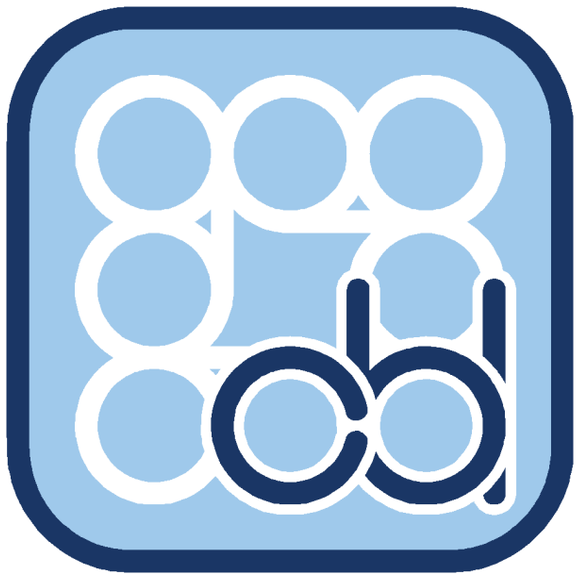
Machine Learning with Objective Priors

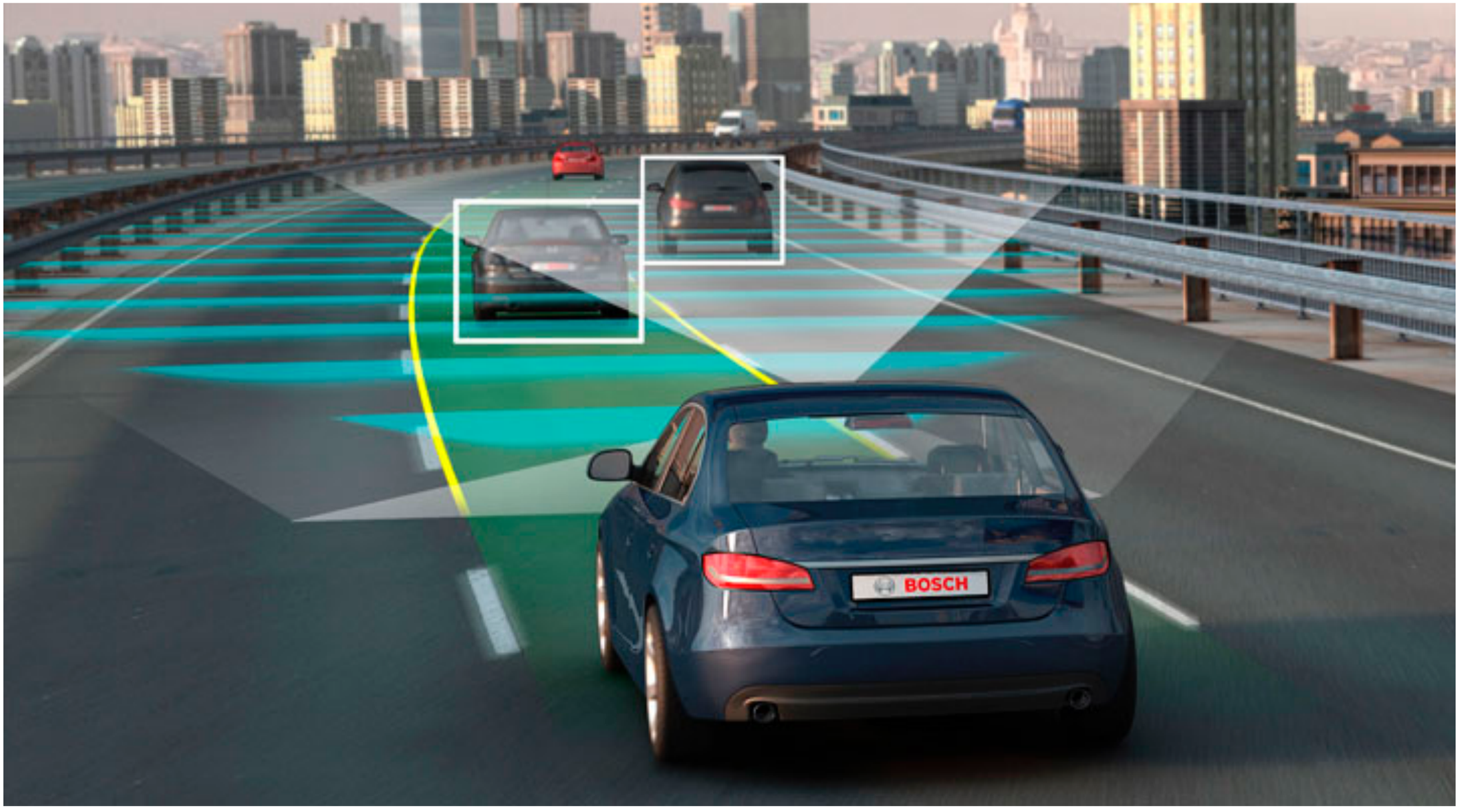
