The Amortized Bootstrap

Eric Nalisnick
University of California, Irvine

In collaboration with

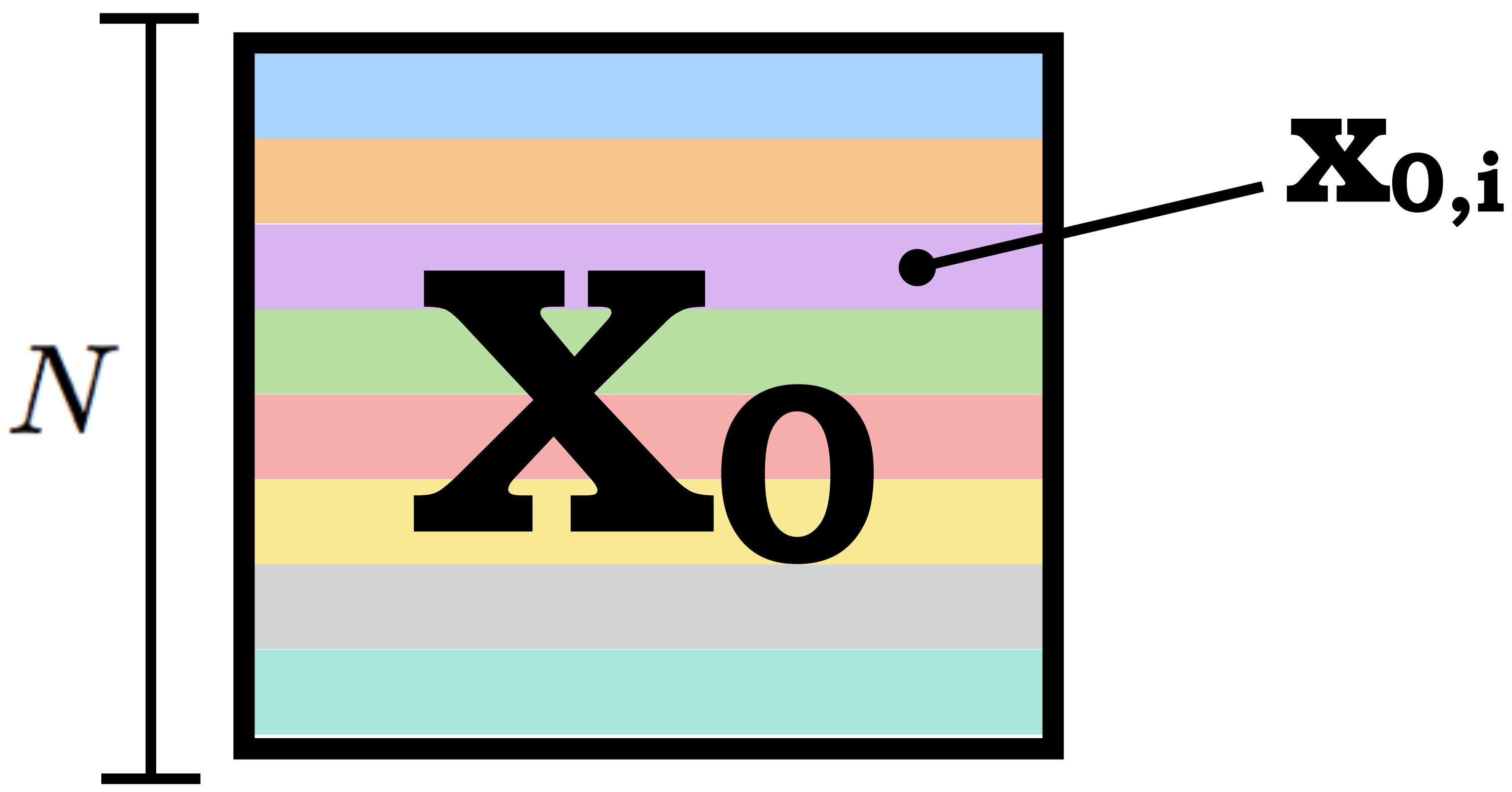
Padhraic Smyth

The Bootstrap

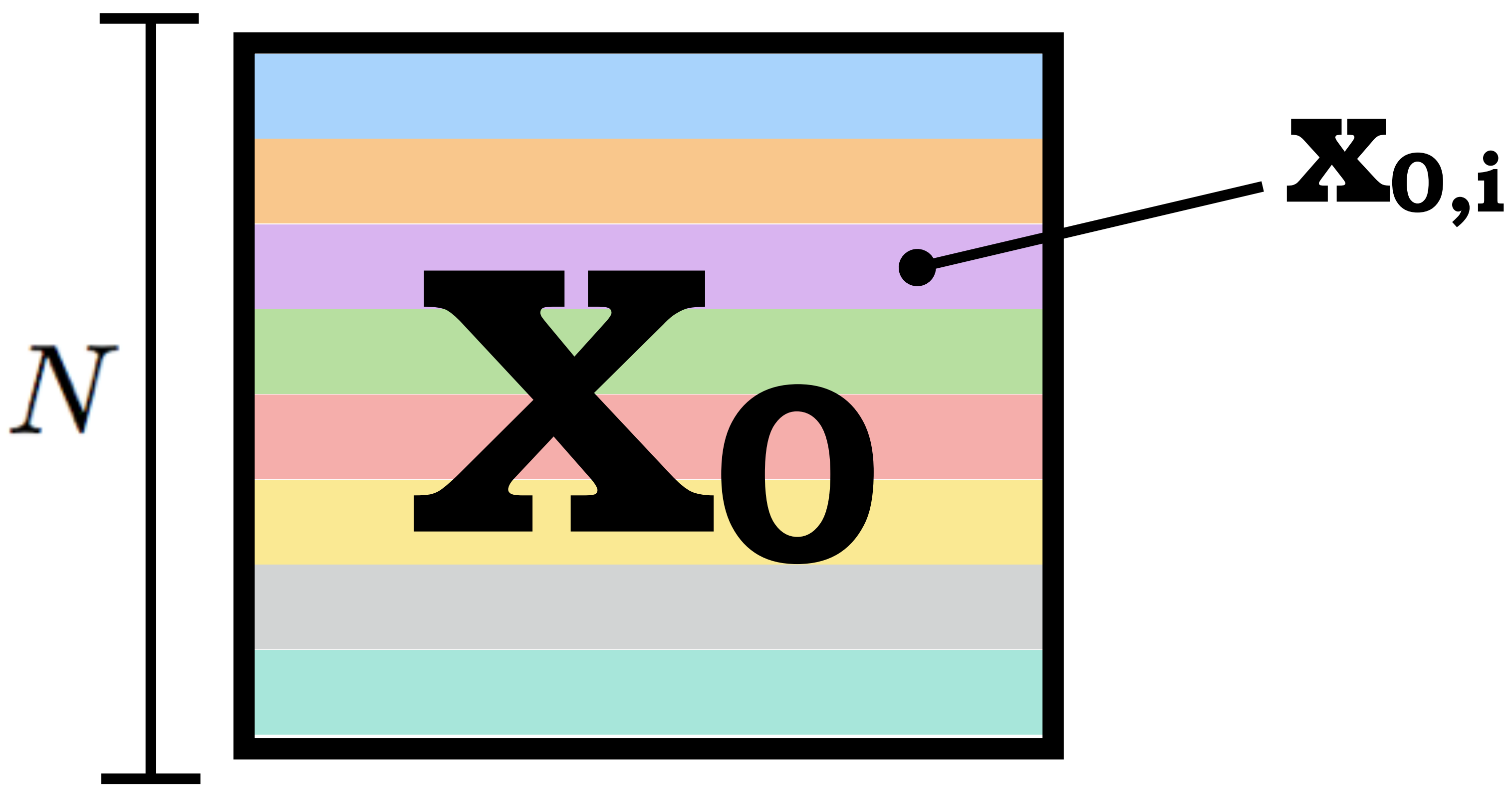
Bootstrap Resampling

