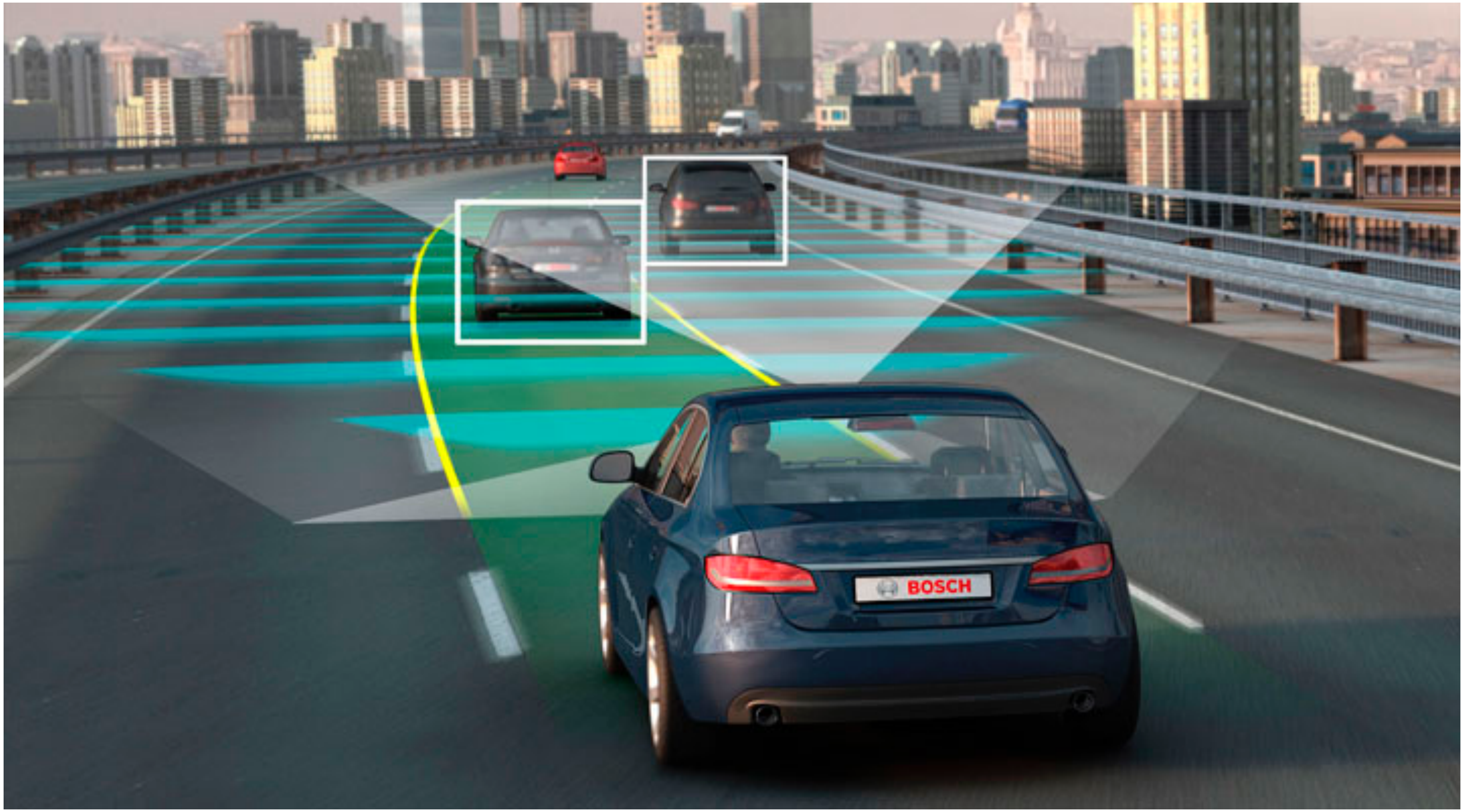
Approximate Inference for Frequentist Uncertainty Estimation

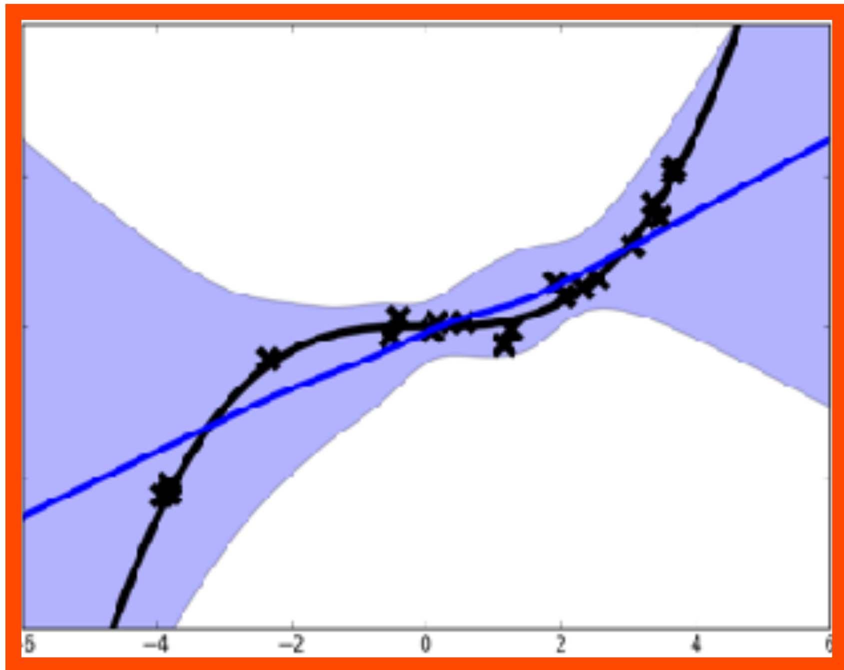
Eric Nalisnick
University of California, Irvine

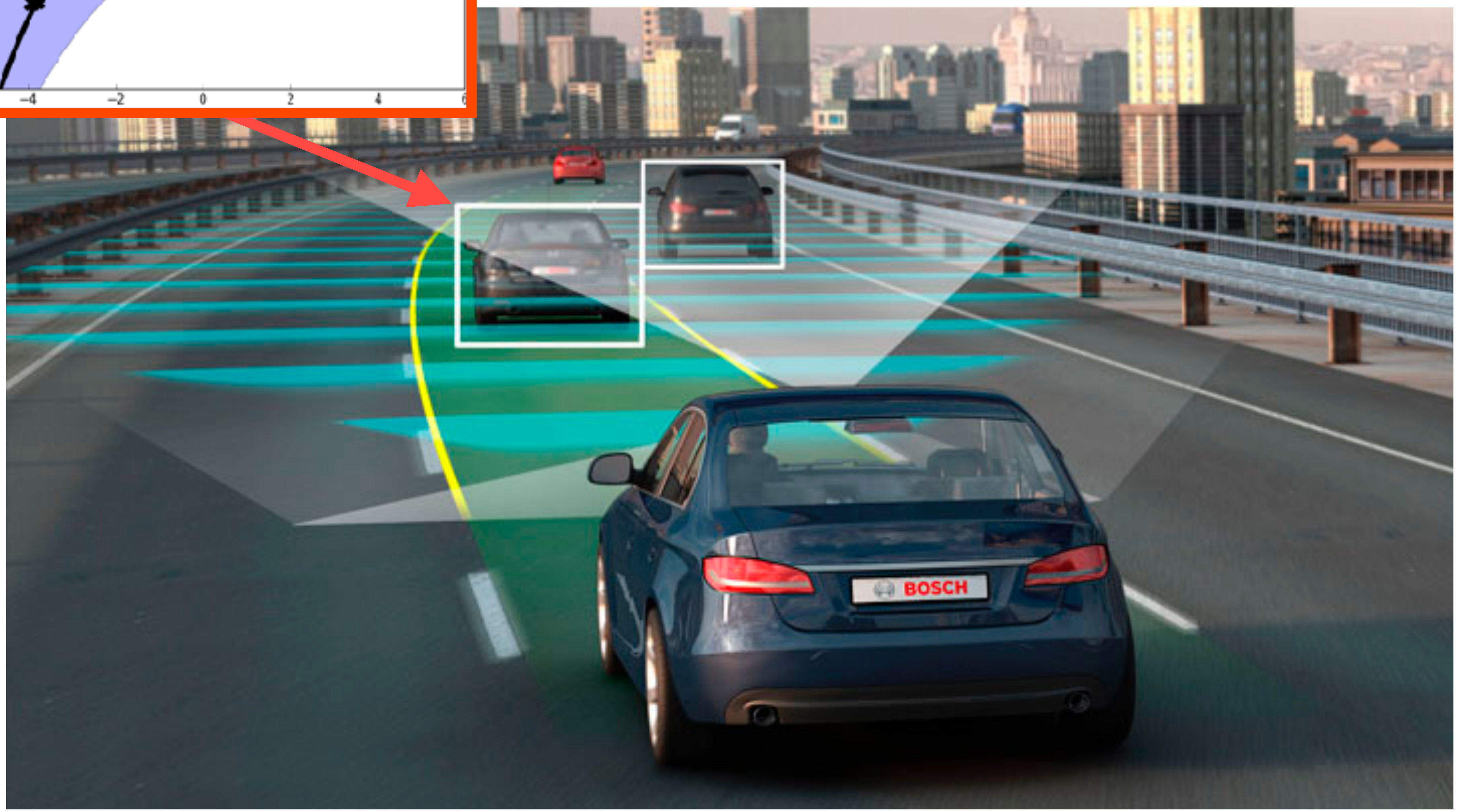
Why should we care about uncertainty estimation?



https://www.rac.co.uk/drive/features/will-self-driving-cars-mean-we-wont-need-car-insurance-anymore/

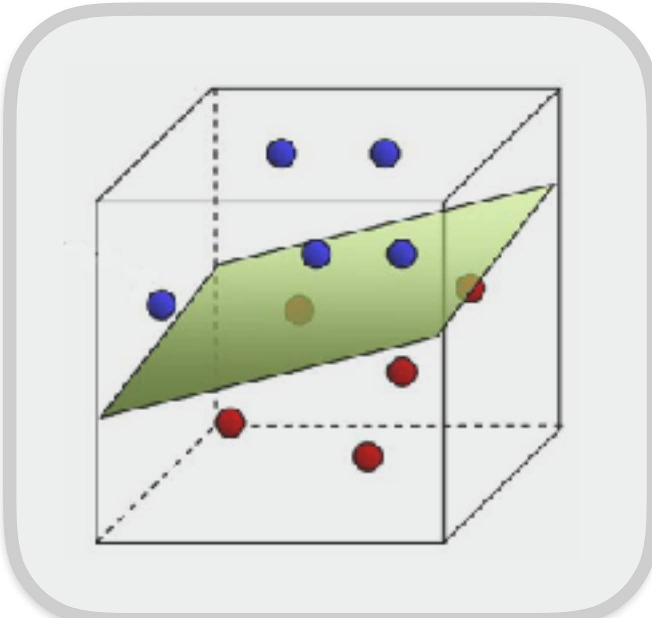
should we care about ertainty estimation?





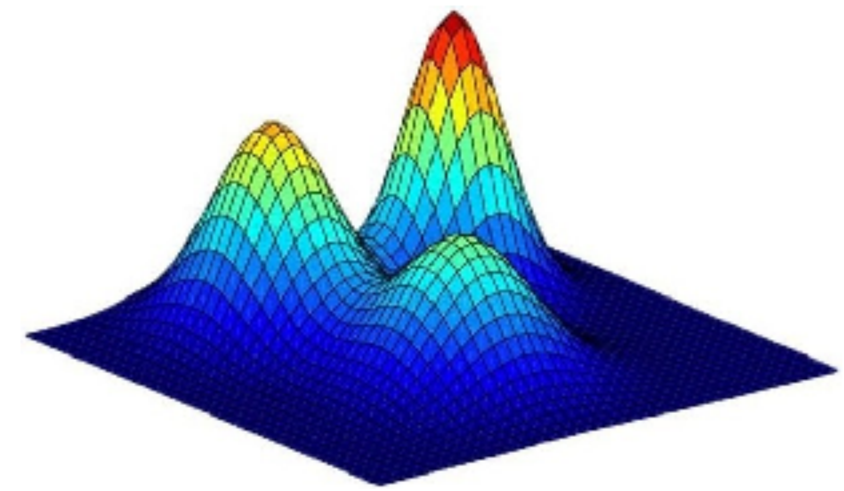
https://www.rac.co.uk/drive/features/will-self-driving-cars-mean-we-wont-need-car-insurance-anymore/

ML MODEL



INFERENCE

$\longrightarrow$



$$p(\mathcal{D}|\boldsymbol{\theta}) \qquad\qquad \pi(\boldsymbol{\theta})$$

PREDICTIVE DISTRIBUTION

$$p(\mathbf{x}^*|\mathcal{D}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta})\, p(\mathbf{x}^*|\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})\; p(\boldsymbol{\theta})}{p(\mathcal{D})}$$


THOMAS BAYES

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

THOMAS BAYES

Surge of progress in scalable / approximate Bayesian inference.

VARIATIONAL INFERENCE

Inference Models / Amortization
- Regression (Salimans & Knowles, 2014)
- Neural Networks (Kingma & Welling, 2014) (Rezende et al., 2014)
- Gaussian Processes (Tran et al., 2016)

Approximations via Transformation
- Normalizing Flows (Rezende & Mohamed, 2015)
- Hamiltonian Flow (Salimans et al, 2015)
- Inv. Auto-Regressive (Kingma et al., 2016)

Implicit Posterior Approximations
- Stein Particle Descent (Liu & Wang, 2016)
- Operator VI (Ranganath et al., 2016)
- Adversarial VB (Mescheder et al., 2017)

BAYESIAN NEURAL NETS

Scalable Posterior Inference
- Prob. Backpropagation (Hernández-Lobato & Adams, 2015)
- Bayes by Backprop. (Blundell et al., 2015)
- Matrix Gauss. Approx. (Louizos & Welling, 2016)

Latent Variable Models
- Variational Autoencoders (Kingma & Welling, 2014)
- Structured VAEs (Johnson et al., 2017)
- Bayesian GANs (Saatchi & Wilson, 2017)

"X as Bayesian Inference"
- Dropout as Bayesian Approx. (Gal & Ghahramani, 2016)
- Posterior Distillation (Balan et al., 2015)

What about Frequentism?

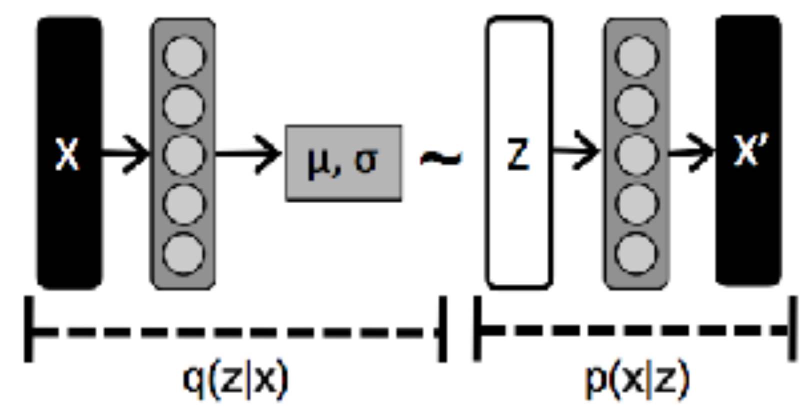
Why Be Frequentist?