Eric Nalisnick



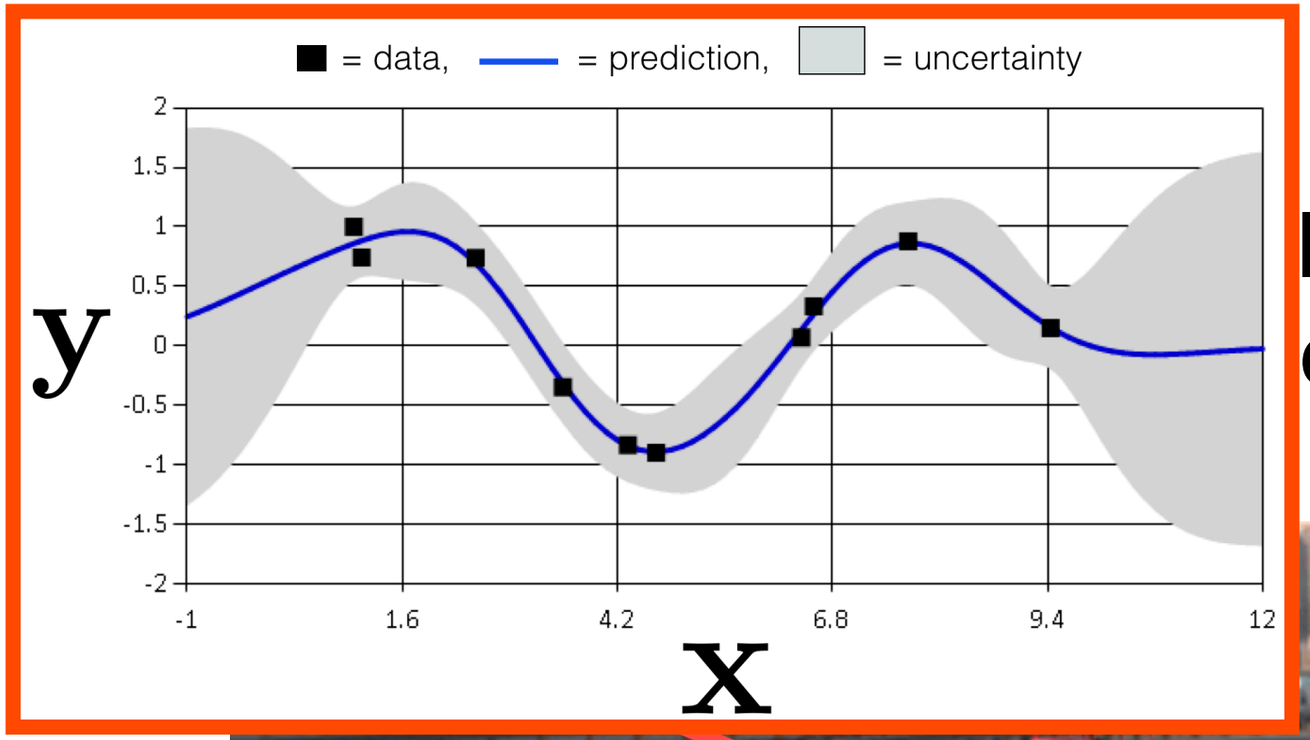
Computational and Biological Learning
University of Cambridge

