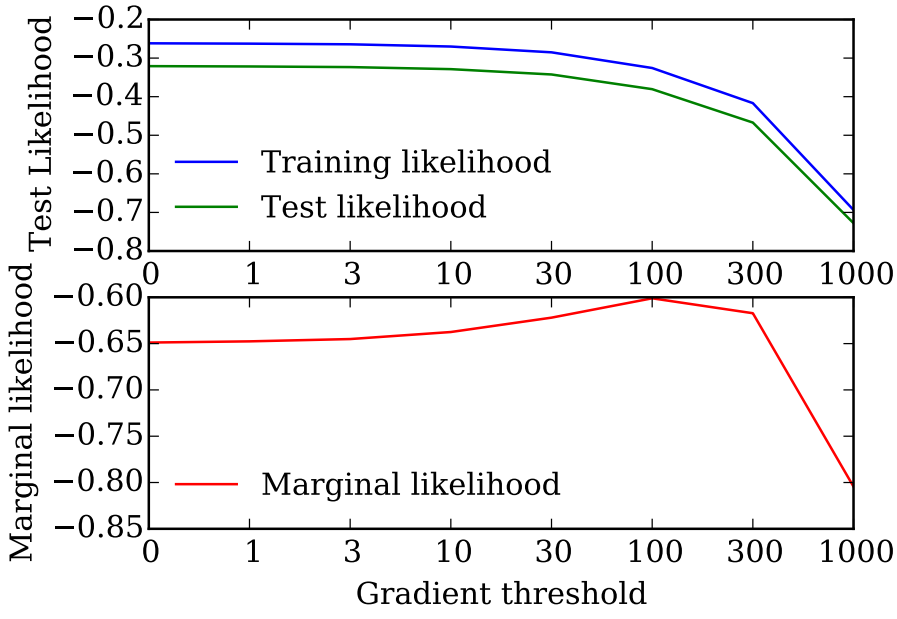
Experiments

- Choosing the number of hidden units
- Top: Training and test-set likelihood as a function of the number of hidden units in the first layer of a neural network.
- Bottom: Stochastic gradient descent marginal likelihood estimates.
- Side result: Looks like inter-sample variance is low, surprisngly!



Experiments

- Choosing the number of hidden units
- Top: Training and test-set likelihood as a function of the gradient threshold.
- Marginal likelihood as a function of the gradient threshold. A gradient threshold of zero corresponds to standard SGD.



Main Takeaways

- Optimization with random starts implies nonparametric intermediate distributions
- Can estimate lower bound on model evidence during optimization using minibatches