No priors: choice of prior affects the marginal likelihood, if not the posterior too.

Why Be Frequentist?

No priors: choice of prior affects the marginal likelihood, if not the posterior too.

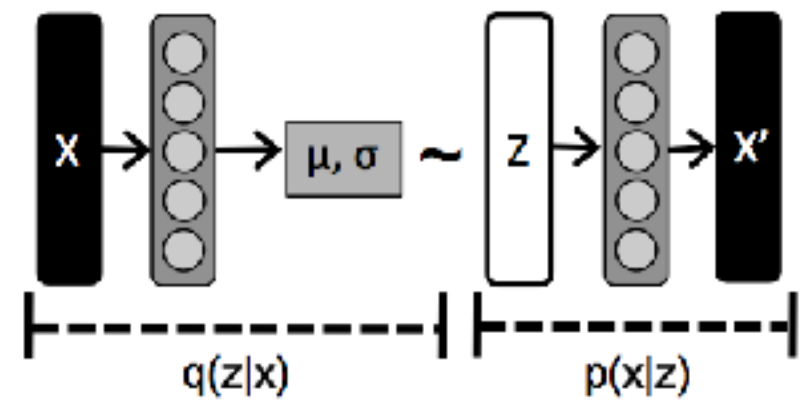
Specifically, some problems for Variational Autoencoders (VAEs)…



Why Be Frequentist?

No priors: choice of prior affects the marginal likelihood, if not the posterior too.

Specifically, some problems for Variational Autoencoders (VAEs)…



> "…to improve our variational bounds we should improve our priors and not just the encoder and decoder….perhaps we should investigate multimodal priors…"

M. Hoffman & M. Johnson. "ELBO Surgery". NIPS 2016 Workshop on *Advances in Approx. Bayesian Inference*.

Other work showing deficiencies with prior / marginal matching:

(Kingma et al., NIPS 2015), (Chen et al., ICLR 2017), (Tomczak & Welling, ArXiv 2017), (Zhao et al., ArXiv 2017)

Frequentist Methods: Two Routes

1 Knowledge of asymptotic behavior.

Maximum Likelihood: $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} \rightarrow \mathrm{N}(\boldsymbol{\theta}_0,\ \mathcal{I}(\boldsymbol{\theta}))$

'Objective' Bayesian Priors: $p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\, I(\boldsymbol{\theta}, \mathcal{D})$

Frequentist Methods: Two Routes

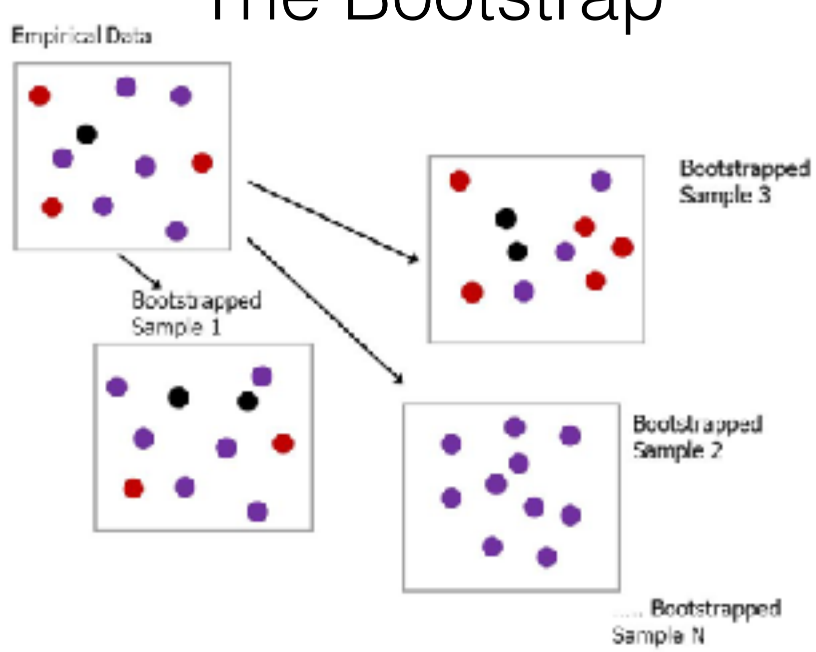
1 Knowledge of asymptotic behavior.

Maximum Likelihood: $\hat{\theta}_{\mathrm{MLE}} \rightarrow \mathrm{N}(\theta_0,\ \mathcal{I}(\theta))$

'Objective' Bayesian Priors: $p^*(\theta) = \underset{p(\theta)}{\operatorname{argmax}}\, I(\theta, \mathcal{D})$

2 Simulation of sampling process.

The Bootstrap



Other examples: jackknife, cross-validation, permutation tests, Monte Carlo tests…

Image from "Resampling Techniques" by Joel Edwards

Frequentist Methods: Two Routes

1 Knowledge of asymptotic behavior.

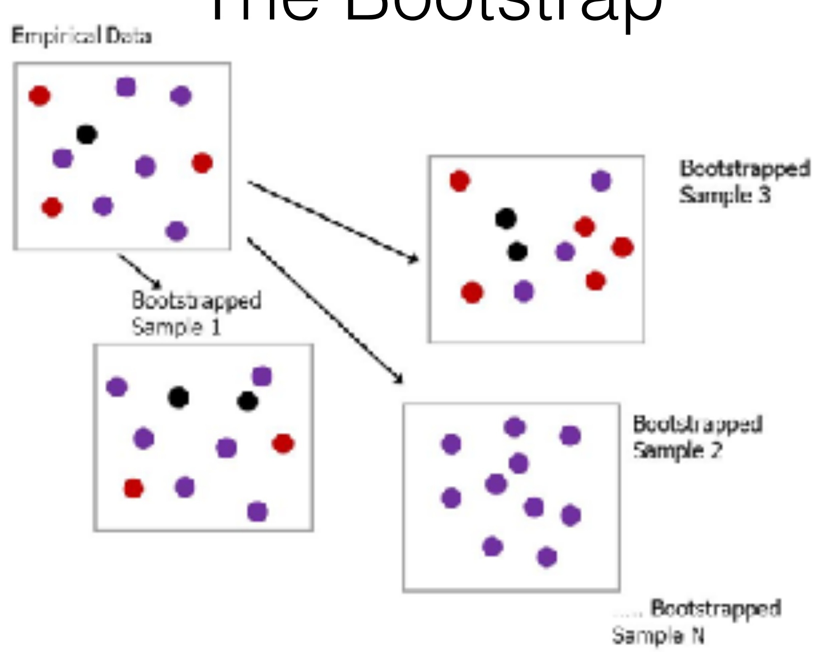
Maximum Likelihood: $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} \to \mathrm{N}(\boldsymbol{\theta}_0, \mathcal{I}(\boldsymbol{\theta}))$

PROBLEM: Analytically Intractable

'Objective' Bayesian Priors: $p^*(\boldsymbol{\theta}) = \operatorname*{argmax}_{p(\boldsymbol{\theta})} I(\boldsymbol{\theta}, \mathcal{D})$

2 Simulation of sampling process.

The Bootstrap



Other examples: jackknife, cross-validation, permutation tests, Monte Carlo tests…

Image from "Resampling Techniques" by Joel Edwards

Frequentist Methods: Two Routes

1 Knowledge of asymptotic behavior.

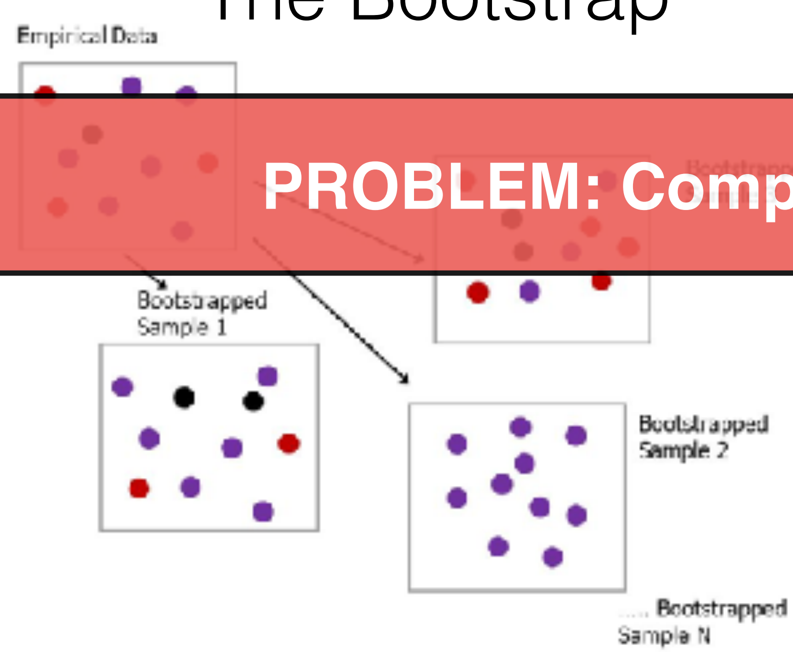
Maximum Likelihood: $\hat{\theta}_{\mathrm{MLE}} \rightarrow \mathrm{N}(\theta_0, \mathcal{I}(\theta))$

PROBLEM: Analytically Intractable

'Objective' Bayesian Priors: $p^*(\theta) = \underset{p(\theta)}{\operatorname{argmax}}\, I(\theta, \mathcal{D})$

2 Simulation of sampling process.

The Bootstrap



PROBLEM: Computationally Expensive

Other examples: jackknife, cross-validation, permutation tests, Monte Carlo tests…

Image from "Resampling Techniques" by Joel Edwards

Using Advances in Approx. Inference for Frequentism

1 Knowledge of asymptotic behavior.

CONTRIBUTION

Approximating Objective Bayesian Priors

(Nalisnick & Smyth, UAI 2017)

Use variational bound to find an approximate prior.

2 Simulation of sampling process.

CONTRIBUTION

The Amortized Bootstrap

(Nalisnick & Smyth, SoCalML 2017)

Use implicit models to approximate bootstrap distribution.

Approximating Reference Priors

(Nalisnick & Smyth, UAI 2017)

Objective Bayesian Priors

Reference Priors (Bernardo, 1979):

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\, I(\boldsymbol{\theta}, \mathcal{D})$$

$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathcal{D}} p(\mathcal{D})\mathrm{KLD}[p(\boldsymbol{\theta}|\mathcal{D}) \;||\; p(\boldsymbol{\theta})]d\mathcal{D}.$$

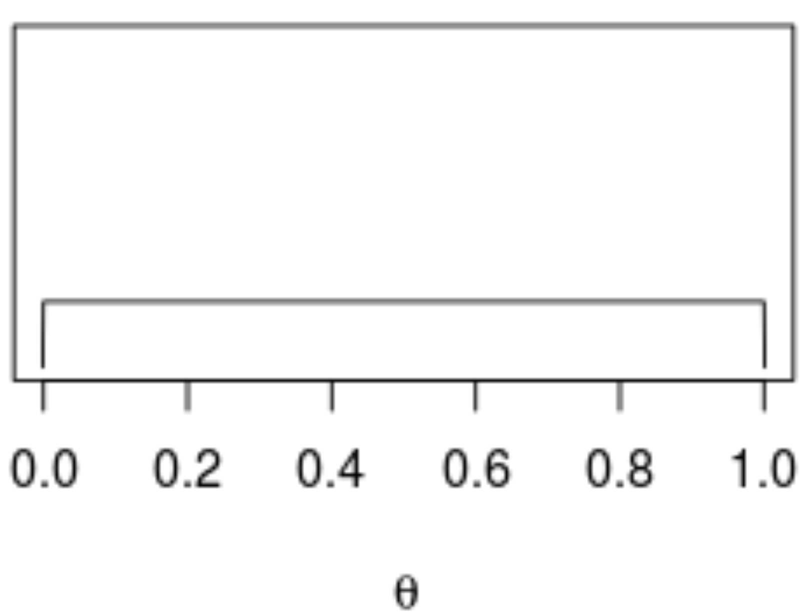
Objective Bayesian Priors

Reference Priors (Bernardo, 1979):

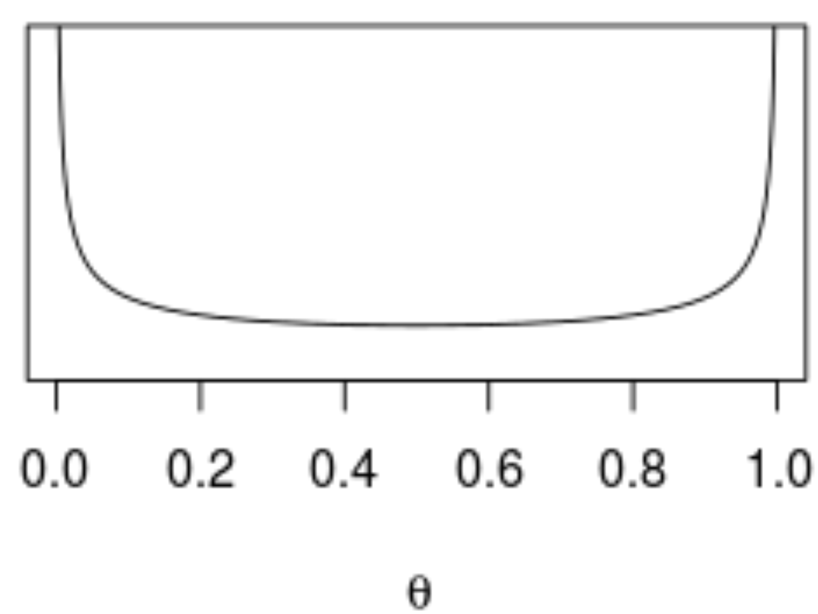
$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\, I(\boldsymbol{\theta}, \mathcal{D})$$
$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathcal{D}} p(\mathcal{D})\mathrm{KLD}[p(\boldsymbol{\theta}|\mathcal{D}) \mid\mid p(\boldsymbol{\theta})] d\mathcal{D}.$$

Equivalent to *Jeffreys* priors in one-dimension.

FLAT PRIOR



REFERENCE / JEFFREYS PRIOR



Reference Priors

- Prior that results in the data having max affect on the posterior.
- Posterior credible intervals match the corresponding confidence intervals*.
- Called ‘reference’ because they serve as a point of comparison for subjective priors.