Div F Conference
20.3.2019

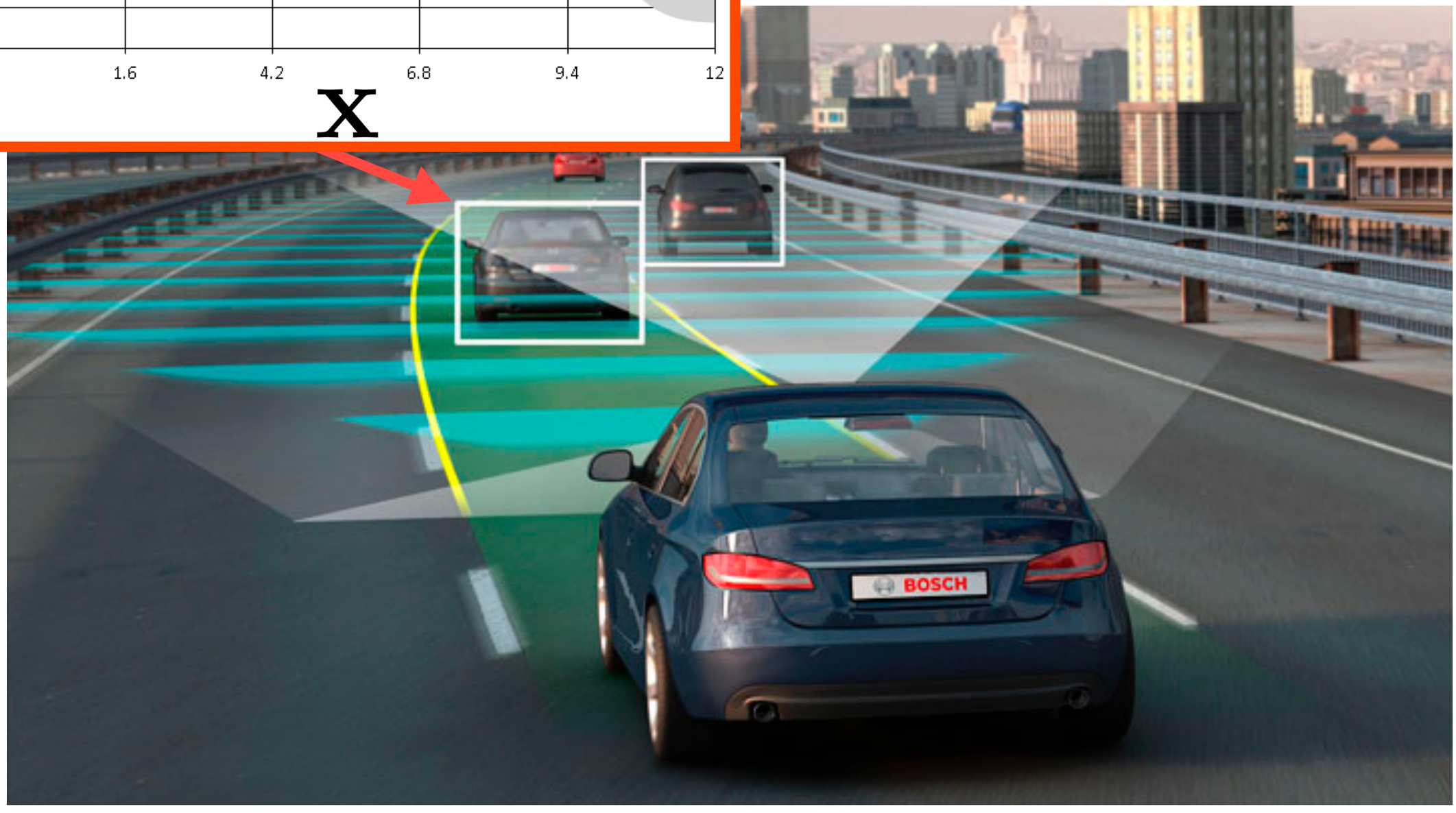
Modern machine learning requires uncertainty estimation.



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



learning requires estimation.



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

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})\ \ p(\boldsymbol{\theta})}{p(\mathbf{X})}$$

THOMAS BAYES

Likelihood

Prior

Posterior

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})\ p(\boldsymbol{\theta})}{p(\mathbf{X})}$$

Marginal Likelihood

(a.k.a. model evidence)
(a.k.a. normalizing constant)
(a.k.a. partition function)

$$p(\mathbf{X}) = \int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta})\ p(\boldsymbol{\theta})\ d\boldsymbol{\theta}$$

THOMAS BAYES

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})\ \ p(\boldsymbol{\theta})}{p(\mathbf{X})}$$

THOMAS BAYES

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathbf{X})}$$

Garbage in: arbitrary priors
Garbage out: uncontrollable error bars

~ Michael I. Jordan, *MLSS* (2017)

THOMAS BAYES

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathbf{X})}$$

How should we think about priors for machine learning models?