$$\boxed{\mathbf{x}_0}$$

Bootstrap Resampling

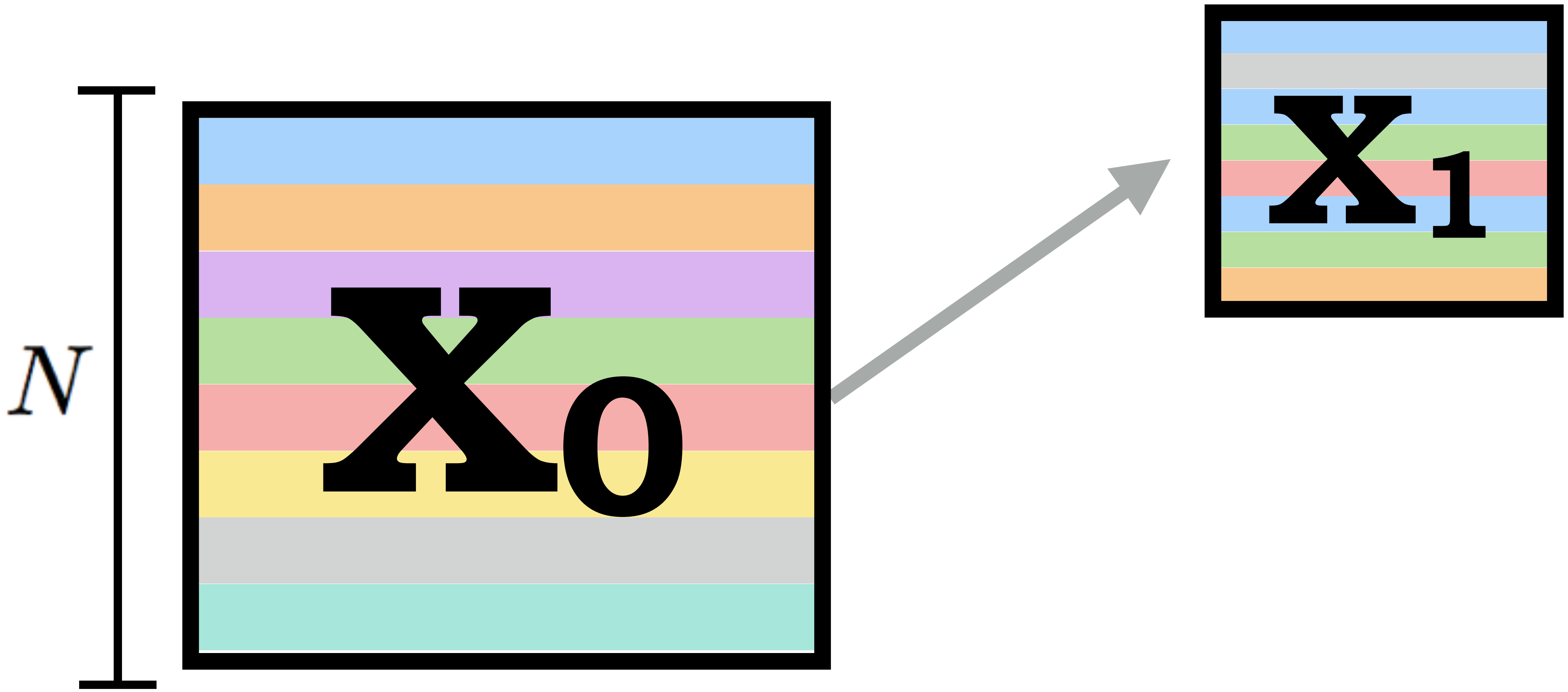


Bootstrap Resampling



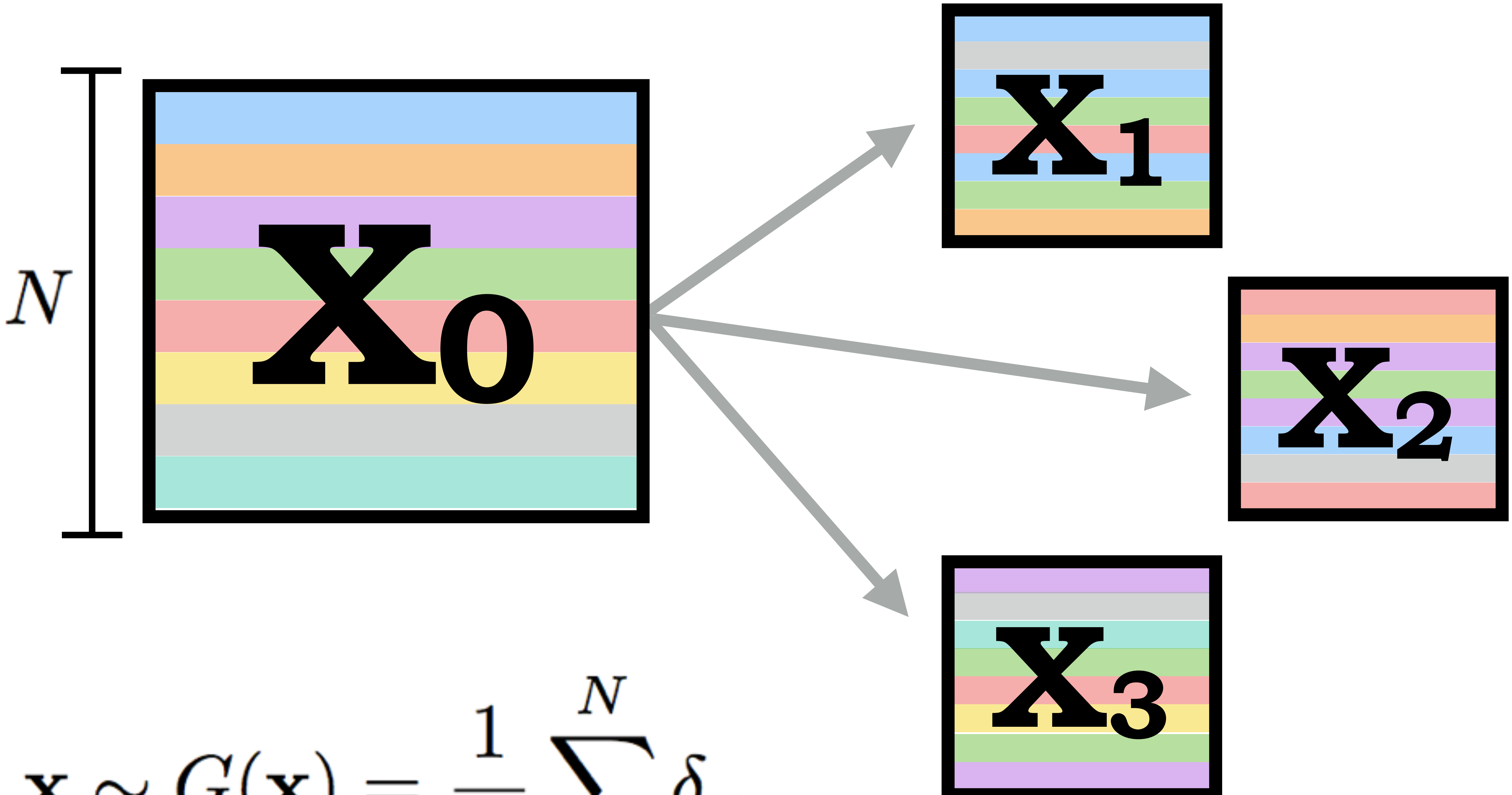
$$\mathbf{x} \sim G(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_{0,i}}$$