* conditions apply

Reference Priors

We can lower-bound the mutual information with the following Monte Carlo objective:

$$I(\boldsymbol{\theta}, \mathcal{D}) \geq \mathcal{J}_{\text{RP}}(\boldsymbol{\lambda})$$

Reference Priors

We can lower-bound the mutual information with the following Monte Carlo objective:

$$I(\boldsymbol{\theta}, \mathcal{D}) \geq \mathcal{J}_{\text{RP}}(\boldsymbol{\lambda})$$

$$= \mathbb{E}_{\boldsymbol{\theta}_{\boldsymbol{\lambda}}} \left[-\mathbb{H}_{\mathcal{D}|\boldsymbol{\theta}}[\mathcal{D}] - \mathbb{E}_{\mathcal{D}|\boldsymbol{\theta}}[\max_{s} \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)] \right]$$

where $\hat{\theta}_s \sim p_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$

Reference Priors

We can lower-bound the mutual information with the following Monte Carlo objective:

$$I(\boldsymbol{\theta}, \mathcal{D}) \geq \mathcal{J}_{\text{RP}}(\boldsymbol{\lambda})$$

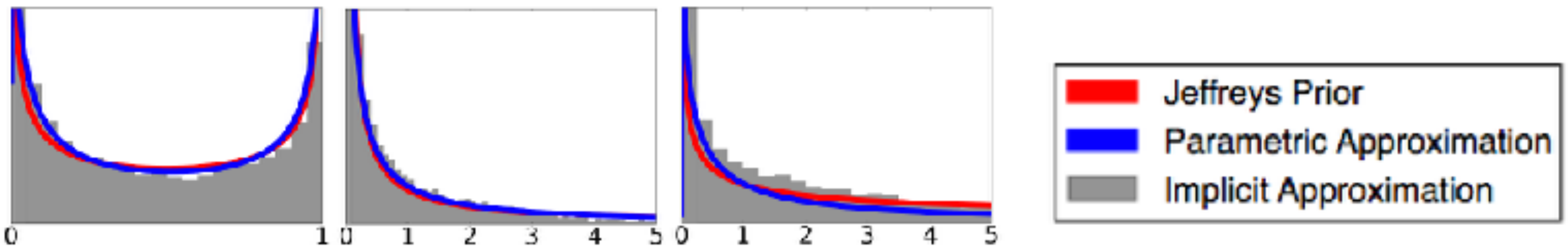
$$= \mathbb{E}_{\boldsymbol{\theta}_{\boldsymbol{\lambda}}} \left[-\mathbb{H}_{\mathcal{D}|\boldsymbol{\theta}}[\mathcal{D}] - \mathbb{E}_{\mathcal{D}|\boldsymbol{\theta}}[\max_{s} \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)] \right]$$

$$= \frac{1}{S} \sum_{s=1}^{S} \text{KLD}[p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s) \mid\mid p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{\max})]$$

where $\hat{\theta}_s \sim p_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$

Reference Priors

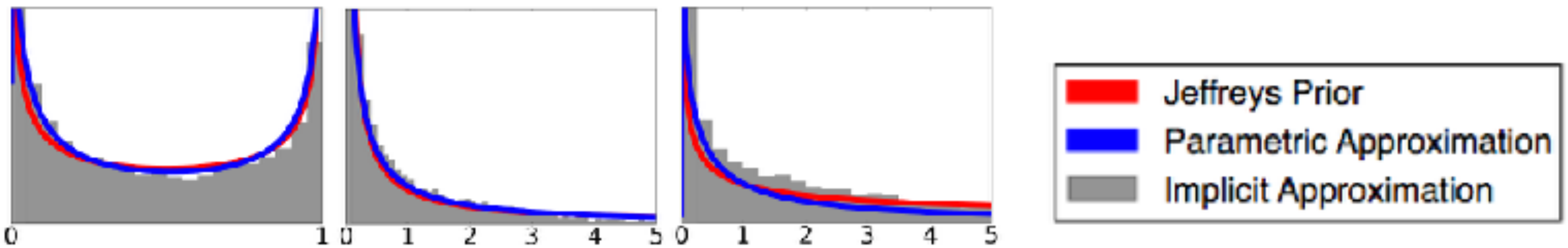
Recovering Jeffreys priors:



(a) Bernoulli (b) Gaussian Scale (c) Poisson

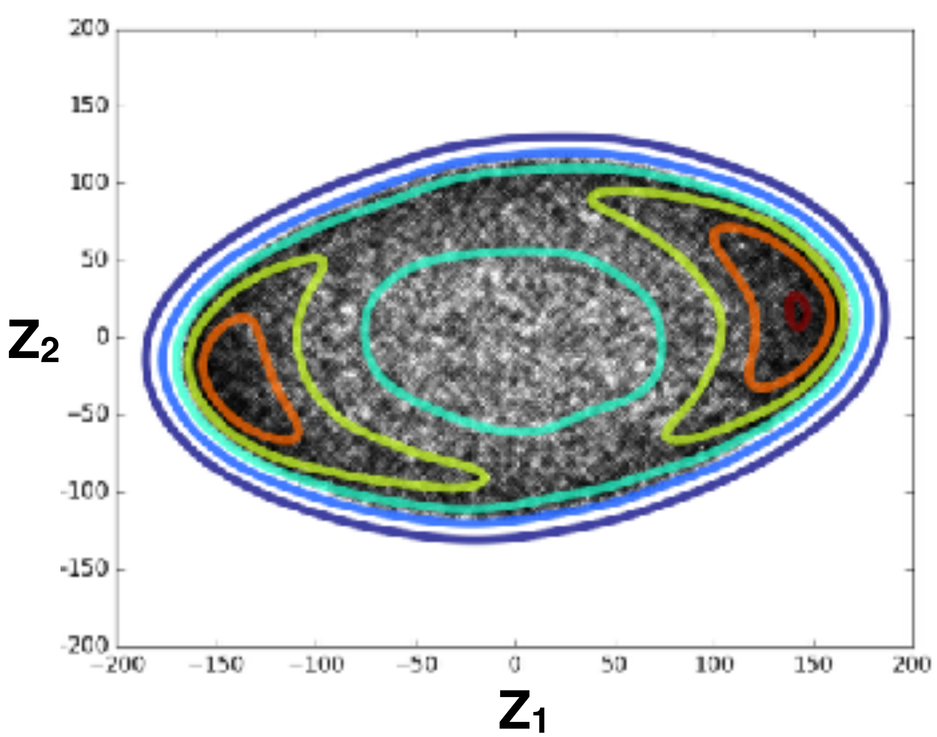
Reference Priors

Recovering Jeffreys priors:



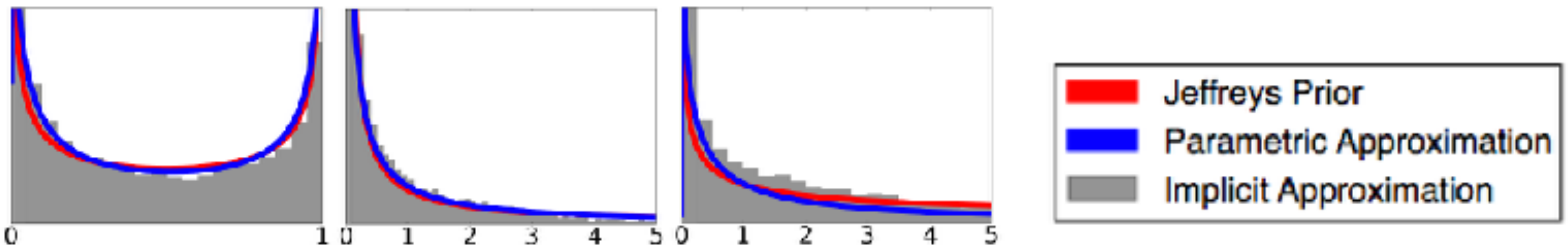
(a) Bernoulli (b) Gaussian Scale (c) Poisson

Variational Autoencoder's reference prior:



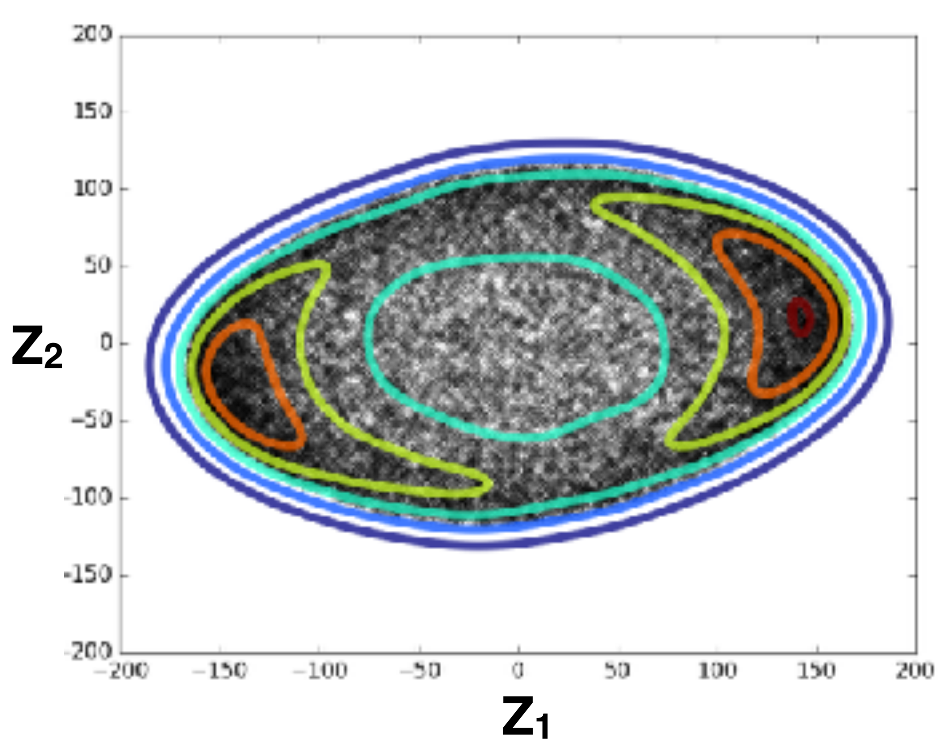
Reference Priors

Recovering Jeffreys priors:



(a) Bernoulli (b) Gaussian Scale (c) Poisson

Variational Autoencoder's reference prior:



Improves performance for low-dimensional (< 15) latent spaces but gives (approx.) identical performance for 50 dims, the size commonly used.

The Amortized Bootstrap

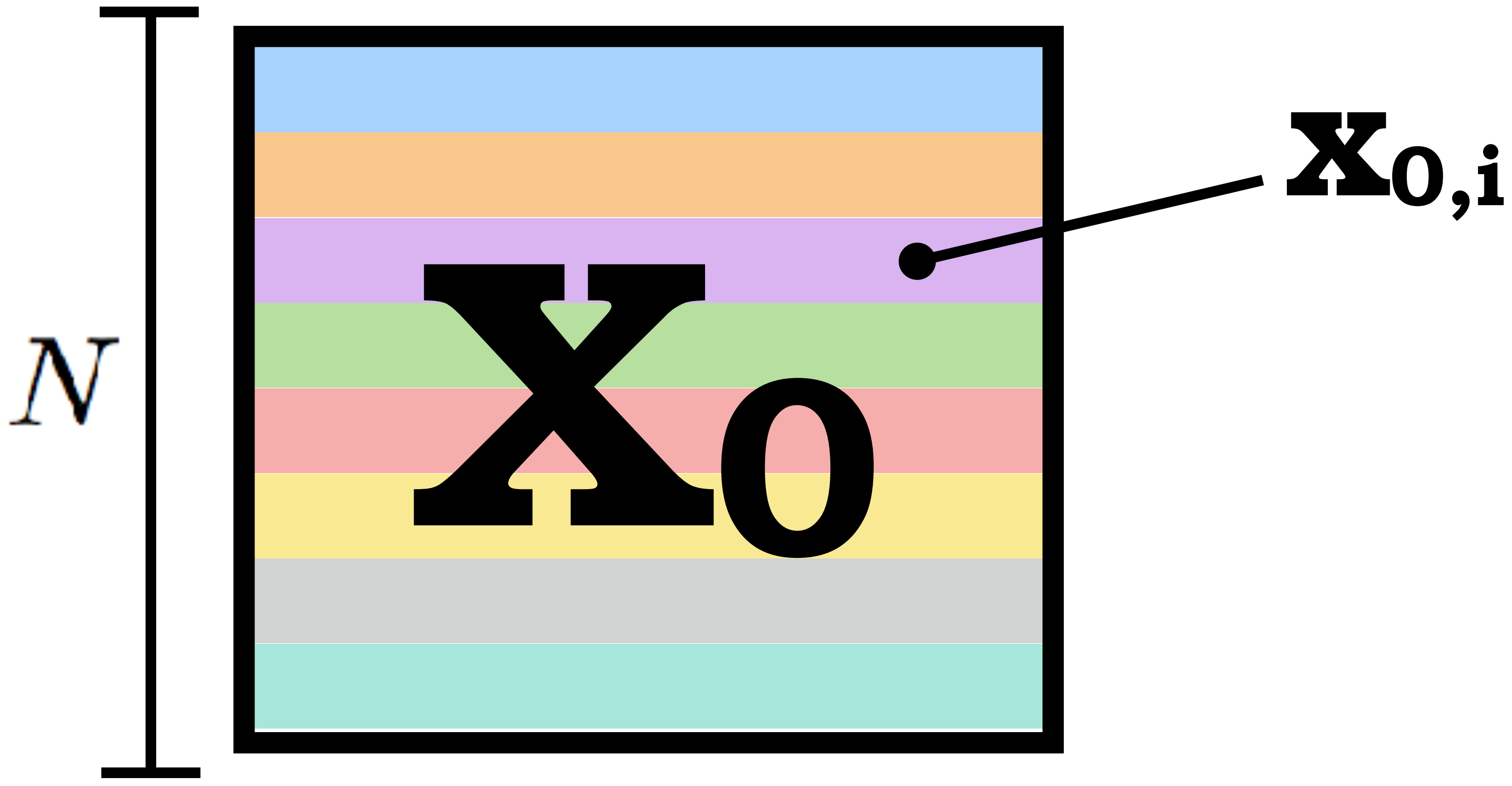
(Nalisnick & Smyth, SoCalML 2017)