THOMAS BAYES

EXAMPLE

Priors for Modeling Coin Tosses

Priors for Modeling Coin Tosses

$$\mathbf{X} = \{X_n \in \{0, 1\}\}_{n=1}^{N}$$

$$p(\mathbf{X}|\theta) = \texttt{Binomial}(N, \theta)$$

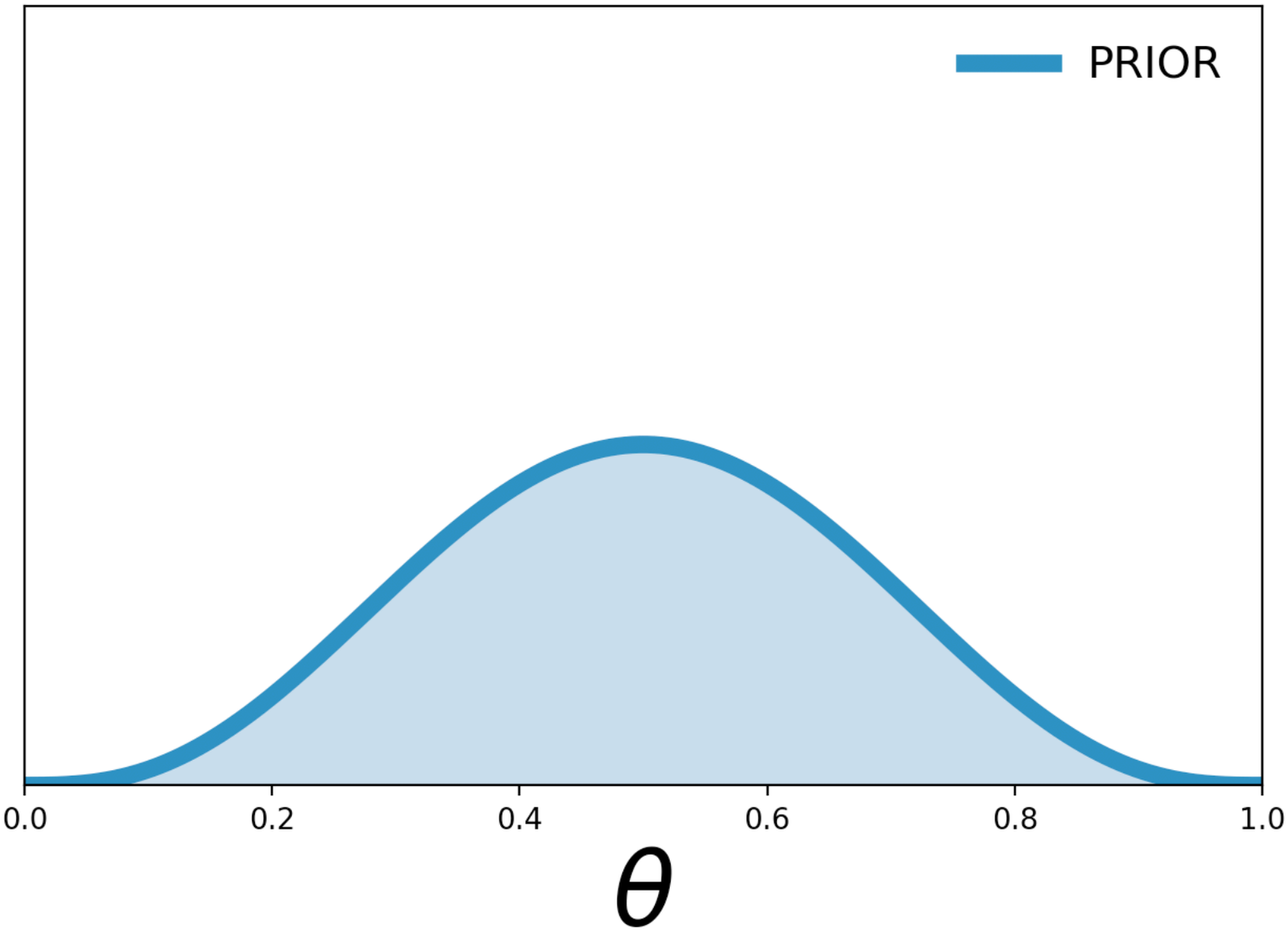
$$p(\theta) = \texttt{Beta}(\alpha_0, \beta_0)$$

Priors for Modeling Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(\alpha_0, \beta_0)$$