Bootstrap Resampling



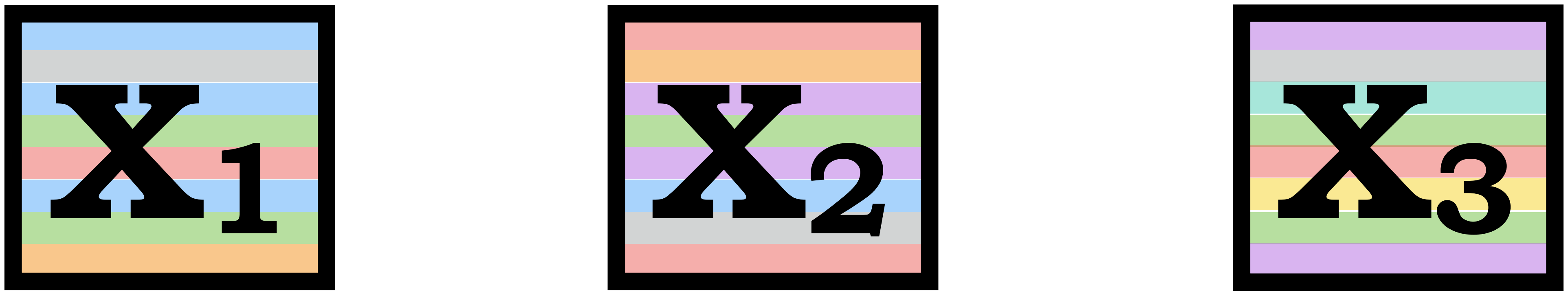
$$\mathbf{x} \sim G(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_{0,i}}$$

Bootstrap Resampling

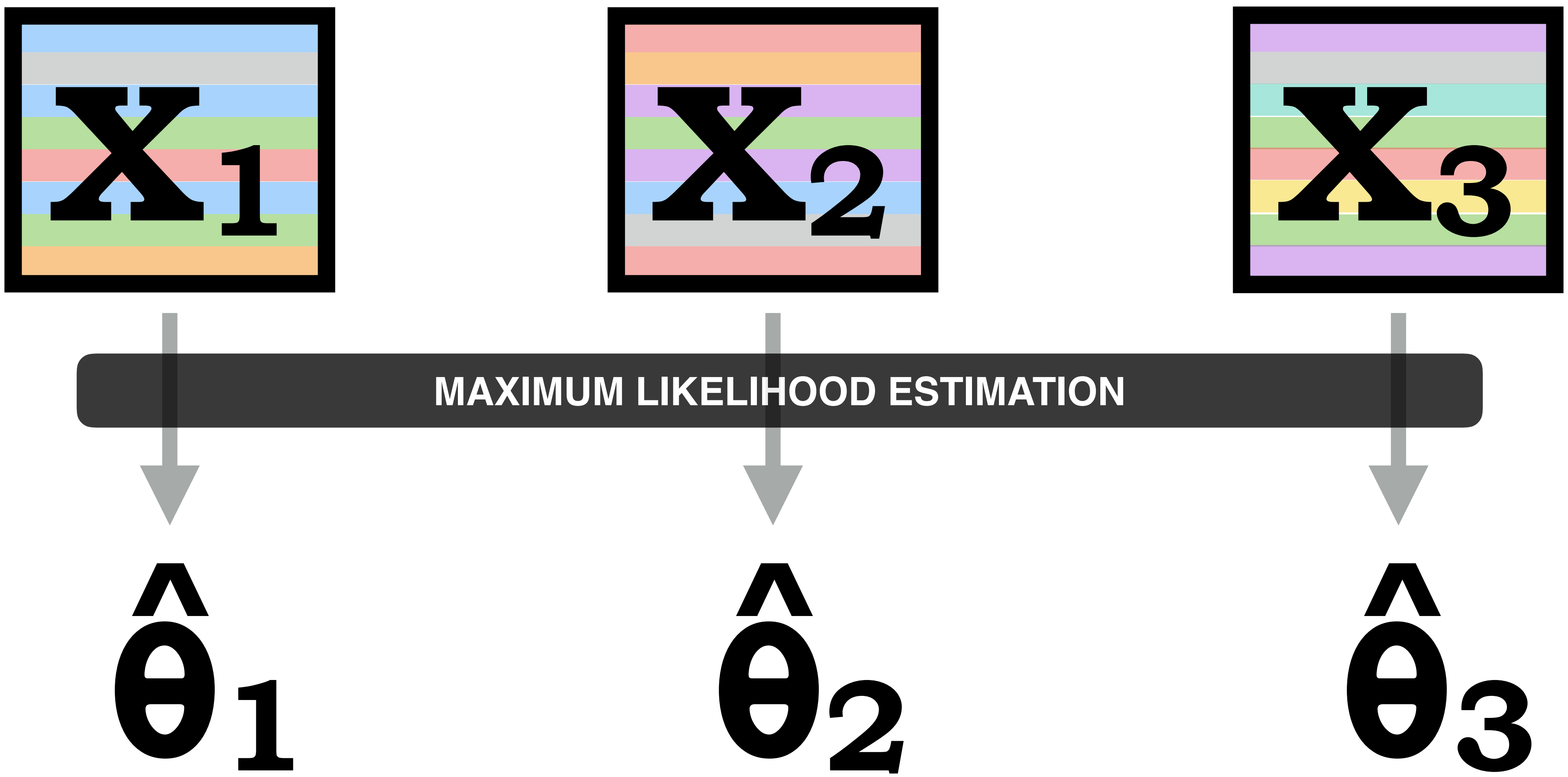


$$\mathbf{x} \sim G(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathbf{x}_{0,i}}$$