The Bootstrap

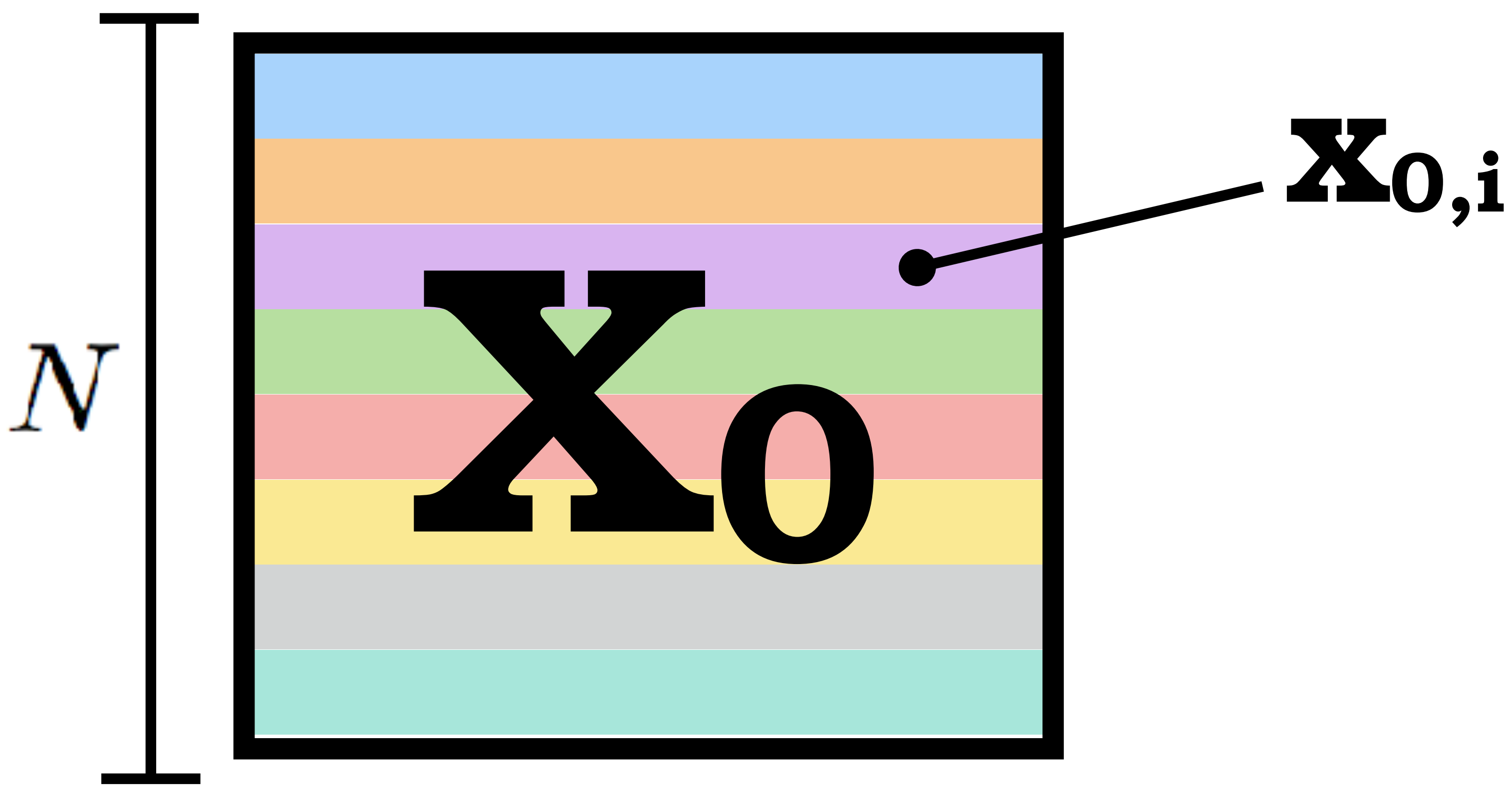
Bootstrap Resampling

$$\boxed{X_0}$$

Bootstrap Resampling

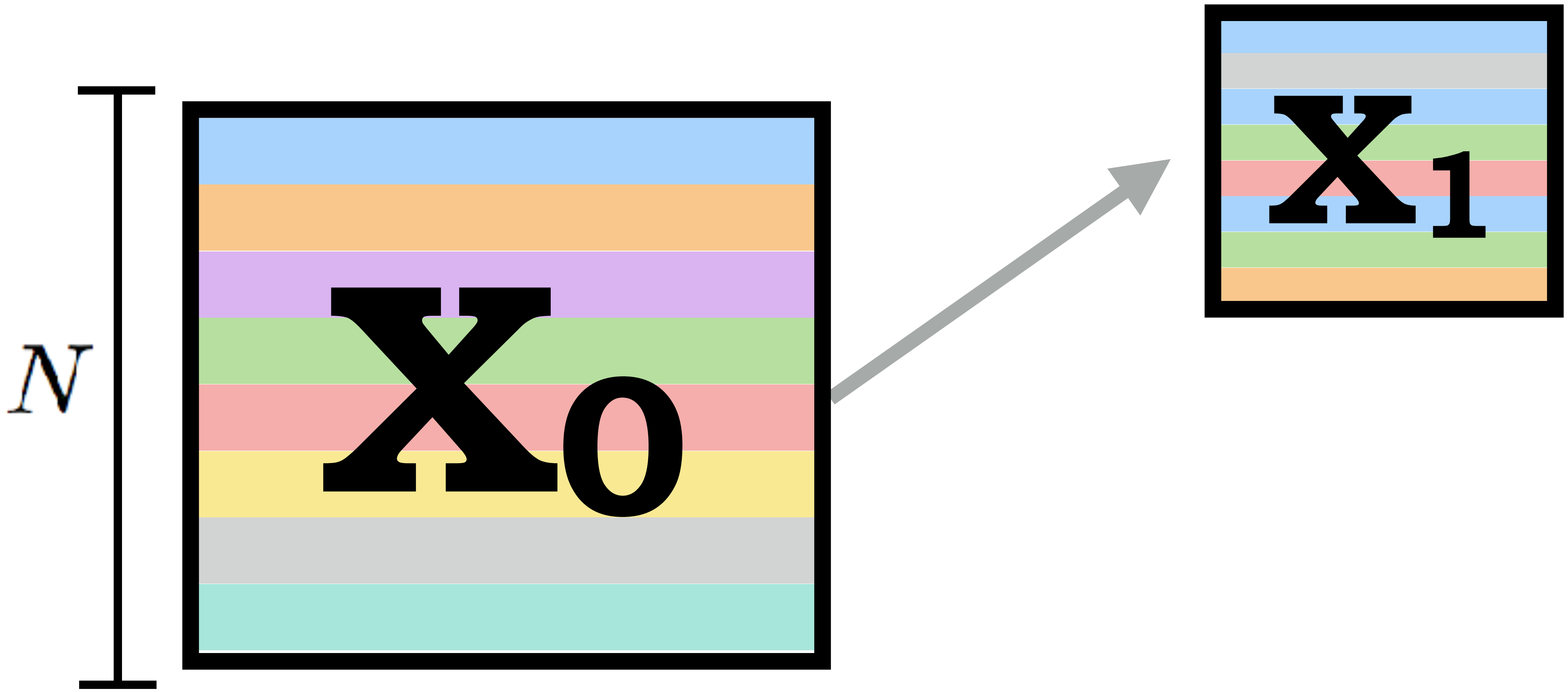


Bootstrap Resampling



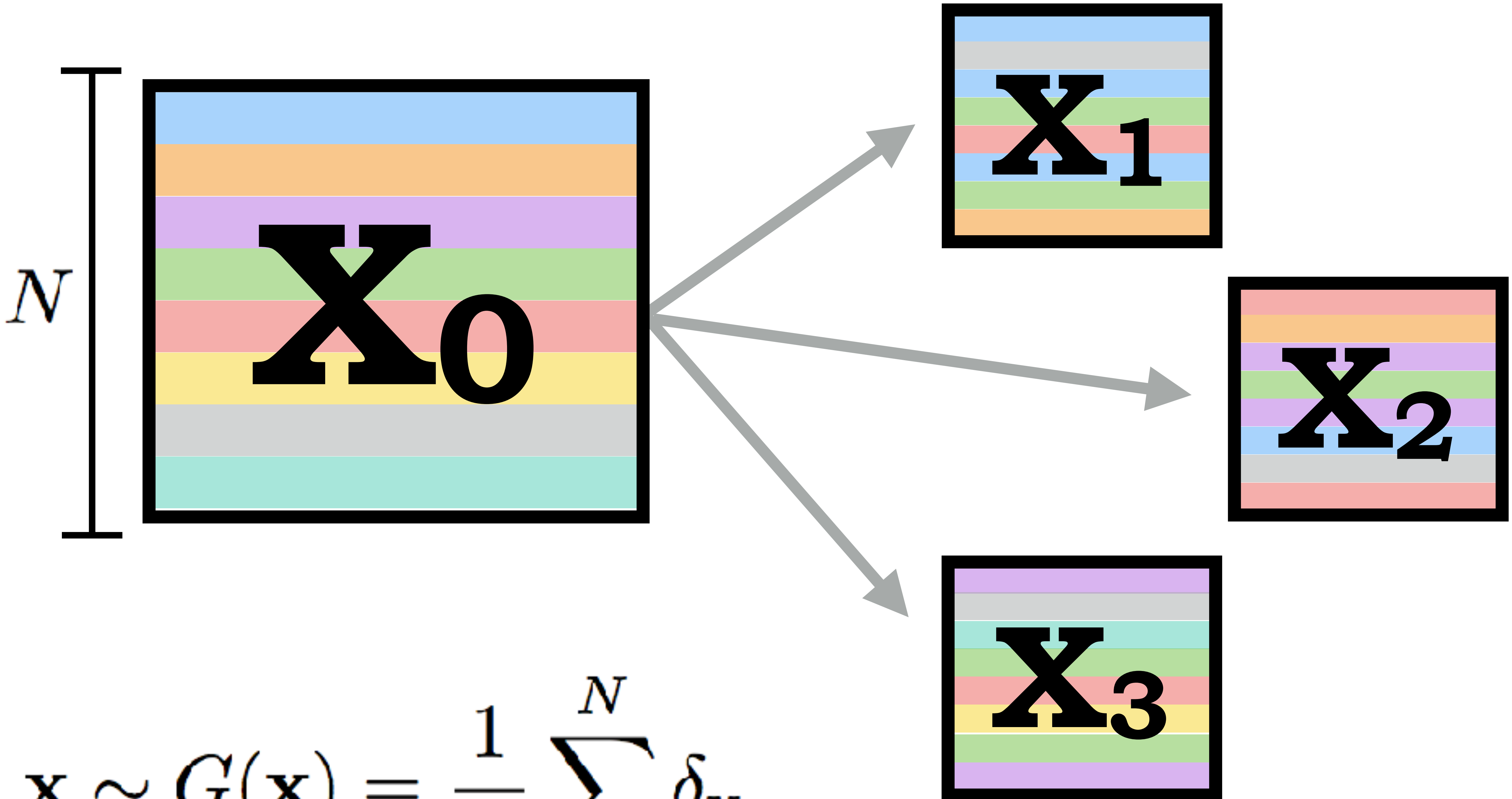
$$\mathbf{x} \sim G(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_{0,i}}$$

Bootstrap Resampling



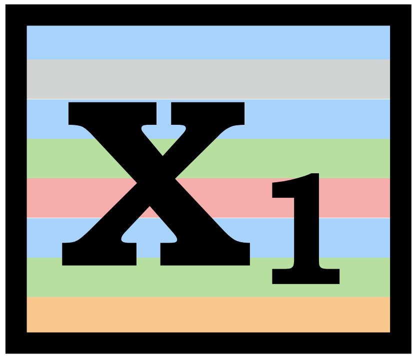
$$\mathbf{x} \sim G(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_{0,i}}$$

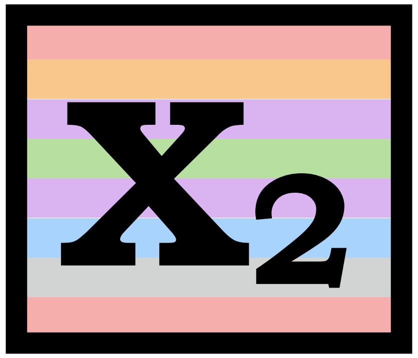
Bootstrap Resampling

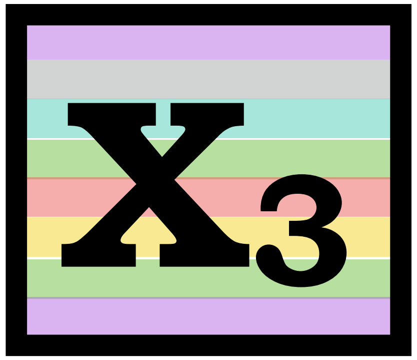


$$\mathbf{x} \sim G(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_{0,i}}$$