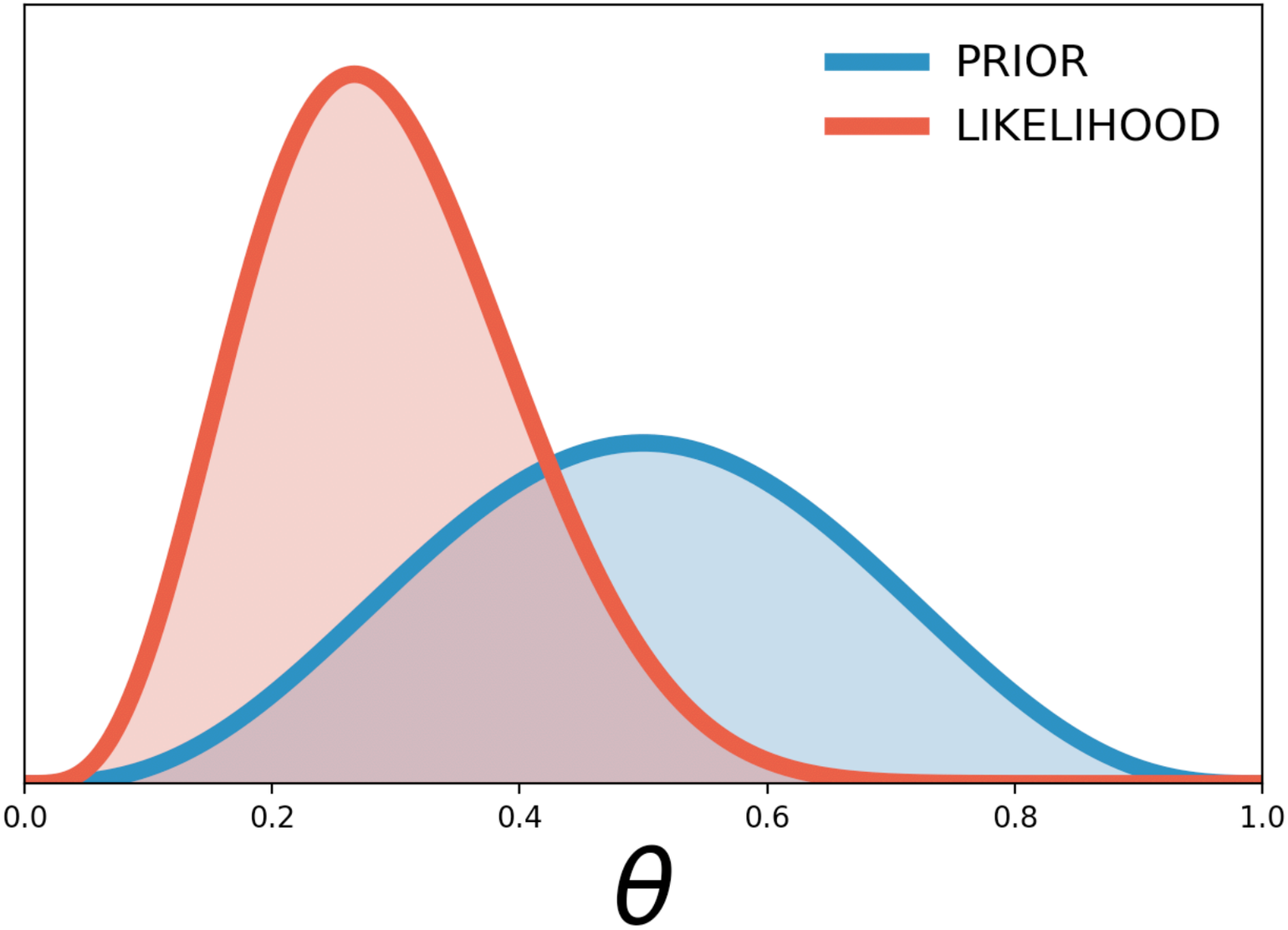
Priors for Modeling Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(4, 4)$$