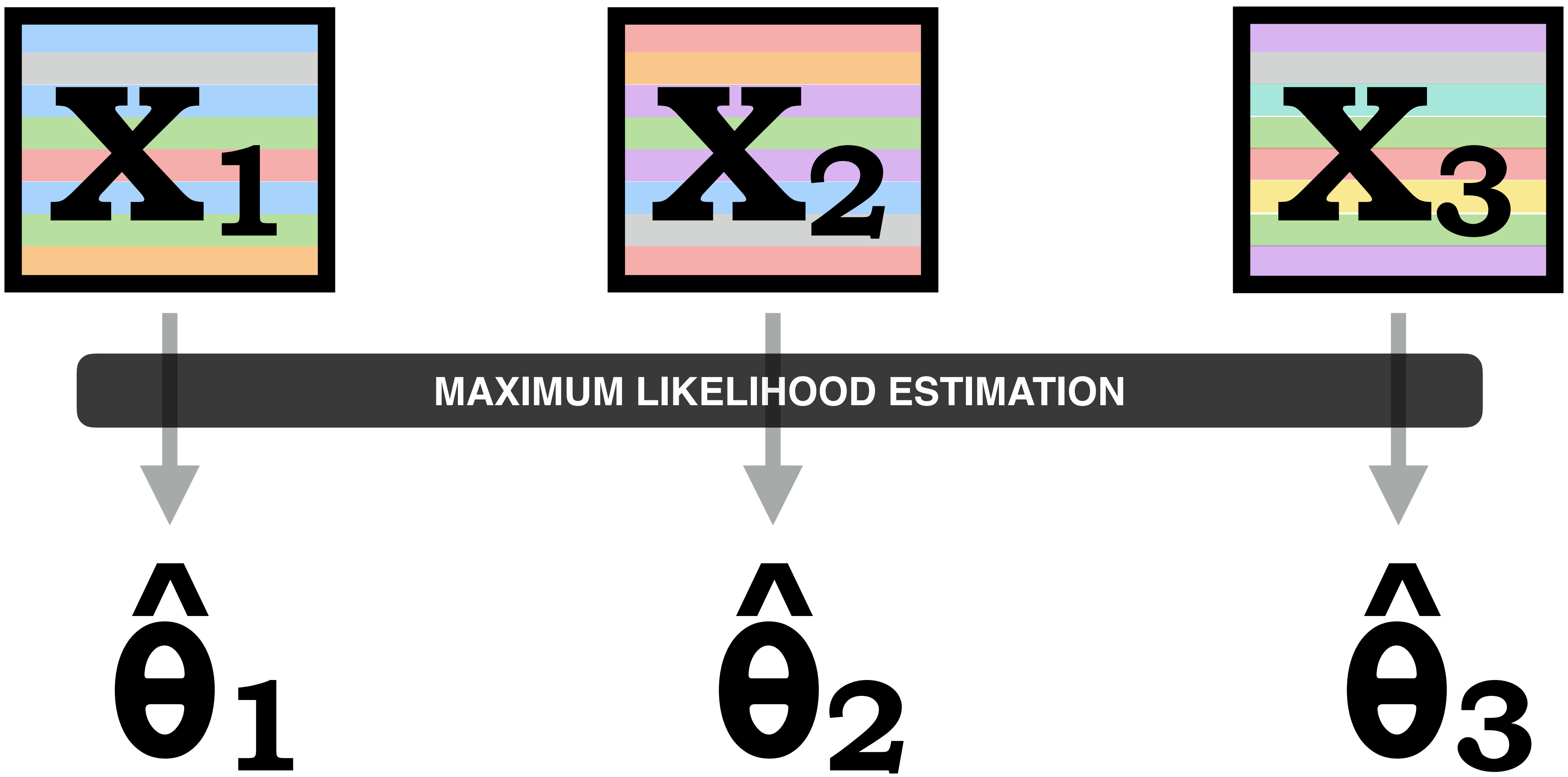
Bootstrap Distribution



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 $F(\boldsymbol{\theta})$ 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**).

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

Modeling the Bootstrap Distribution

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

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

Modeling the Bootstrap Distribution

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

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

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

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

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

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

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

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

Learning the Bootstrap Model

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

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

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

Learning the Bootstrap Model

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

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

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

Learning the Bootstrap Model

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

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

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.

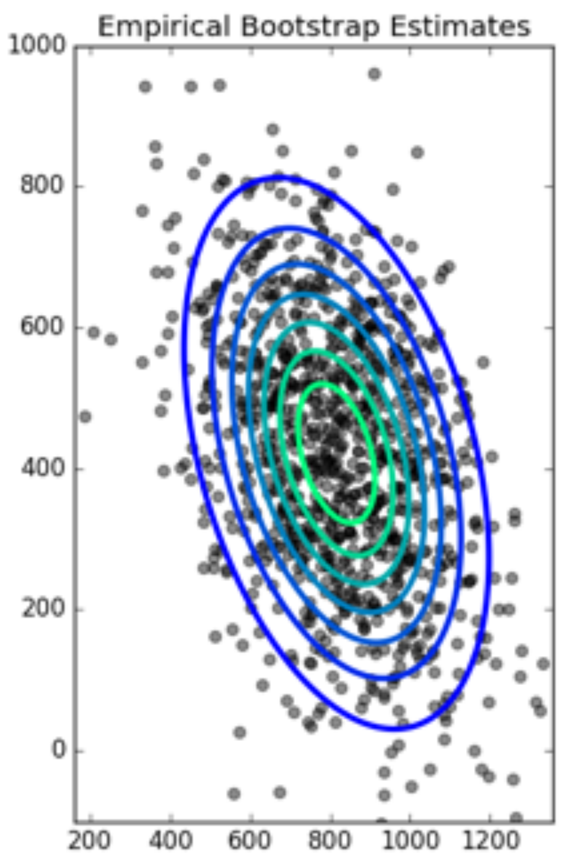
$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\mathbf{X}|\boldsymbol{\theta})] - \mathbf{KLD}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta})]$$