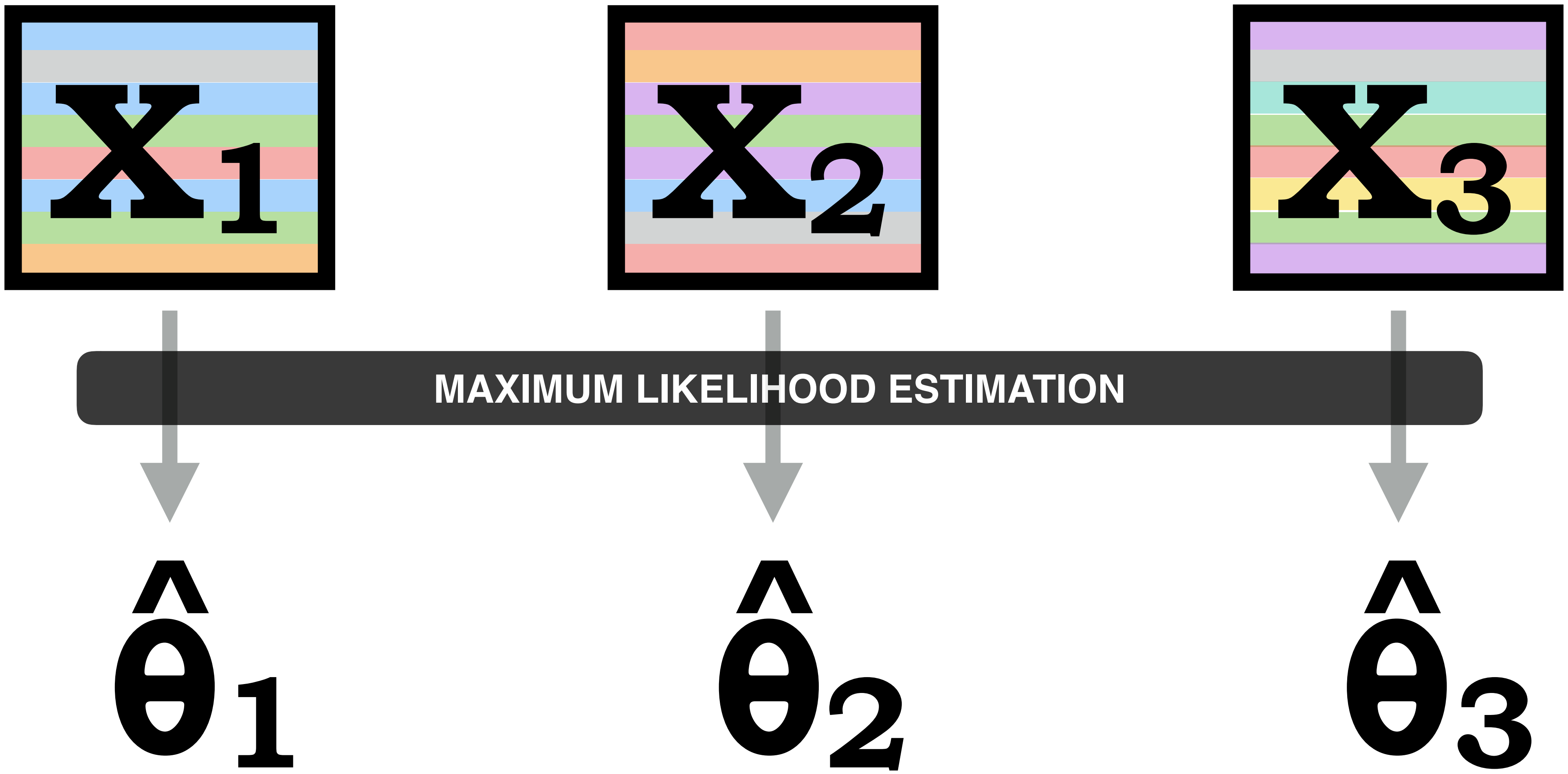
Bootstrap Distribution



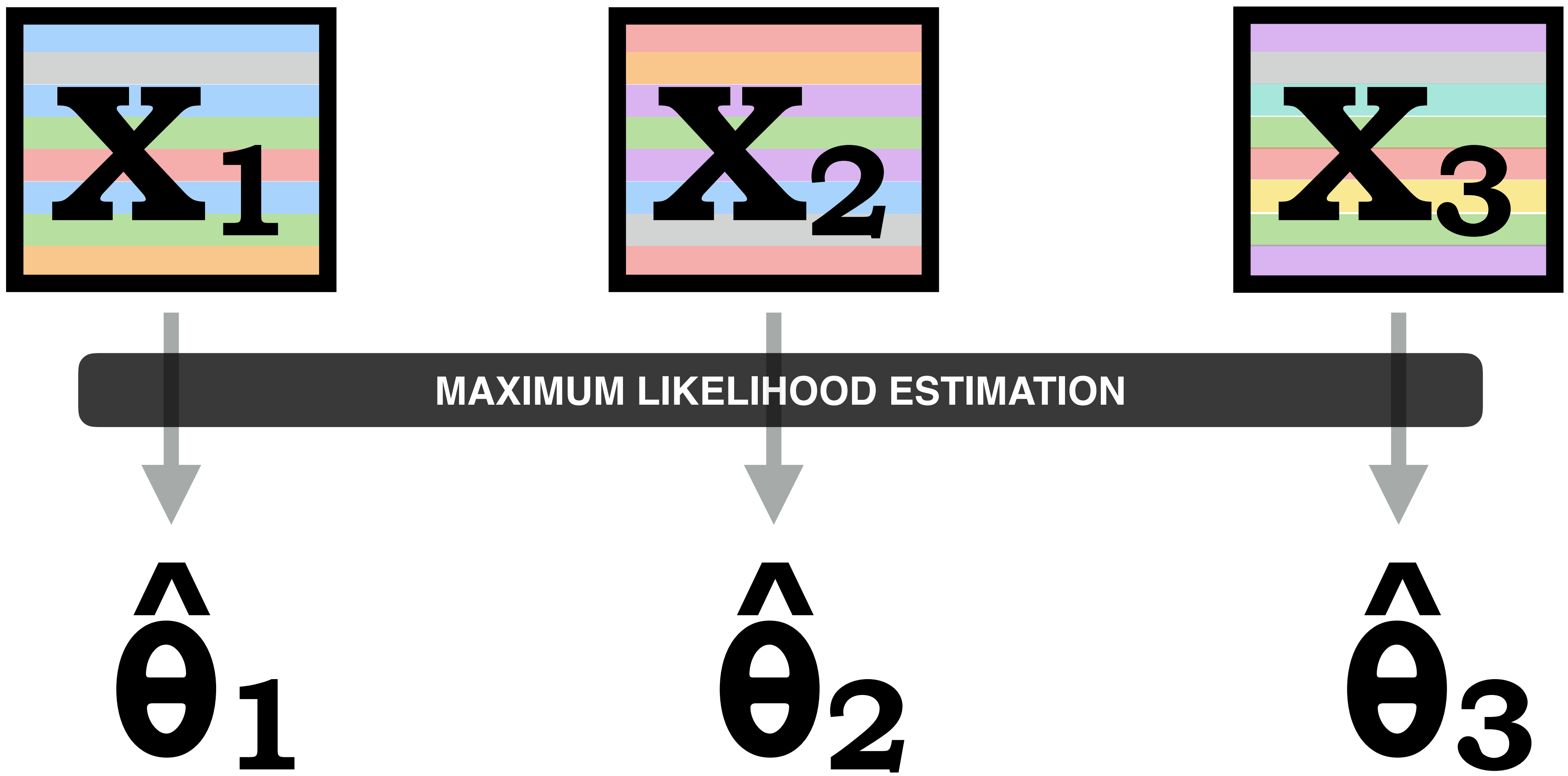




Bootstrap Distribution



Bootstrap Distribution



$$\boldsymbol{\theta} \sim F(\boldsymbol{\theta}) = \frac{1}{K} \sum_{k=1}^{K} \delta_{\hat{\boldsymbol{\theta}}_k}$$

The Amortized Bootstrap

Modeling the Bootstrap Distribution

QUESTION: Can we approximate the bootstrap distribution F(**θ**) with a model (like in variational inference for Bayesian posteriors)?

Modeling the Bootstrap Distribution

QUESTION: Can we approximate the bootstrap distribution F(**θ**) with a model (like in variational inference for Bayesian posteriors)?

IDEA: Use an *implicit model* to approximate F(**θ**).

$$\hat{\boldsymbol{\theta}} = f_{\phi}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \sim p_0$$

Modeling the Bootstrap Distribution

IDEA: Use an *implicit model* to approximate F(**θ**).

$$\hat{\boldsymbol{\theta}} = f_{\boldsymbol{\phi}}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \sim p_0$$

Modeling the Bootstrap Distribution

IDEA: Use an *implicit model* to approximate F(**θ**).

$$\hat{\boldsymbol{\theta}} = f_{\phi}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \sim p_0$$

PROS

- [ ] **Amortized Inference:** share statistical strength across dataset replications / generate unlimited samples.
- [ ] Results in **bootstrap smoothing** (Efron & Tibshirani, 1997).

Modeling the Bootstrap Distribution

IDEA: Use an *implicit model* to approximate F(**θ**).

$$\hat{\boldsymbol{\theta}} = f_{\phi}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \sim p_0$$

PROS

- **Amortized Inference:** share statistical strength across dataset replications / generate unlimited samples.
- Results in **bootstrap smoothing** (Efron & Tibshirani, 1997).

CONS

- Breaks bootstrap theory. Can recover only an approximation.
- Can't distribute computation.

Learning the Bootstrap Model

$$\mathcal{J}(\mathbf{X}_0, \phi) = \mathbb{E}_{F_{\boldsymbol{\phi}}(\boldsymbol{\theta})} \mathbb{E}_{G(\mathbf{x})} [\log p(\mathbf{X}|\boldsymbol{\theta})]$$

Learning the Bootstrap Model

$$\mathcal{J}(\mathbf{X}_0, \phi) = \mathbb{E}_{F_{\boldsymbol{\phi}}(\boldsymbol{\theta})}\mathbb{E}_{G(\mathbf{x})}[\log p(\mathbf{X}|\boldsymbol{\theta})]$$

$$\approx \frac{1}{K}\sum_{k=1}^{K} \log p(\mathbf{X}_k|\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi},k})$$

Learning the Bootstrap Model

$$\mathcal{J}(\mathbf{X}_0, \phi) = \mathbb{E}_{F_{\boldsymbol{\phi}}(\boldsymbol{\theta})}\mathbb{E}_{G(\mathbf{x})}[\log p(\mathbf{X}|\boldsymbol{\theta})]$$

$$\approx \frac{1}{K}\sum_{k=1}^{K} \log p(\mathbf{X}_k|\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi},k})$$

$$\frac{\partial \mathcal{J}(\mathbf{X}_0, \phi)}{\partial \boldsymbol{\phi}} = \frac{1}{K}\sum_{k=1}^{K} \frac{\partial \log p(\mathbf{X}_k|\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi},k})}{\partial \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi},k}} \frac{\partial \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi},k}}{\partial \boldsymbol{\phi}}$$

Learning the Bootstrap Model

$$\mathcal{J}(\mathbf{X}_0, \phi) = \mathbb{E}_{F_{\phi}(\boldsymbol{\theta})}\mathbb{E}_{G(\mathbf{x})}[\log p(\mathbf{X}|\boldsymbol{\theta})]$$

$$\approx \frac{1}{K}\sum_{k=1}^{K} \log p(\mathbf{X}_k|\hat{\boldsymbol{\theta}}_{\phi,k})$$

$$\frac{\partial \mathcal{J}(\mathbf{X}_0, \phi)}{\partial \phi} = \frac{1}{K}\sum_{k=1}^{K} \underbrace{\frac{\partial \log p(\mathbf{X}_k|\hat{\boldsymbol{\theta}}_{\phi,k})}{\partial \hat{\boldsymbol{\theta}}_{\phi,k}}}_{\text{Regular bootstrap update}} \frac{\partial \hat{\boldsymbol{\theta}}_{\phi,k}}{\partial \phi}$$

Learning the Bootstrap Model

$$\mathcal{J}(\mathbf{X}_0, \phi) = \mathbb{E}_{F_{\phi}(\boldsymbol{\theta})} \mathbb{E}_{G(\mathbf{x})} [\log p(\mathbf{X}|\boldsymbol{\theta})]$$

$$\approx \frac{1}{K} \sum_{k=1}^{K} \log p(\mathbf{X}_k | \hat{\boldsymbol{\theta}}_{\phi,k})$$

$$\frac{\partial \mathcal{J}(\mathbf{X}_0, \phi)}{\partial \phi} = \frac{1}{K} \sum_{k=1}^{K} \underbrace{\frac{\partial \log p(\mathbf{X}_k | \hat{\boldsymbol{\theta}}_{\phi,k})}{\partial \hat{\boldsymbol{\theta}}_{\phi,k}}}_{\text{Regular bootstrap update}} \underbrace{\frac{\partial \hat{\boldsymbol{\theta}}_{\phi,k}}{\partial \phi}}_{\text{Shared params.}}$$

Learning the Bootstrap Model

$$\mathcal{J}(\mathbf{X}_0, \phi) = \mathbb{E}_{F_{\phi}(\boldsymbol{\theta})} \mathbb{E}_{G(\mathbf{x})} [\log p(\mathbf{X}|\boldsymbol{\theta})]$$

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\boldsymbol{\theta})} [\log p(\mathbf{X}|\boldsymbol{\theta})] - \mathbf{KLD}[q(\boldsymbol{\theta}) || p(\boldsymbol{\theta})]$$

Learning the Bootstrap Model

$$\mathcal{J}(\mathbf{X}_0, \phi) = \mathbb{E}_{F_{\phi}(\boldsymbol{\theta})}\mathbb{E}_{G(\mathbf{x})}[\log p(\mathbf{X}|\boldsymbol{\theta})]$$

Regularization preventing collapse to ML point estimate.

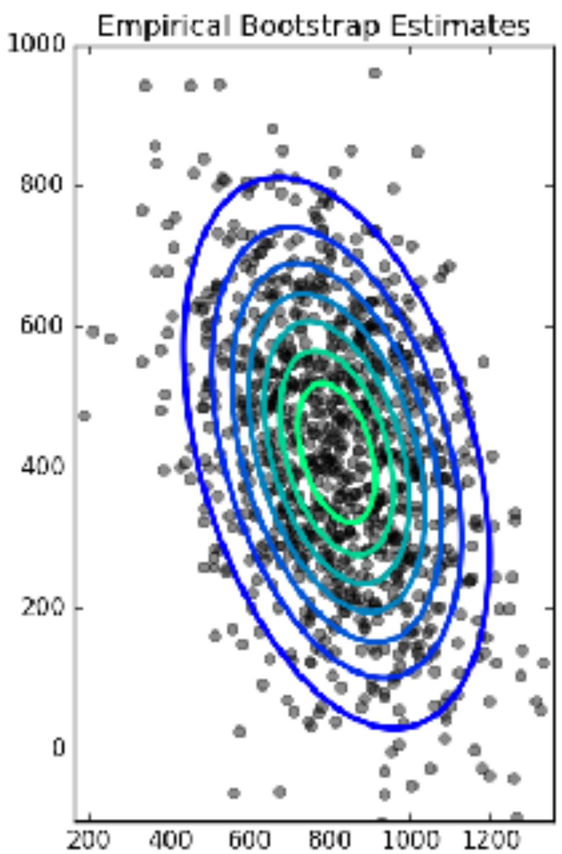
$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\mathbf{X}|\boldsymbol{\theta})] - \mathbf{KLD}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta})]$$

Data-driven uncertainty as opposed to **arbitrary priors** that can hinder performance (Hoffman & Johnson, 2016).

Experiment #1: Diagnostics

Linear Regression

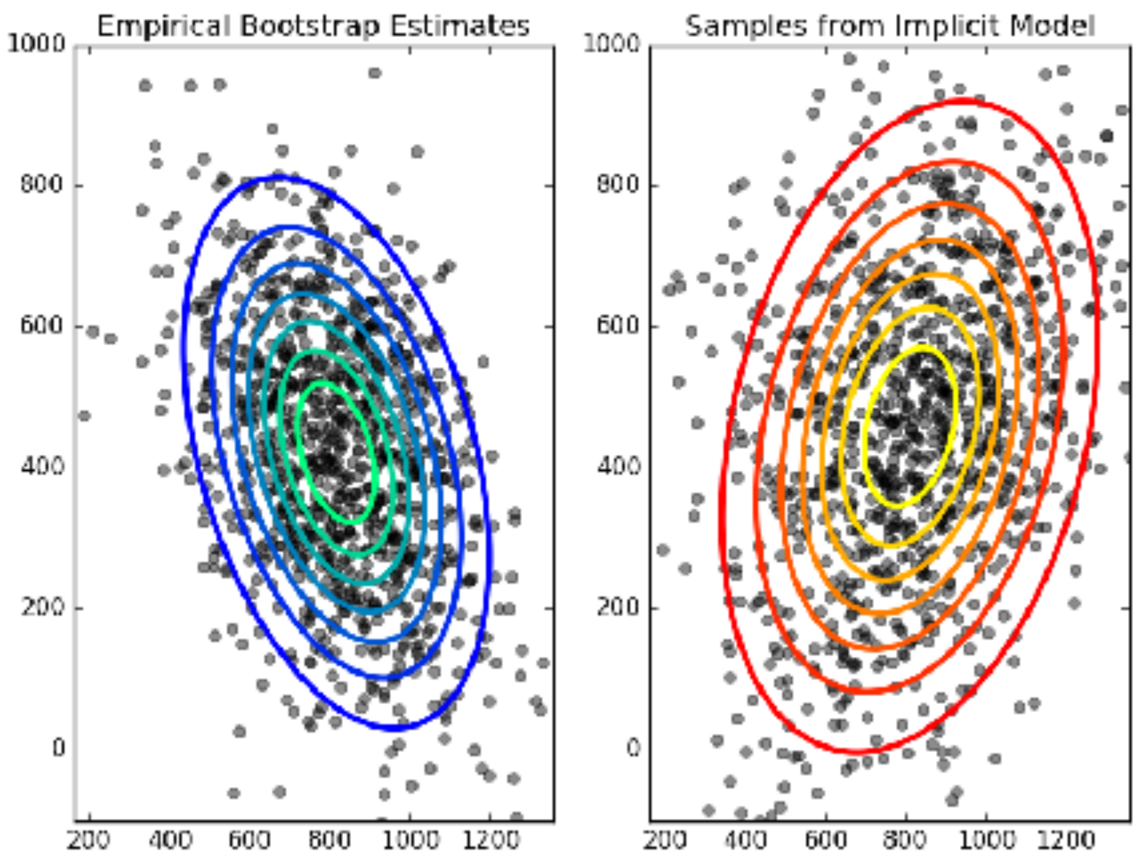
2D Diabetes Dataset



2D Boston Housing Dataset

Linear Regression

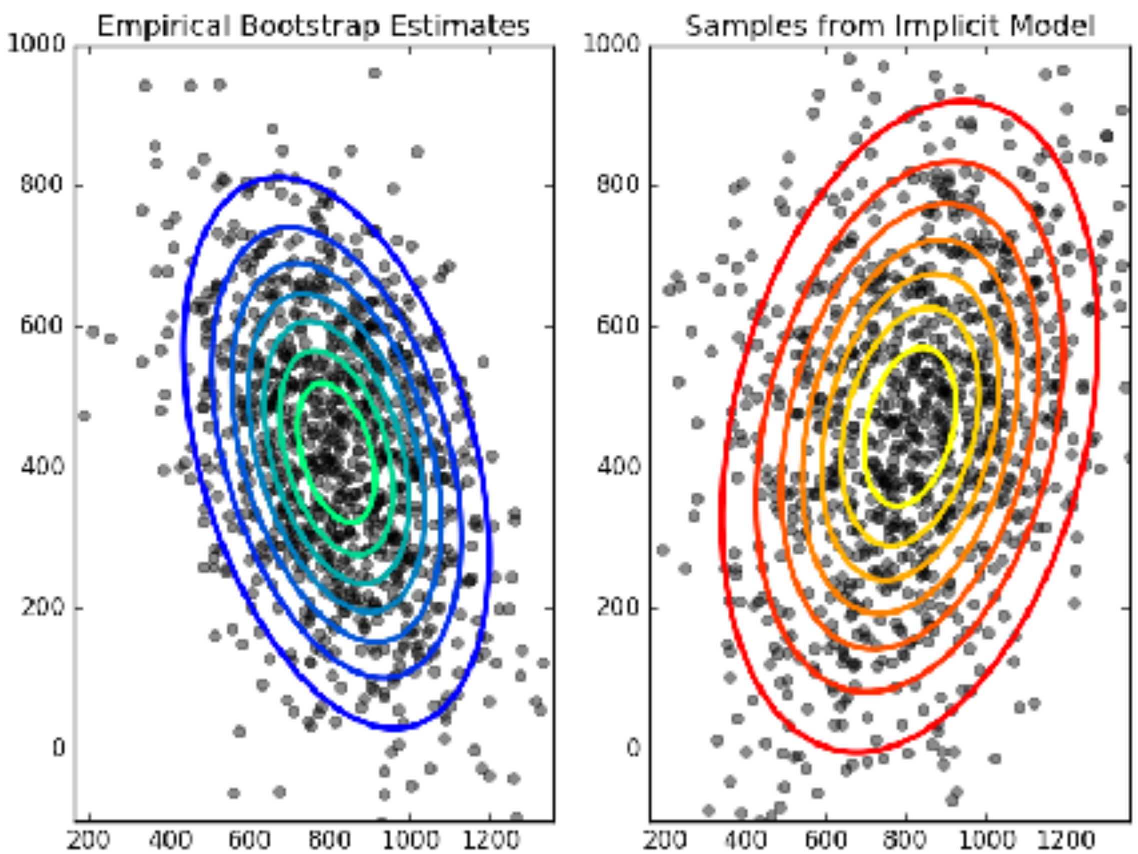
2D Diabetes Dataset



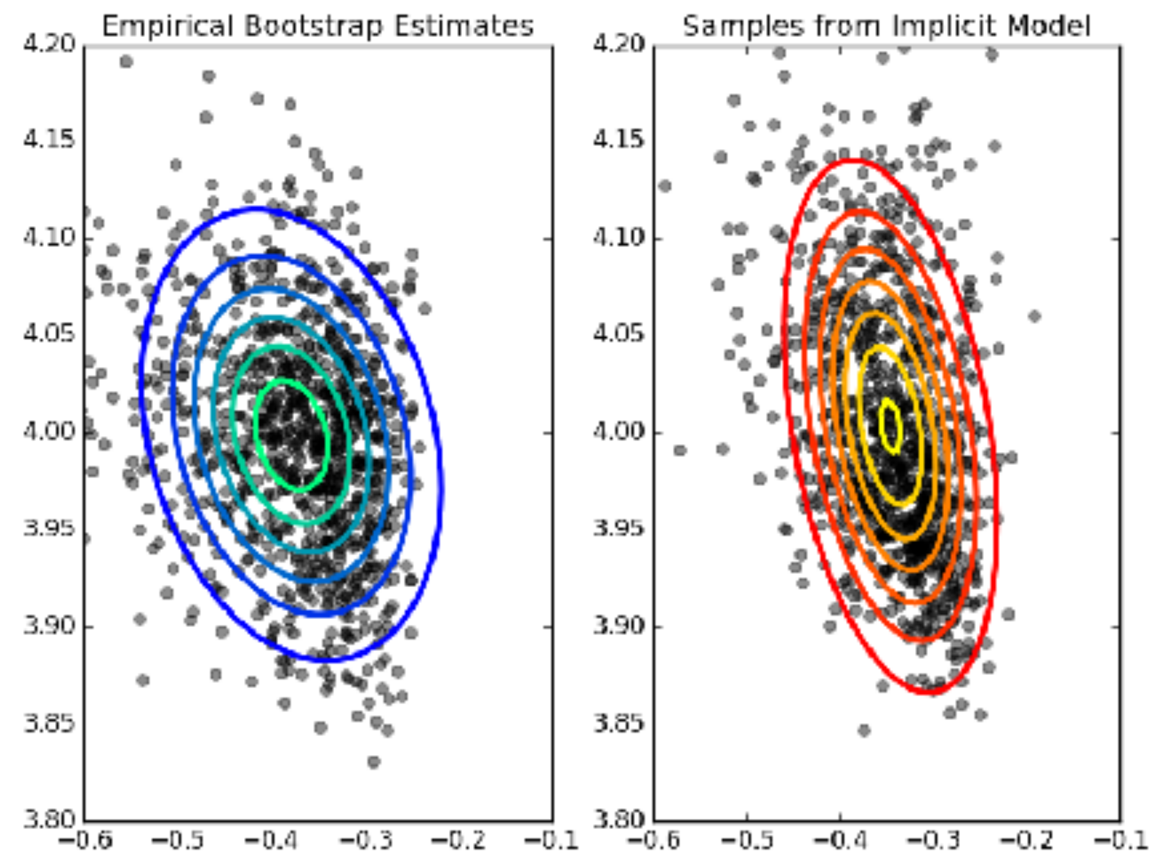
2D Boston Housing Dataset

Linear Regression

2D Diabetes Dataset

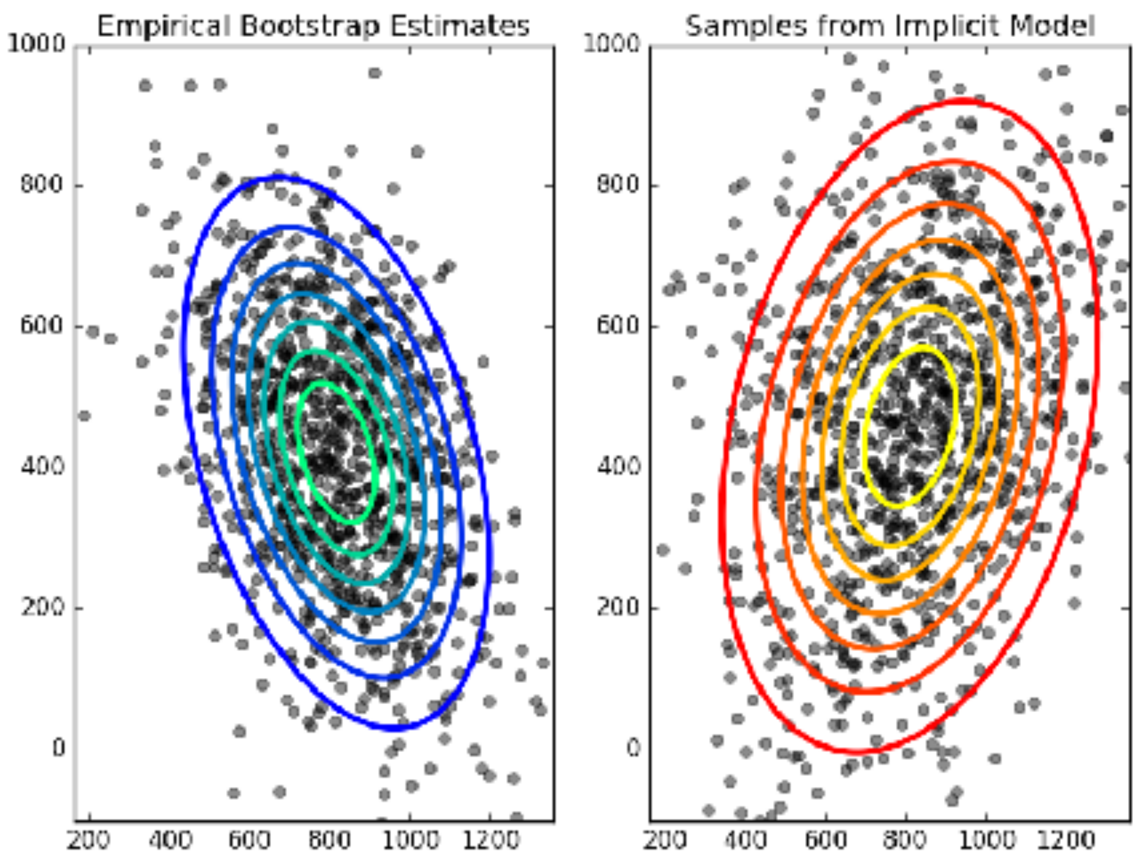


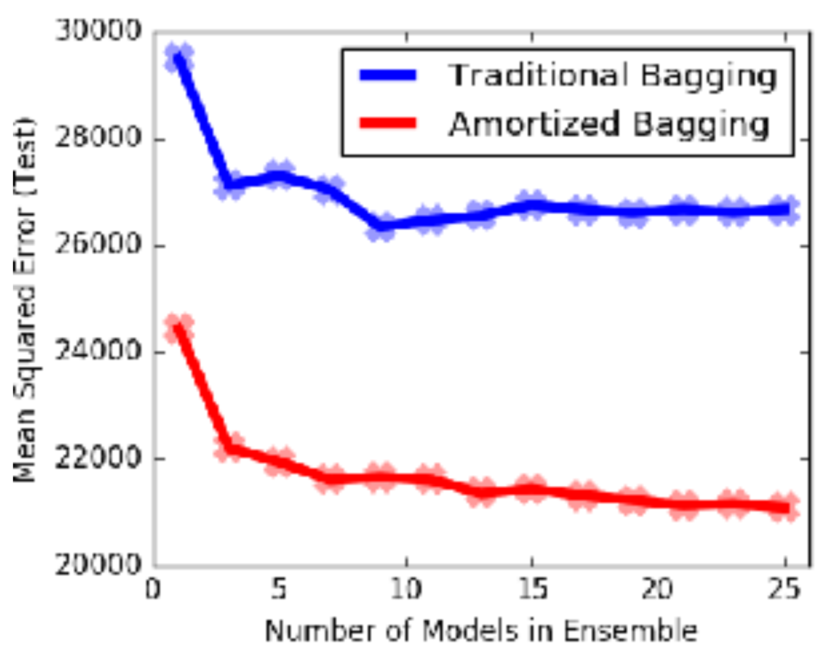
2D Boston Housing Dataset