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

Experiment #1: Sanity Check

Linear Regression

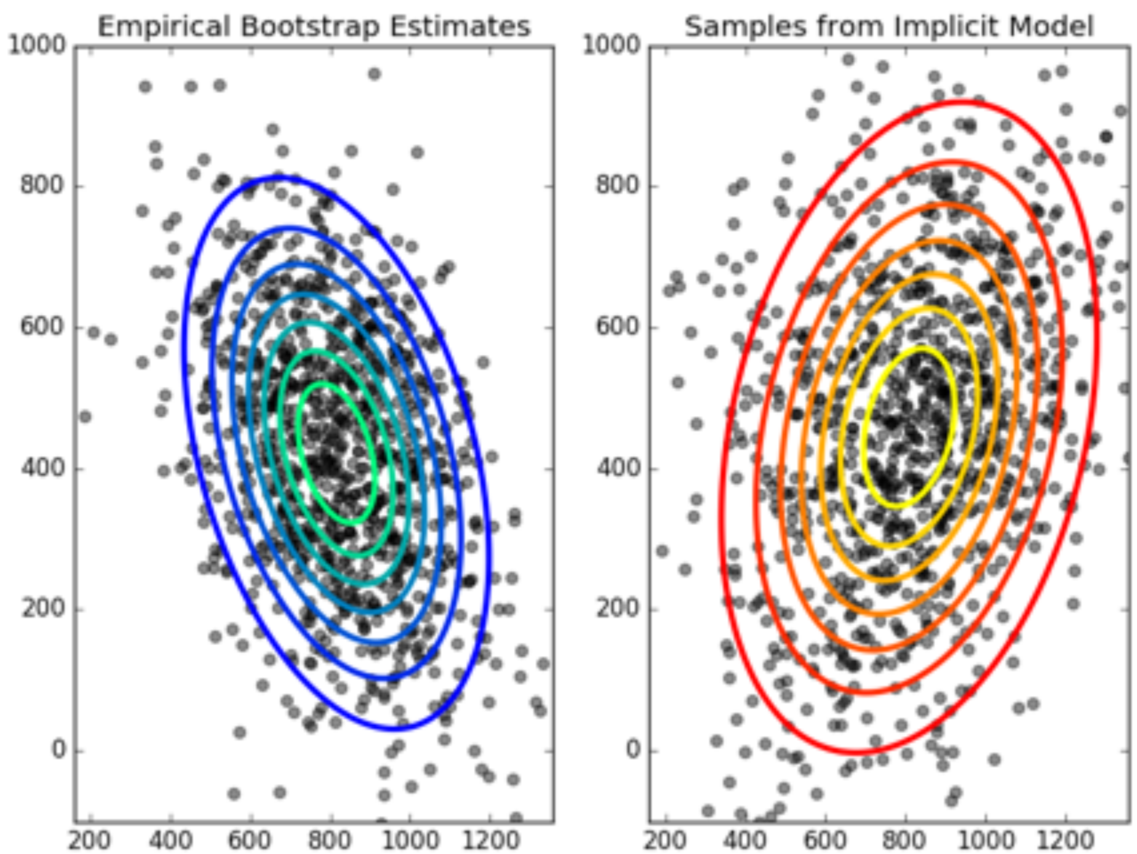
2D Diabetes Dataset



2D Boston Housing Dataset

Linear Regression

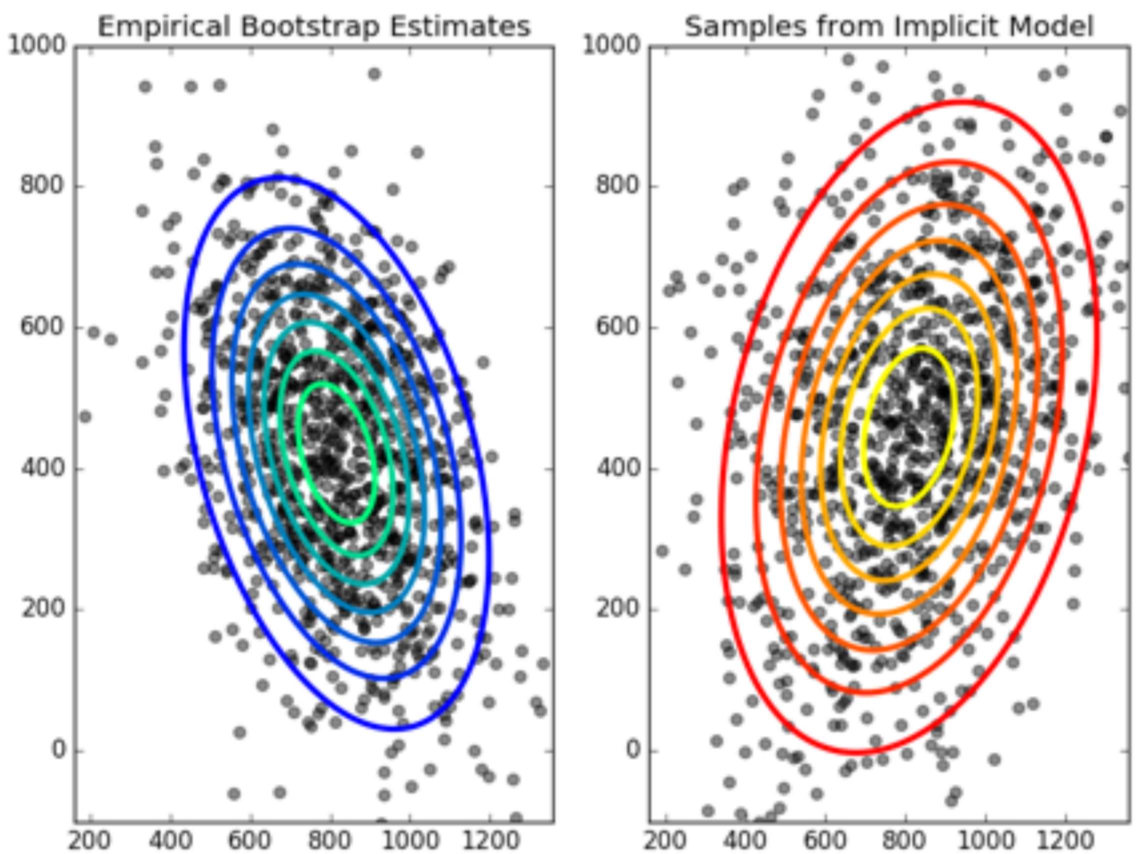
2D Diabetes Dataset



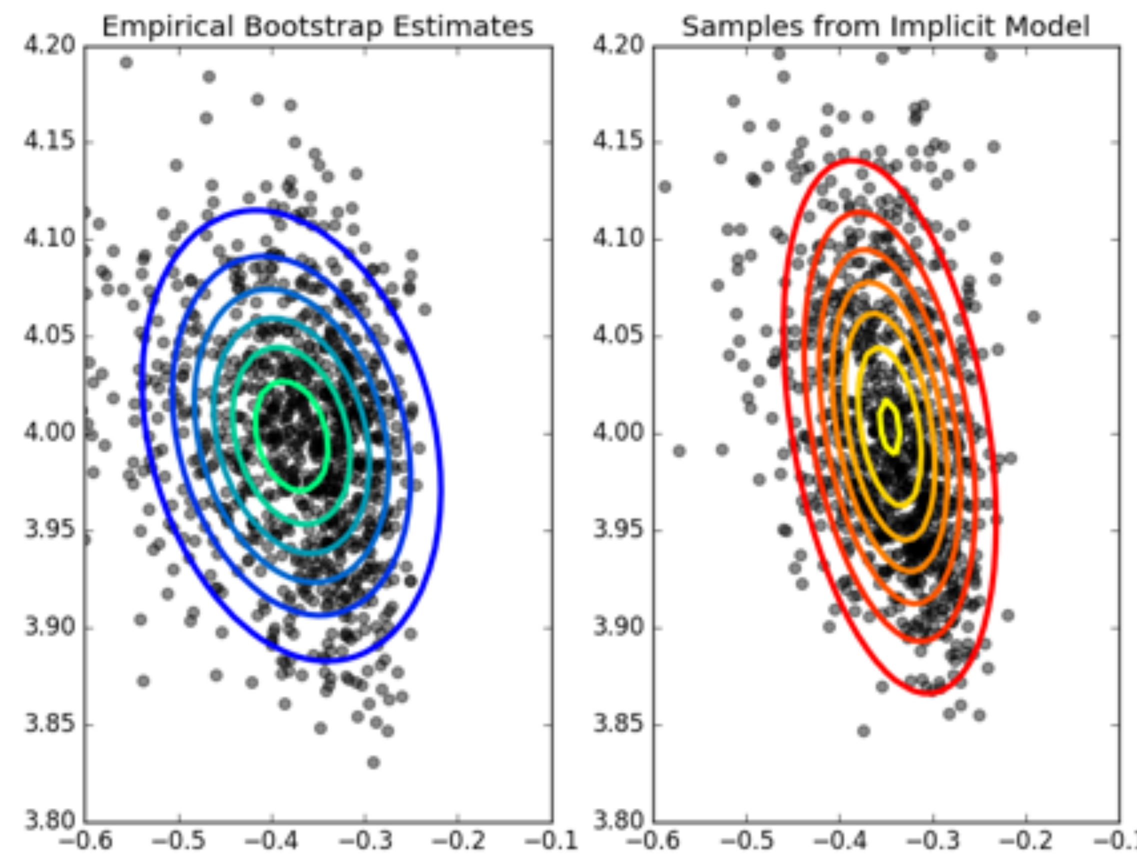
2D Boston Housing Dataset

Linear Regression

2D Diabetes Dataset

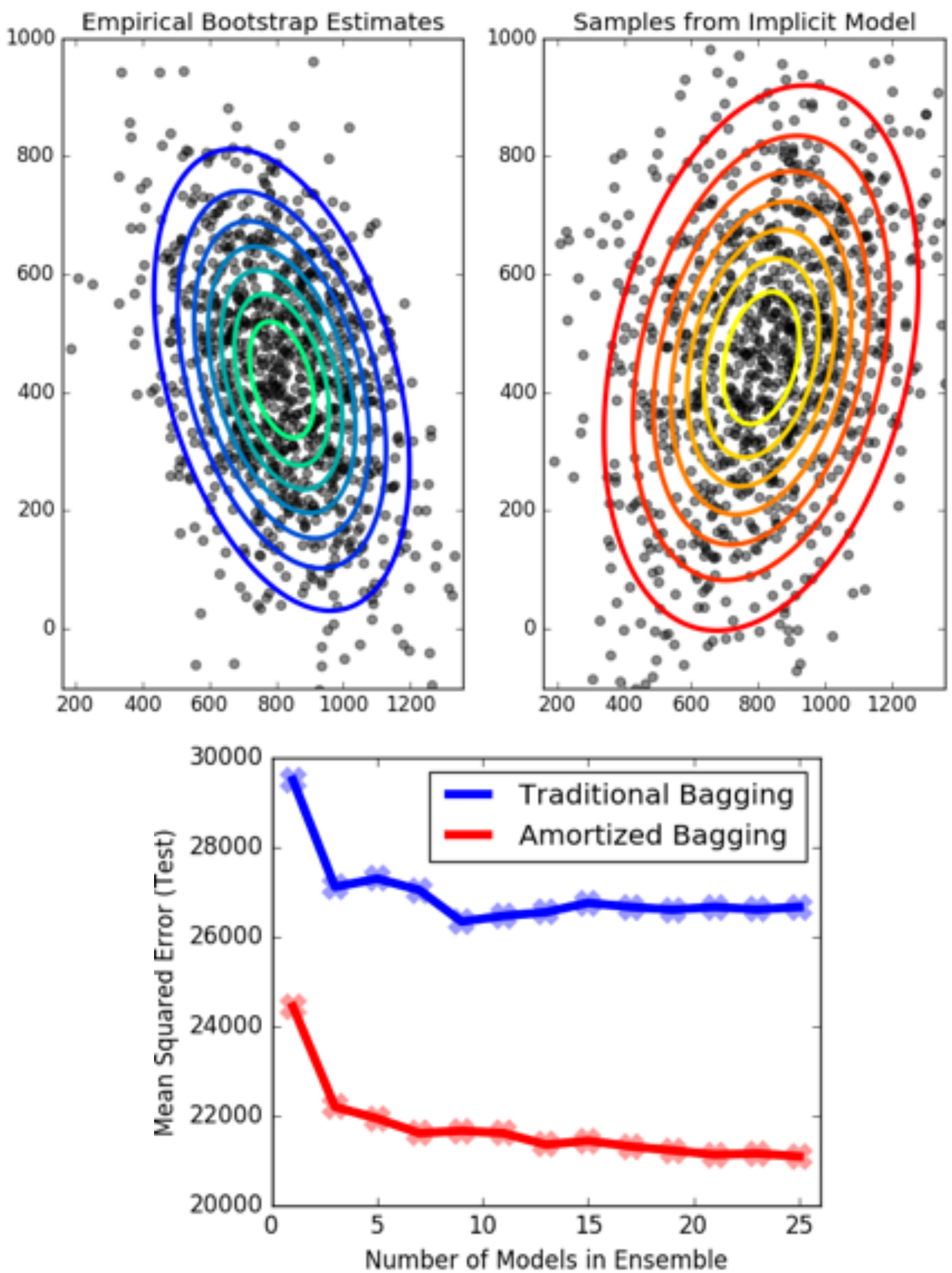


2D Boston Housing Dataset

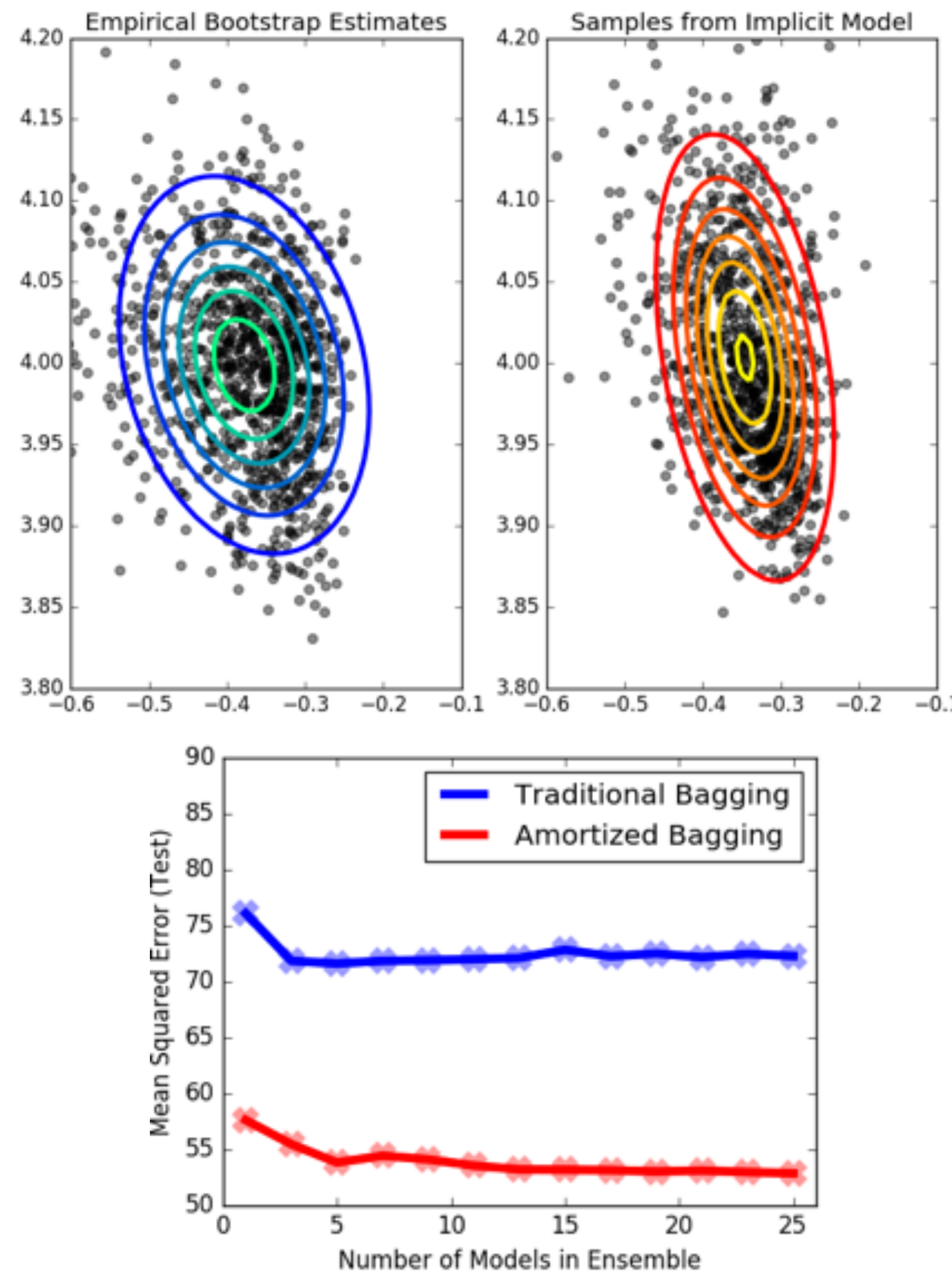


Linear Regression

2D Diabetes Dataset

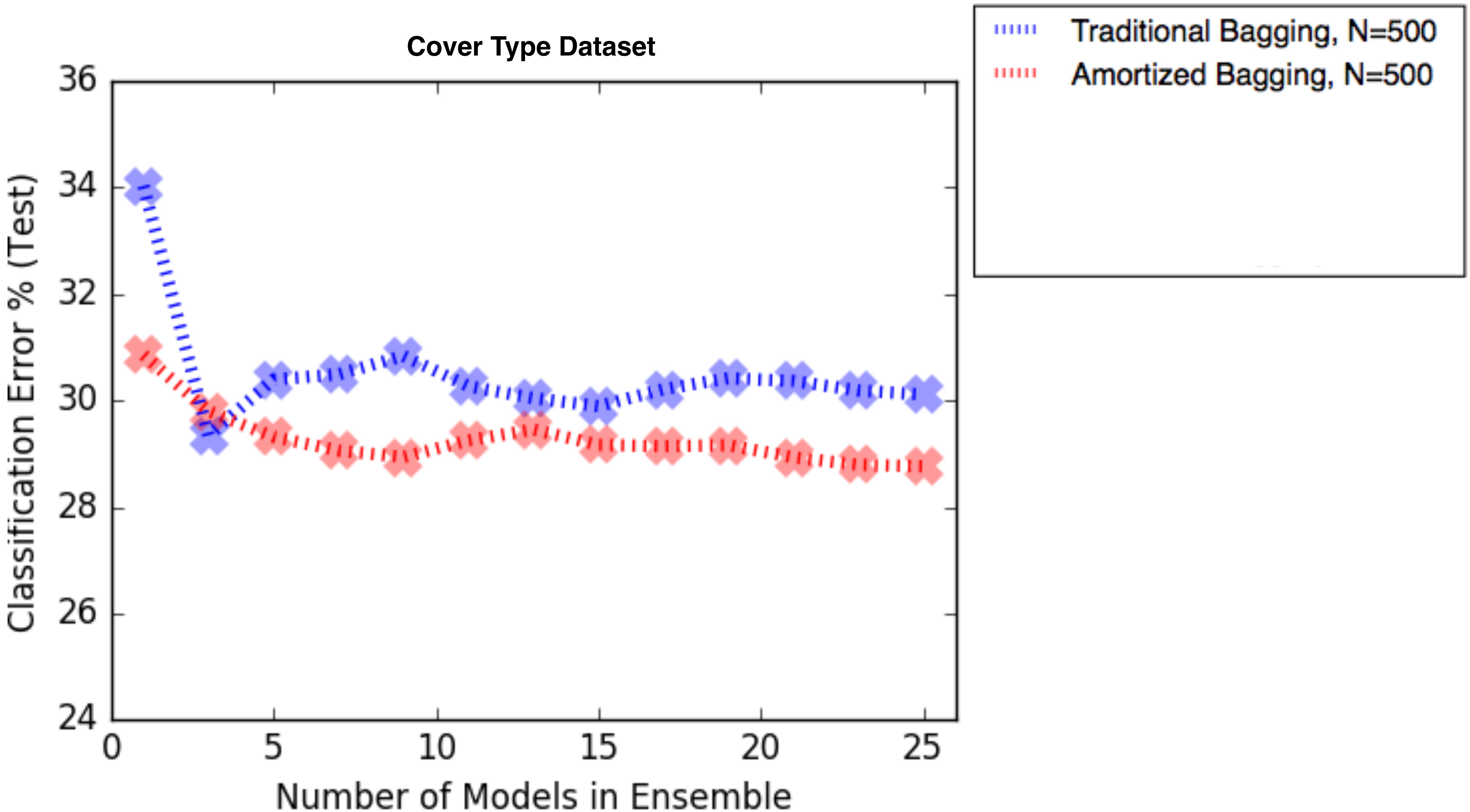


2D Boston Housing Dataset

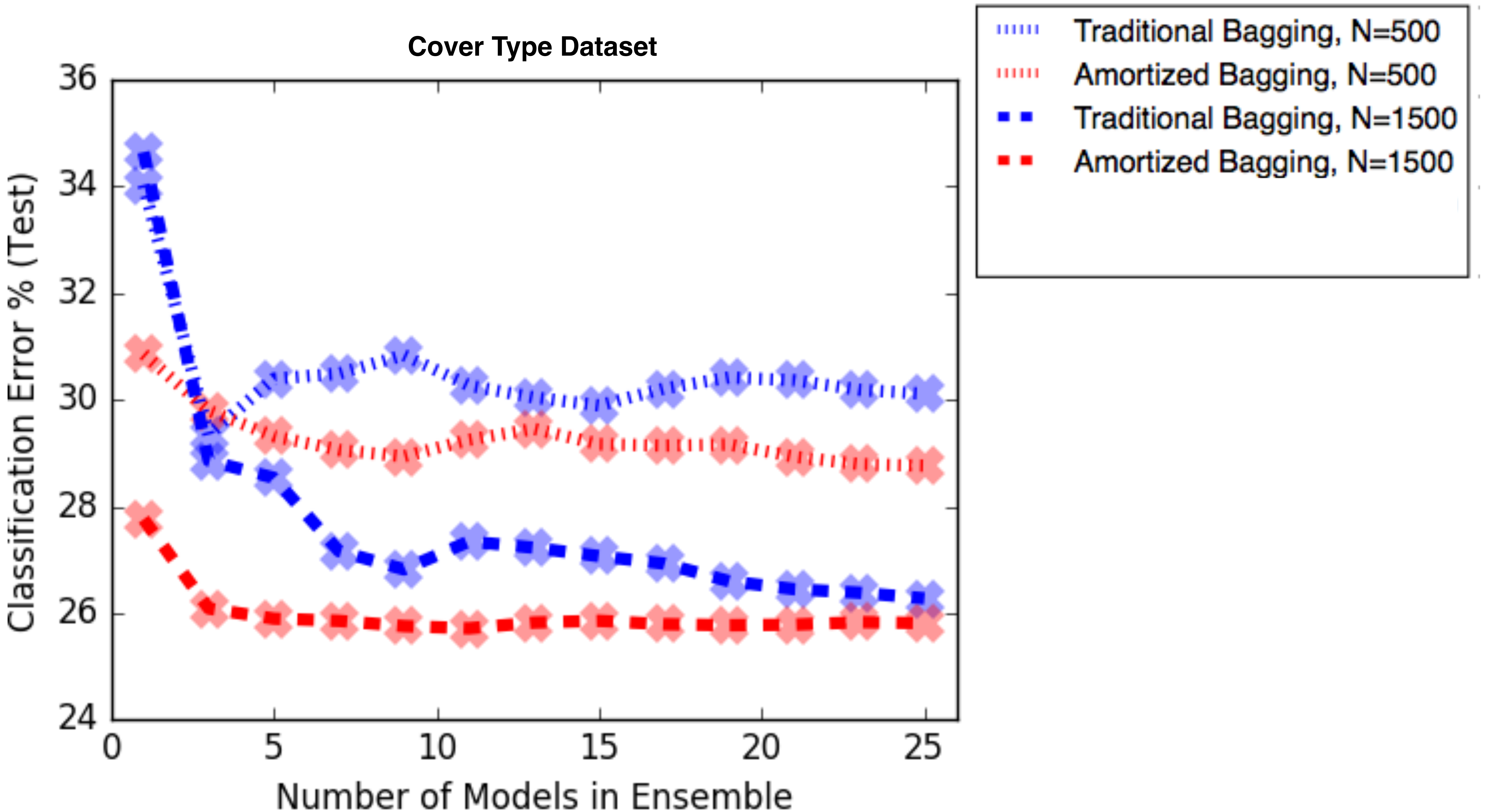


Experiment #2: Varying Dataset Size

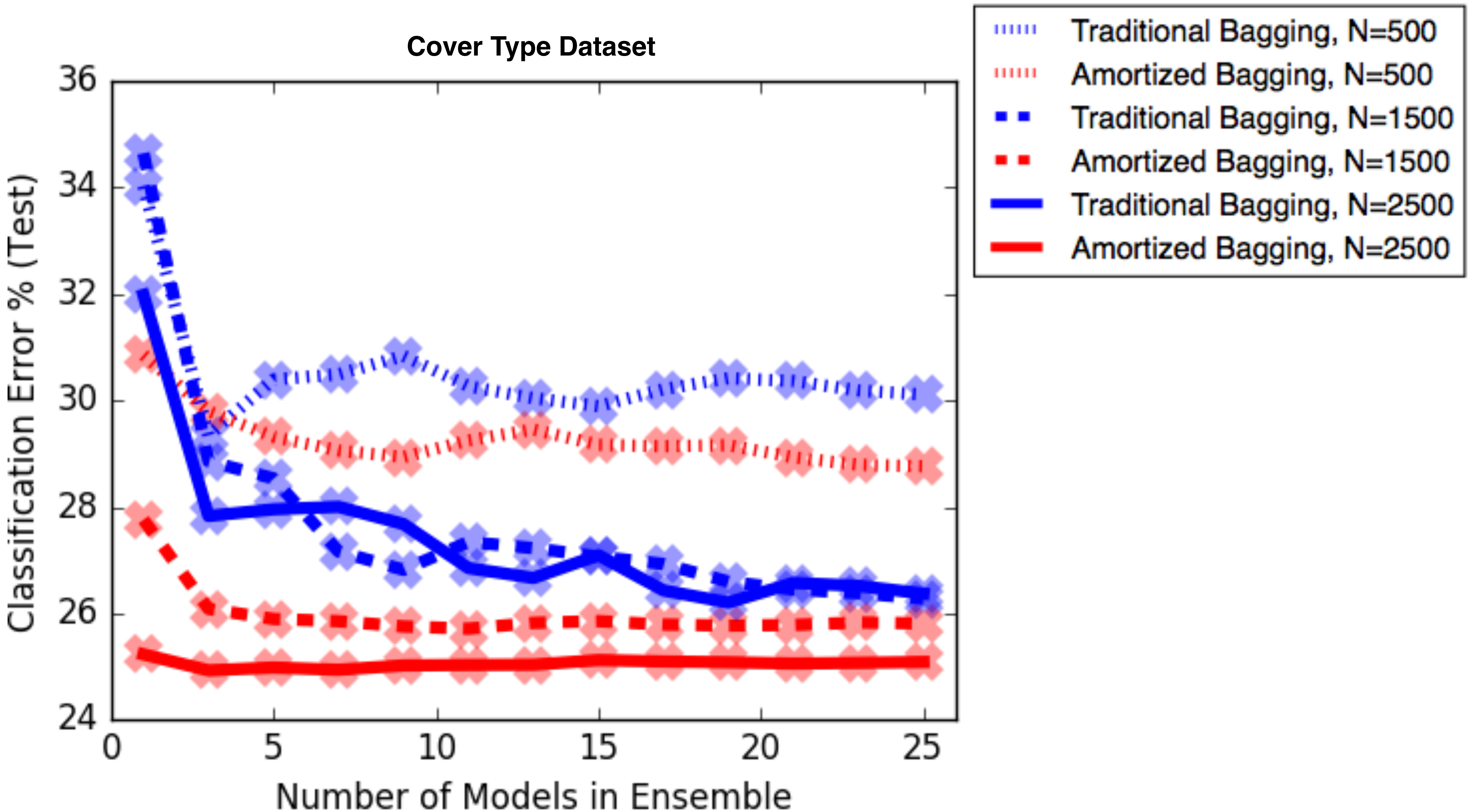
Logistic Regression



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**

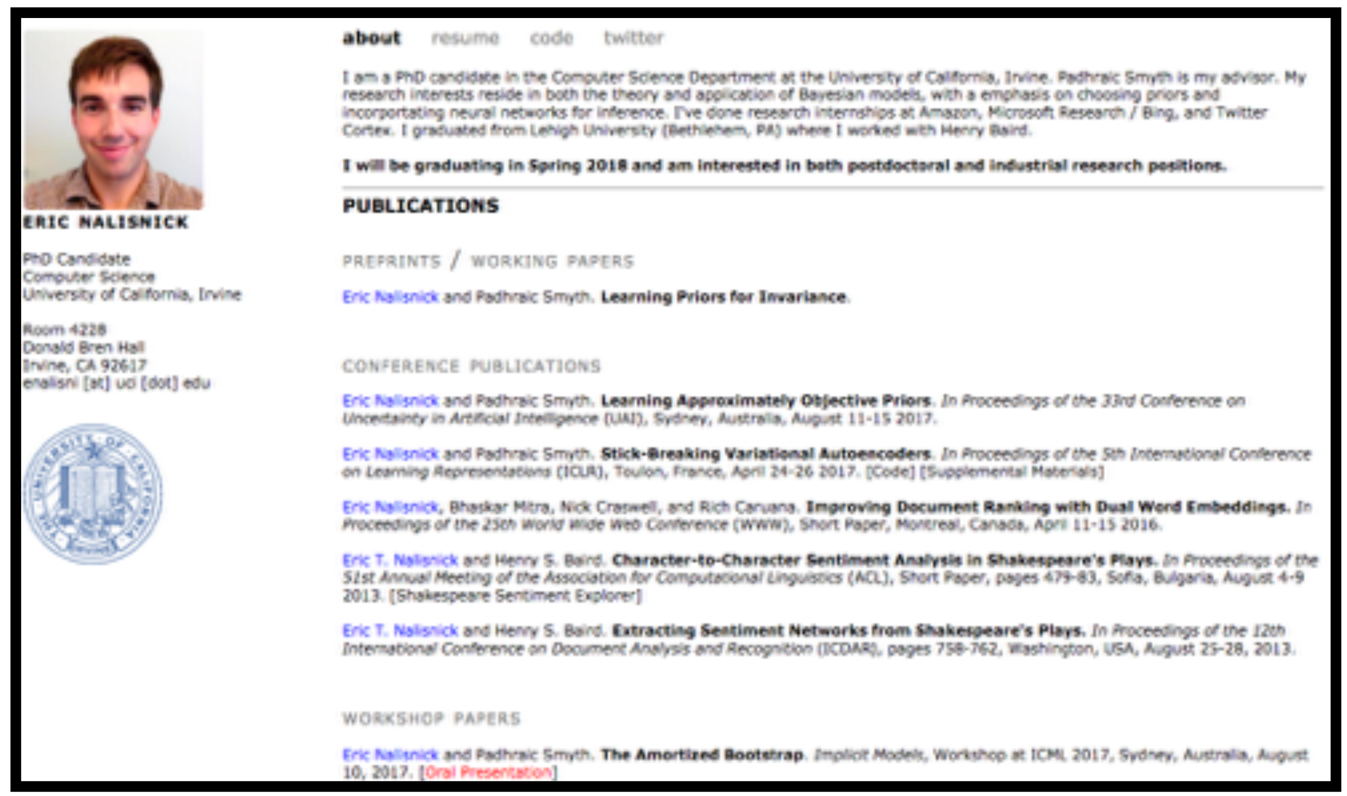
Rotated MNIST Dataset