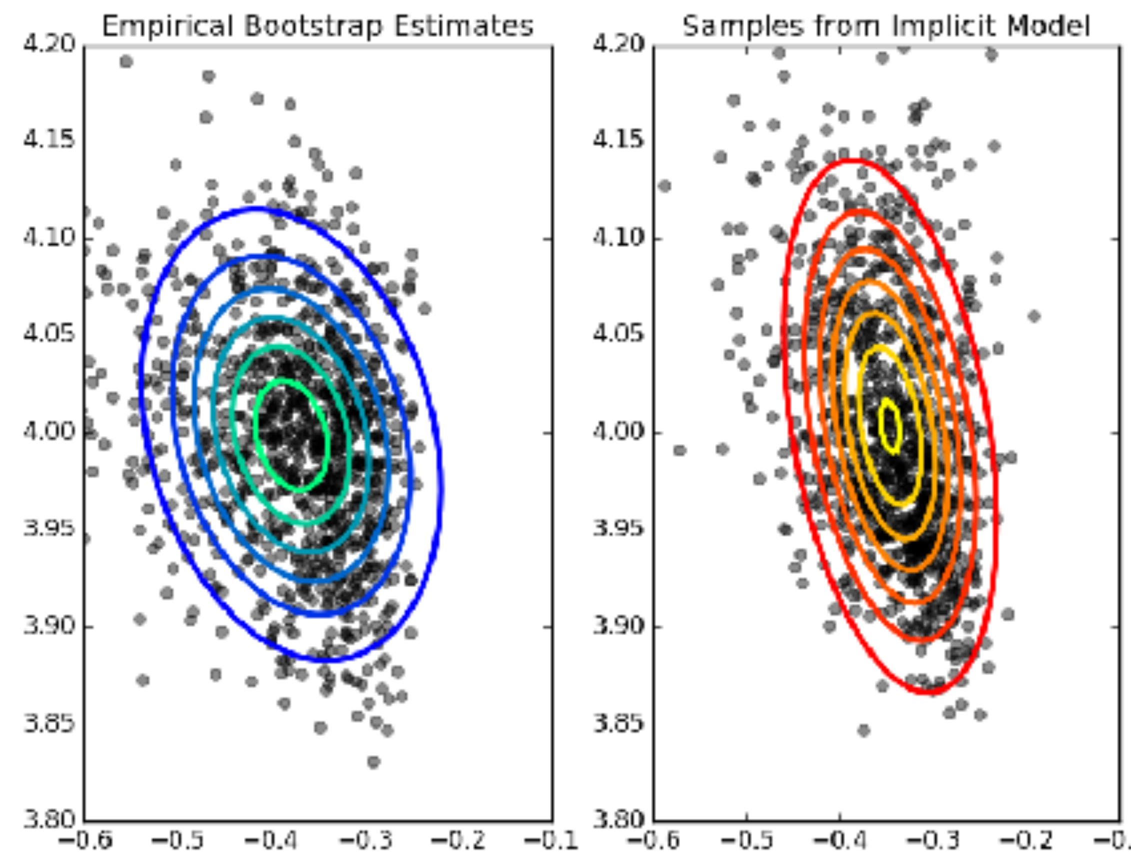
Linear Regression

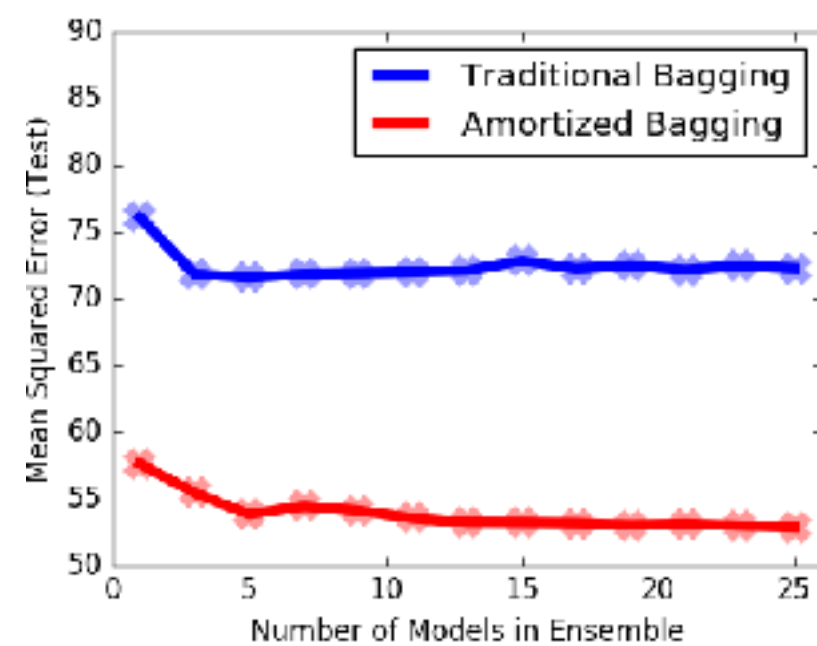
2D Diabetes Dataset





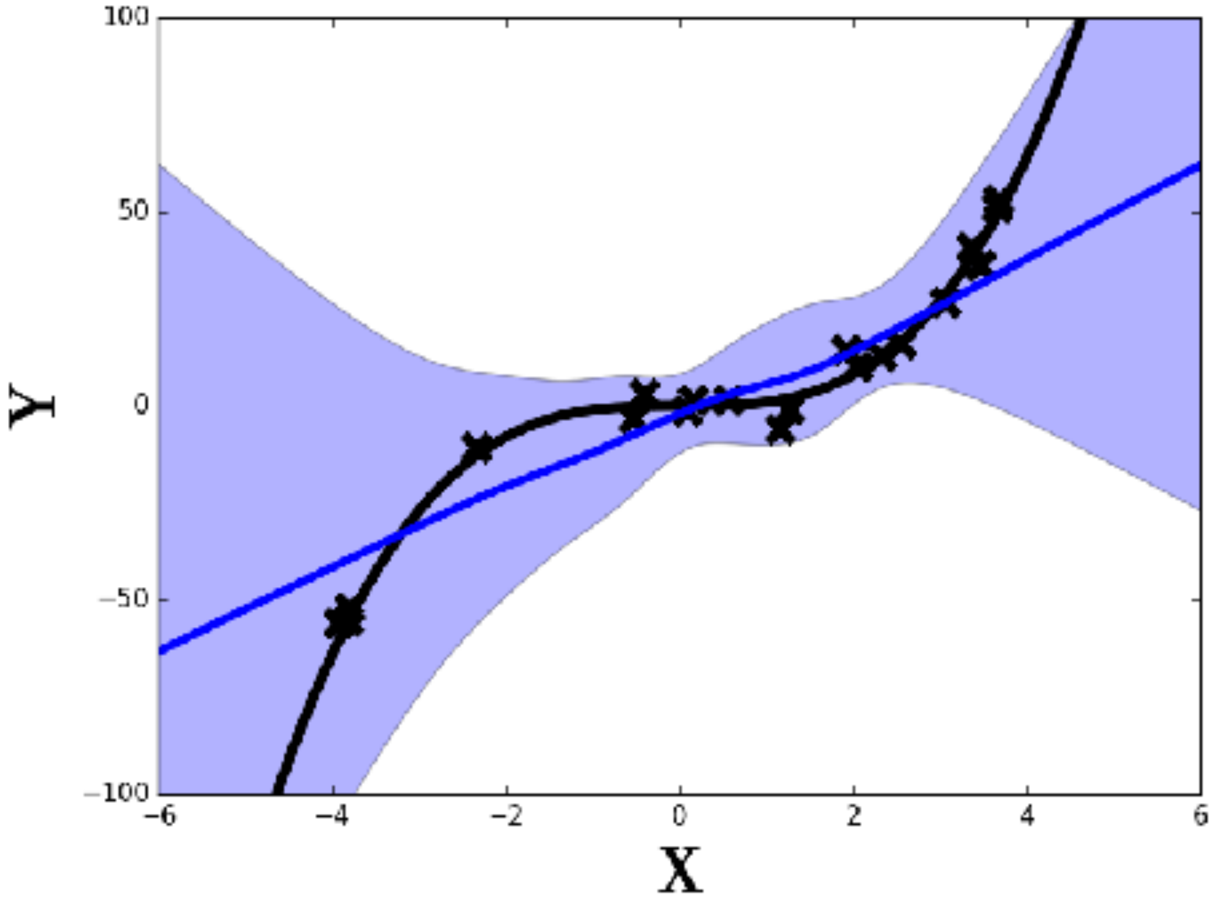
2D Boston Housing Dataset



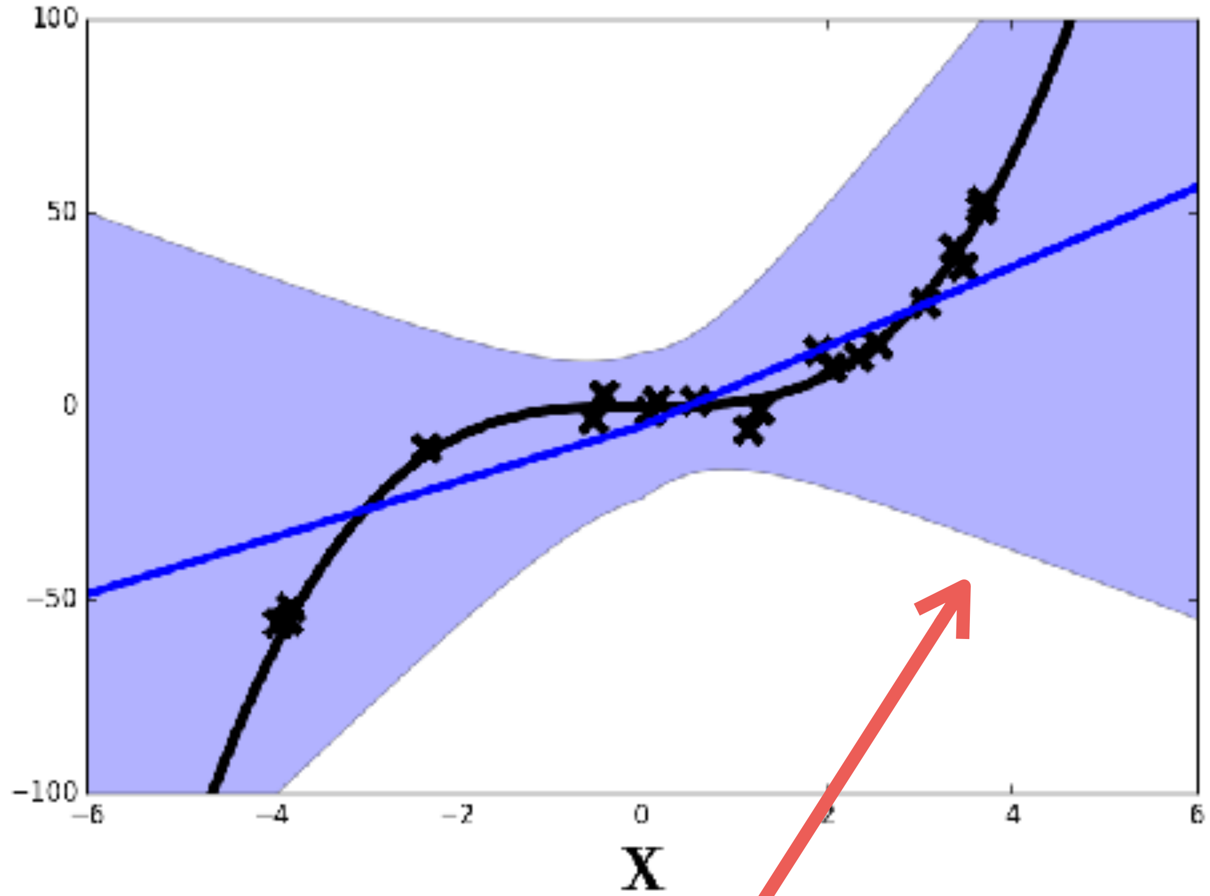


Predictive Uncertainty

TRADITIONAL BOOTSTRAP



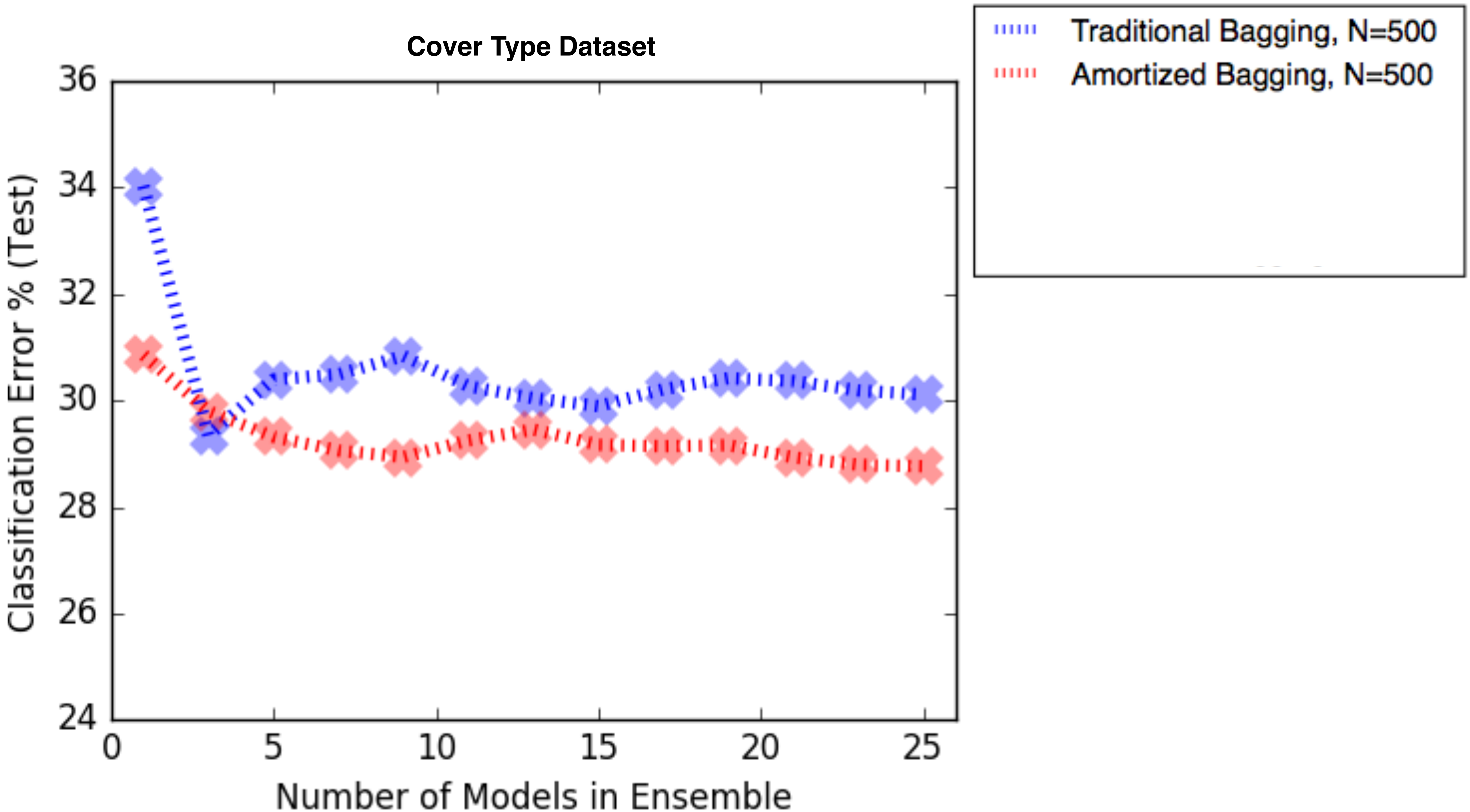
AMORTIZED BOOTSTRAP



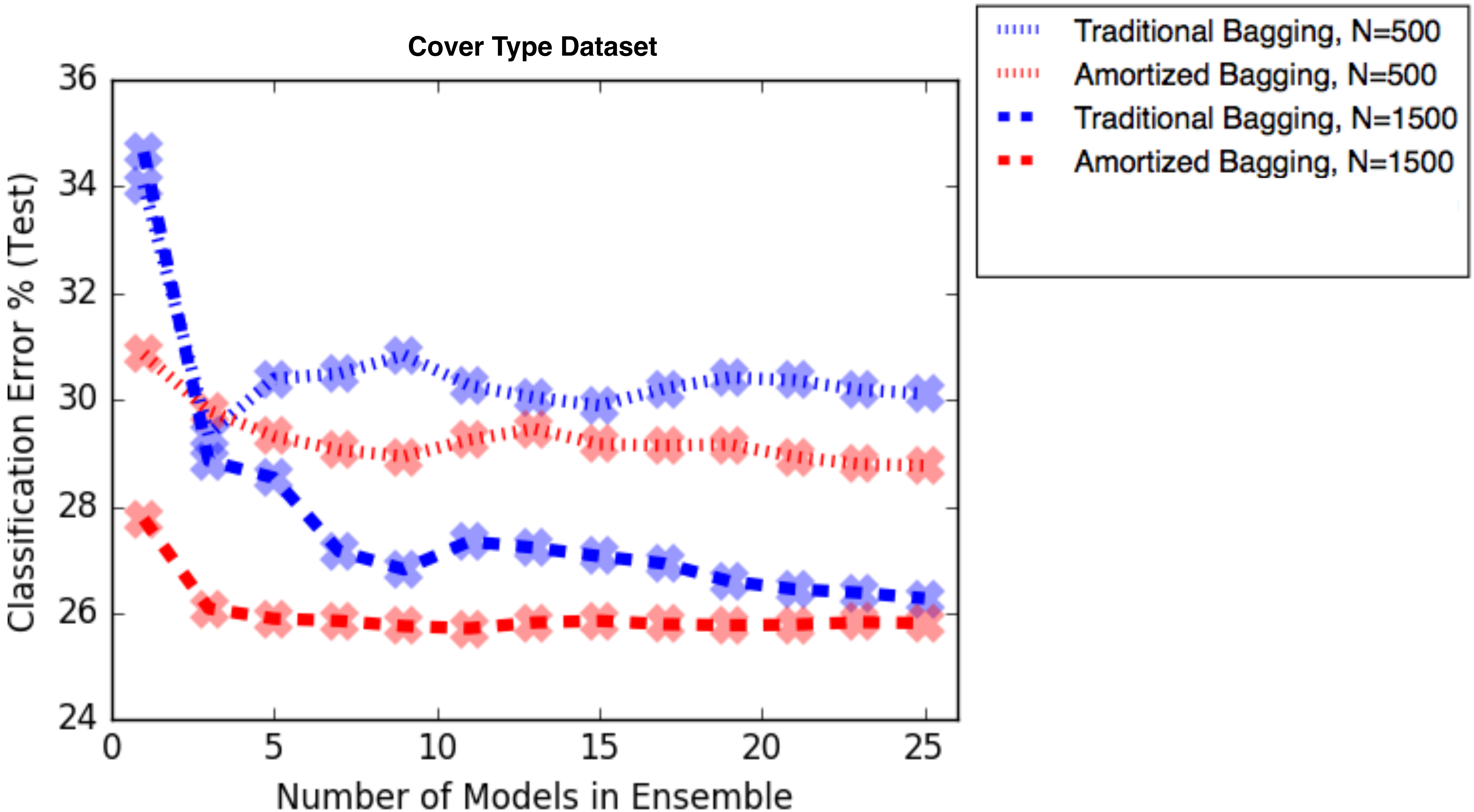
Smooth uncertainty bands, which will likely help in high-dimensions.

Experiment #2: Varying Dataset Size

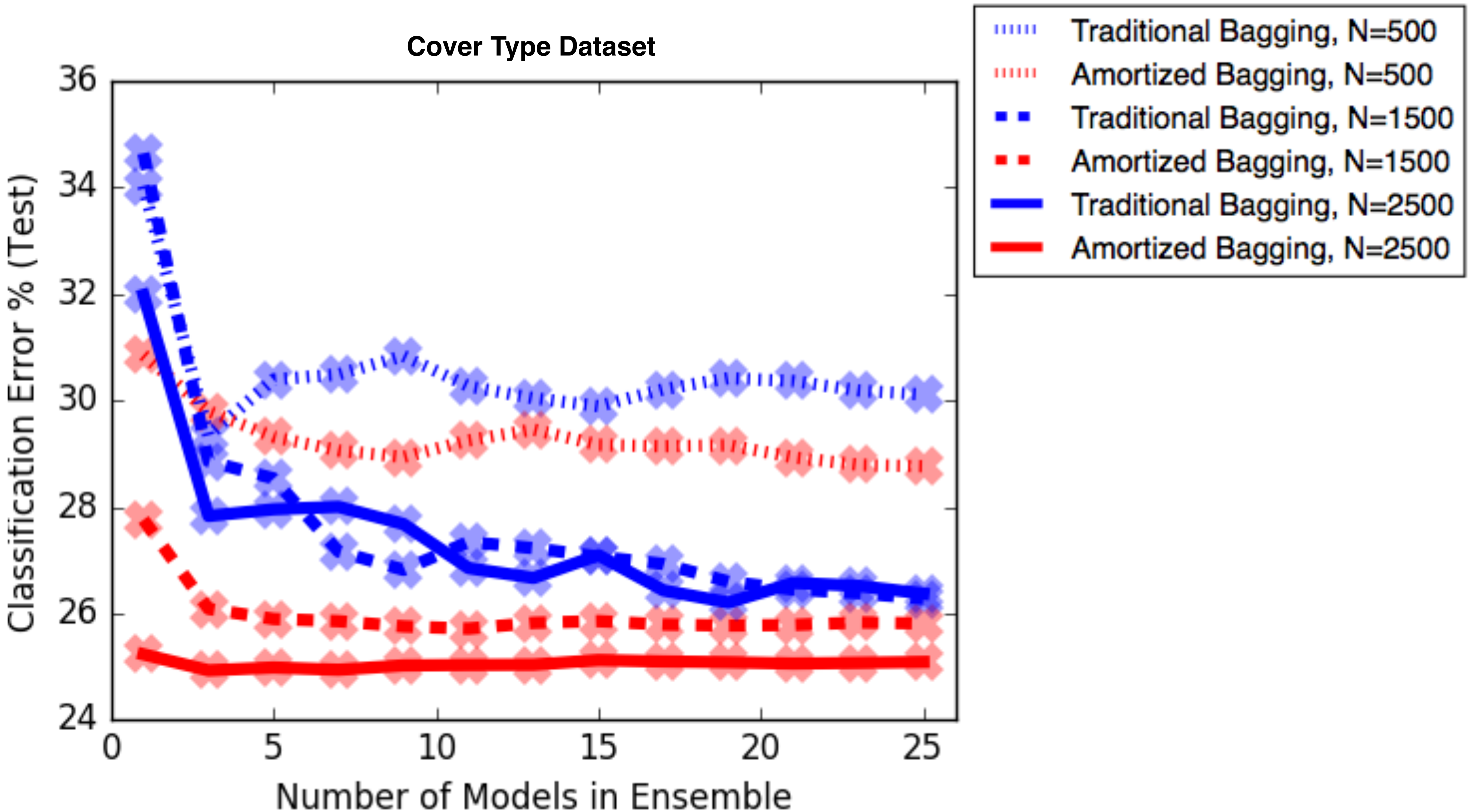
Logistic Regression



Logistic Regression



Logistic Regression



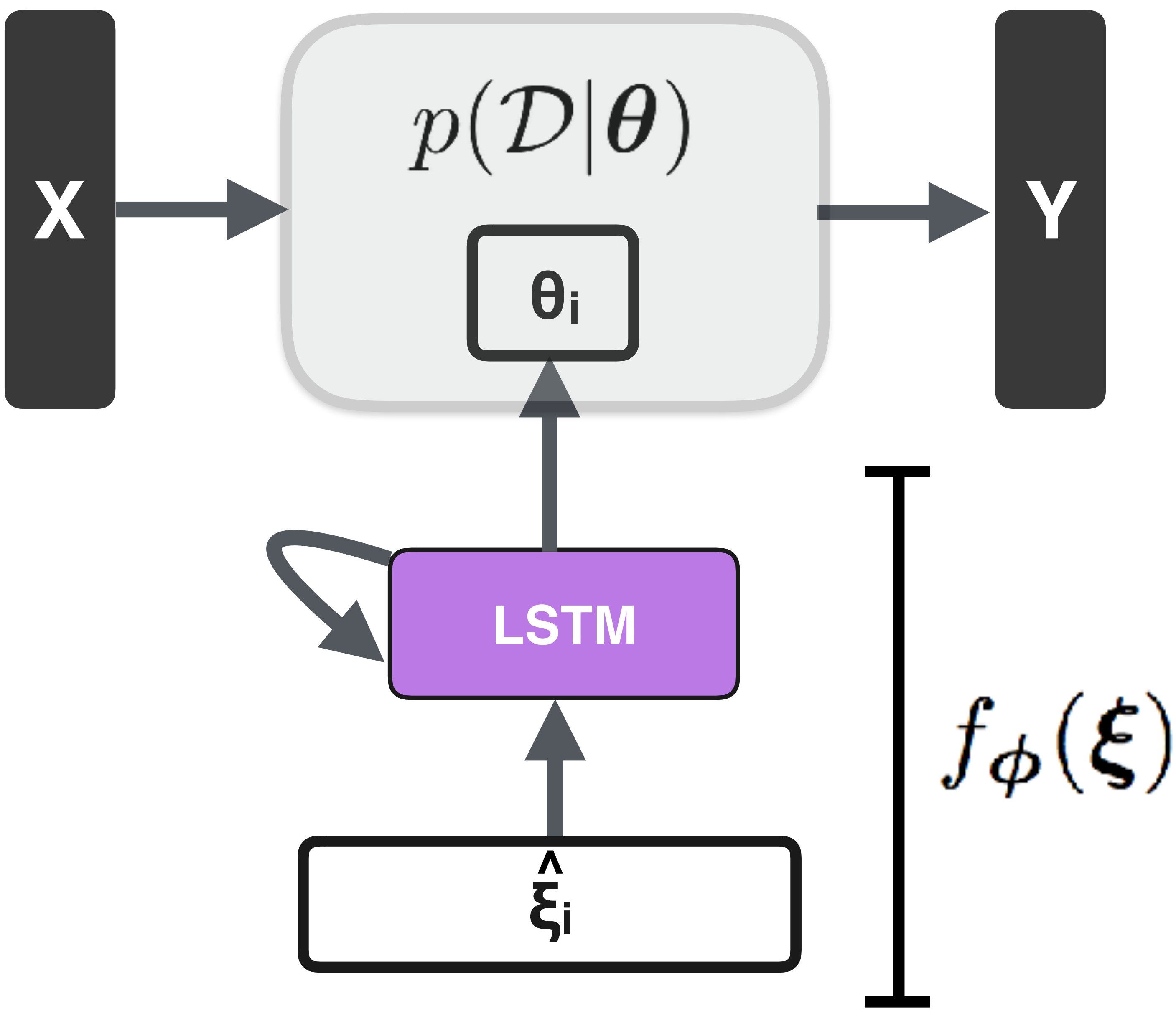
Experiment #3: Classification with NN

Neural Networks

	Test Error for Ensemble of Size K		
	$K = 1$	$K = 5$	$K = 25$
Bagged NNs, Traditional	22.57	19.68	18.57
Bagged NNs, **Amortized**	17.03	16.82	**16.18**

Rotated MNIST Dataset

In-Progress Work: Use RNN Implicit Model



Improve scalability with RNN implicit model.

— NN parameters exhibit low-dim. structure (Denil et al., 2013)

Conclusion: Approx. Inference for Frequentist Tools

- Approximating 'Frequentist-esque' priors.

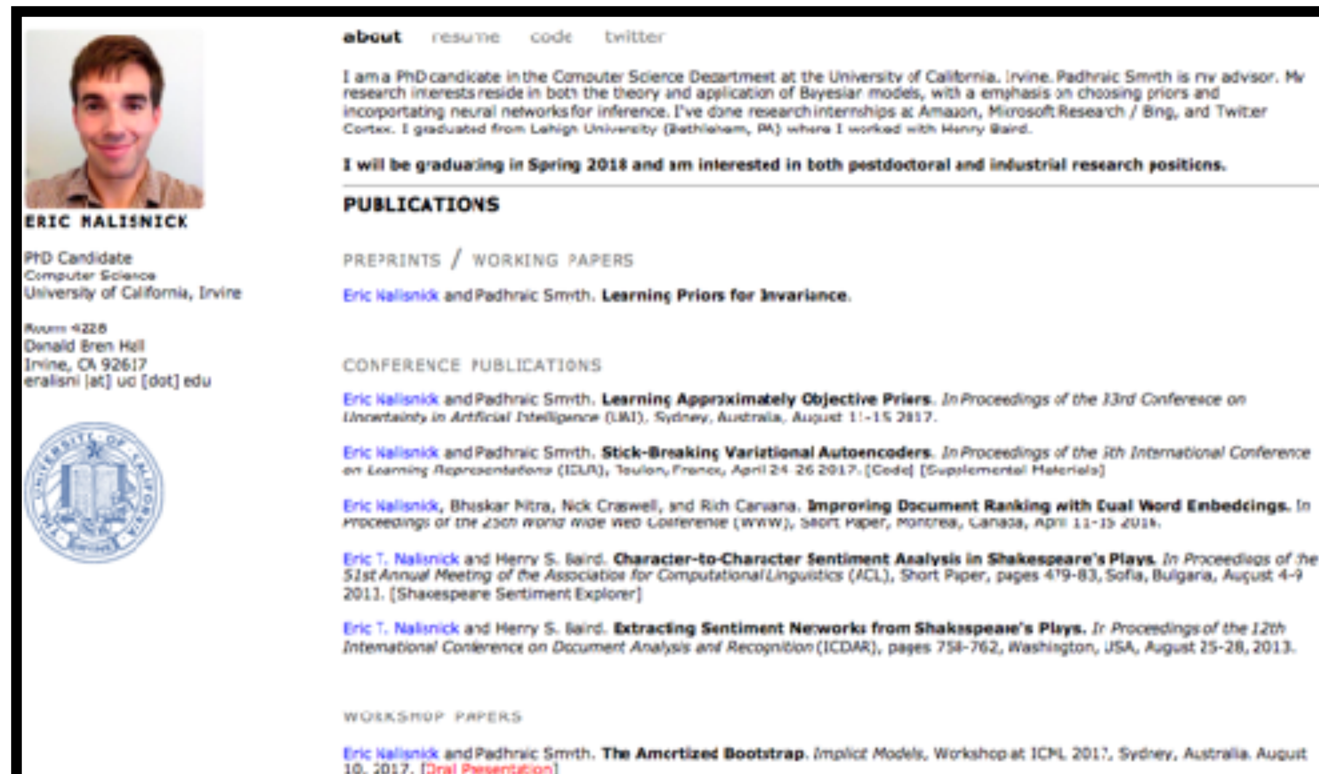
  Obtain data-driven posteriors for complex, formerly intractable models.

- **Amortized bootstrap:** model-based approximation of the bootstrap distribution.

  Results in **superior bagging** performance due (ostensibly) to smoothing and amortization.

Thank you. Questions?

See me at posters #6 and #9.



http://www.ics.uci.edu/~enalisni/

In collaboration with

Padhraic Smyth

References

1. Balan, Anoop Korattikara, et al. "Bayesian dark knowledge." *Advances in Neural Information Processing Systems*. 2015.
2. Bernardo, Jose M. "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society. Series B (Methodological)* (1979): 113-147.
3. Blundell, Charles, et al. "Weight Uncertainty in Neural Network." *International Conference on Machine Learning*. 2015.
4. Denil, Misha, et al. "Predicting parameters in deep learning." *Advances in Neural Information Processing Systems*. 2013.
5. Efron, B. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7.1 (1979): 1-26.
6. Efron, Bradley, and Robert Tibshirani. "Improvements on cross-validation: the 632+ bootstrap method." *Journal of the American Statistical Association* 92.438 (1997): 548-560.
7. Gal, Yarin, and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. 2016.
8. Hernández-Lobato, José Miguel, and Ryan Adams. "Probabilistic backpropagation for scalable learning of bayesian neural networks." *International Conference on Machine Learning*. 2015.
9. Hoffman, Matthew D., and Matthew J. Johnson. "Elbo surgery: yet another way to carve up the variational evidence lower bound." *Workshop in Advances in Approximate Bayesian Inference, NIPS*. 2016.
10. Johnson, Matthew, et al. "Composing graphical models with neural networks for structured representations and fast inference." *Advances in neural information processing systems*. 2016.
11. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *International Conference on Learning Representations (ICLR)* (2014).
12. Kingma, Diederik P., et al. "Improved variational inference with inverse autoregressive flow." *Advances in Neural Information Processing Systems*. 2016.
13. Liu, Qiang, and Dilin Wang. "Stein variational gradient descent: A general purpose bayesian inference algorithm." *Advances In Neural Information Processing Systems*. 2016.
14. Louizos, Christos, and Max Welling. "Structured and efficient variational deep learning with matrix Gaussian posteriors." *International Conference on Machine Learning*. 2016.
15. Mescheder, Lars, Sebastian Nowozin, and Andreas Geiger. "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks." *Proceedings of the 34st International Conference on Machine Learning (ICML-17)*. 2017.
16. Nalisnick, Eric, and Padhraic Smyth. "Learning Approximately Objective Priors.” *Proceedings of the 33rd International Conference on Uncertainty in Artificial Intelligence.* 2017.
17. Ranganath, Rajesh, et al. “Operator variational inference." *Advances in Neural Information Processing Systems*. 2016.
18. Rezende, Danilo, and Shakir Mohamed. "Variational Inference with Normalizing Flows." *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 2015.
19. Rezende, Danilo J., Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models." *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 2014.
20. Saatchi, Yunus, and Andrew Gordon Wilson. "Bayesian GAN." *Advances in neural information processing systems*. 2017.
21. Salimans, Tim, Diederik Kingma, and Max Welling. "Markov chain monte carlo and variational inference: Bridging the gap." *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 2015.
22. Salimans, Tim, and David A. Knowles. "Fixed-form variational posterior approximation through stochastic linear regression." *Bayesian Analysis* 8.4 (2013): 837-882.