Conclusions

- Model-based bootstrap results in **superior bagging** performance due (ostensibly) to smoothing and amortization.

- **Future work:** larger-scale experiments, theoretical analysis, uncertainty quantification.

Thank you. Questions?



ERIC NALISNICK

PhD Candidate
Computer Science
University of California, Irvine

Room 4228
Donald Bren Hall
Irvine, CA 92617
enalisni [at] uci [dot] edu

about resume code twitter

I am a PhD candidate in the Computer Science Department at the University of California, Irvine. Padhraic Smyth is my advisor. My research interests reside in both the theory and application of Bayesian models, with a emphasis on choosing priors and incorporating neural networks for inference. I've done research internships at Amazon, Microsoft Research / Bing, and Twitter Cortex. I graduated from Lehigh University (Bethlehem, PA) where I worked with Henry Baird.

I will be graduating in Spring 2018 and am interested in both postdoctoral and industrial research positions.

PUBLICATIONS

PREPRINTS / WORKING PAPERS

Eric Nalisnick and Padhraic Smyth. **Learning Priors for Invariance**.

CONFERENCE PUBLICATIONS

Eric Nalisnick and Padhraic Smyth. **Learning Approximately Objective Priors**. *In Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence* (UAI), Sydney, Australia, August 11-15 2017.

Eric Nalisnick and Padhraic Smyth. **Stick-Breaking Variational Autoencoders**. *In Proceedings of the 5th International Conference on Learning Representations* (ICLR), Toulon, France, April 24-26 2017. [Code] [Supplemental Materials]

Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. **Improving Document Ranking with Dual Word Embeddings.** *In Proceedings of the 25th World Wide Web Conference* (WWW), Short Paper, Montreal, Canada, April 11-15 2016.

Eric T. Nalisnick and Henry S. Baird. **Character-to-Character Sentiment Analysis in Shakespeare's Plays.** *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL), Short Paper, pages 479-83, Sofia, Bulgaria, August 4-9 2013. [Shakespeare Sentiment Explorer]

Eric T. Nalisnick and Henry S. Baird. **Extracting Sentiment Networks from Shakespeare's Plays.** *In Proceedings of the 12th International Conference on Document Analysis and Recognition* (ICDAR), pages 758-762, Washington, USA, August 25-28, 2013.

WORKSHOP PAPERS

Eric Nalisnick and Padhraic Smyth. **The Amortized Bootstrap**. *Implicit Models*, Workshop at ICML 2017, Sydney, Australia, August 10, 2017. [Oral Presentation]

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

<u>Acknowledgements</u>

Google DeepMind
Travel Award