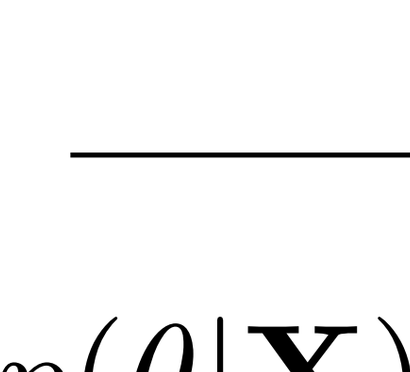
Priors for Modeling Coin Tosses

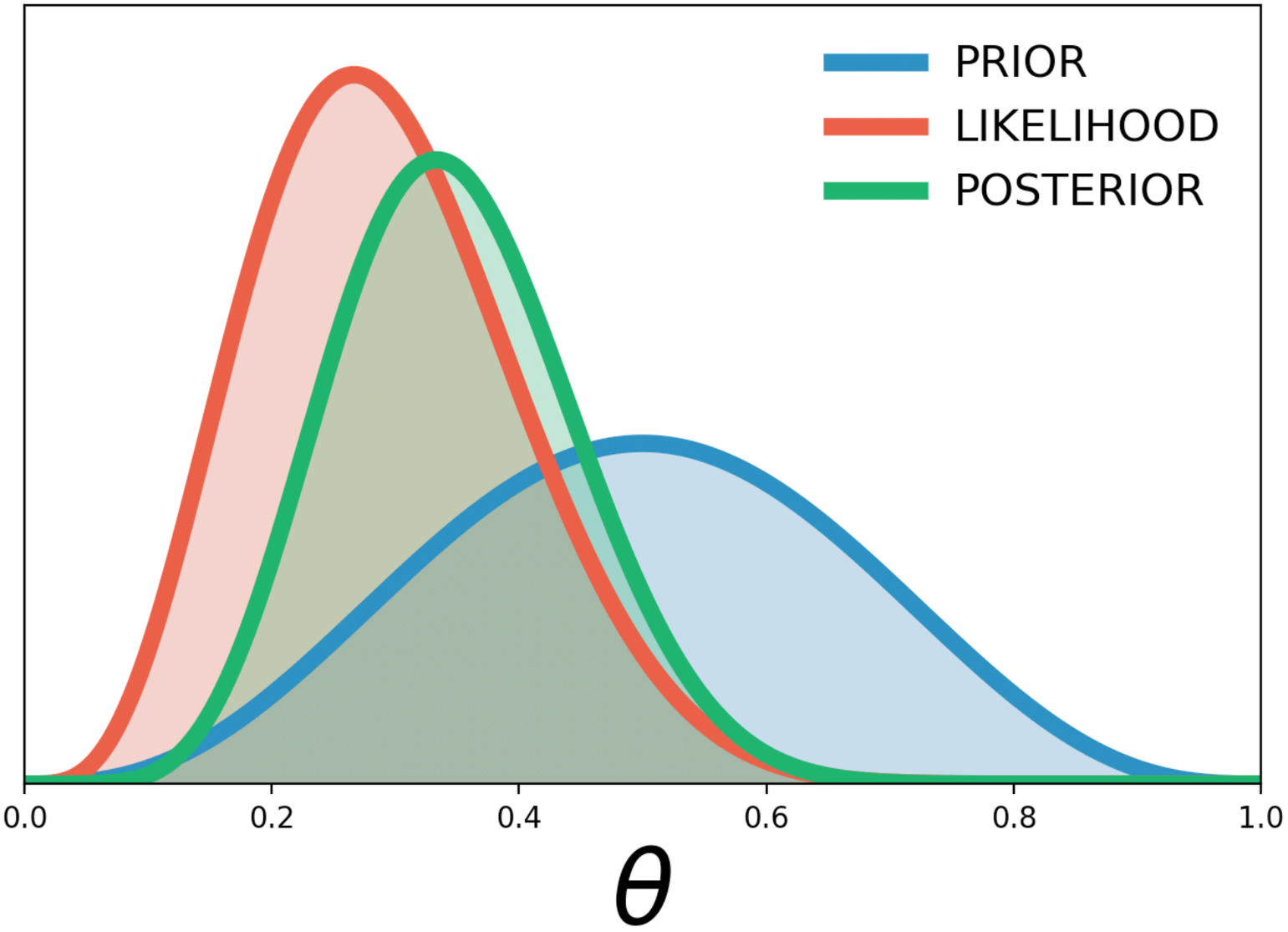
$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(15, \theta)\,\texttt{Beta}(4, 4)$$



Priors for Modeling Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(15, \theta)\,\texttt{Beta}(4, 4)$$





Priors for Modeling Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(\alpha_0, \beta_0)$$

…But what if we have *even stronger* prior information?

Priors for Modeling Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(\alpha_0, \beta_0)$$

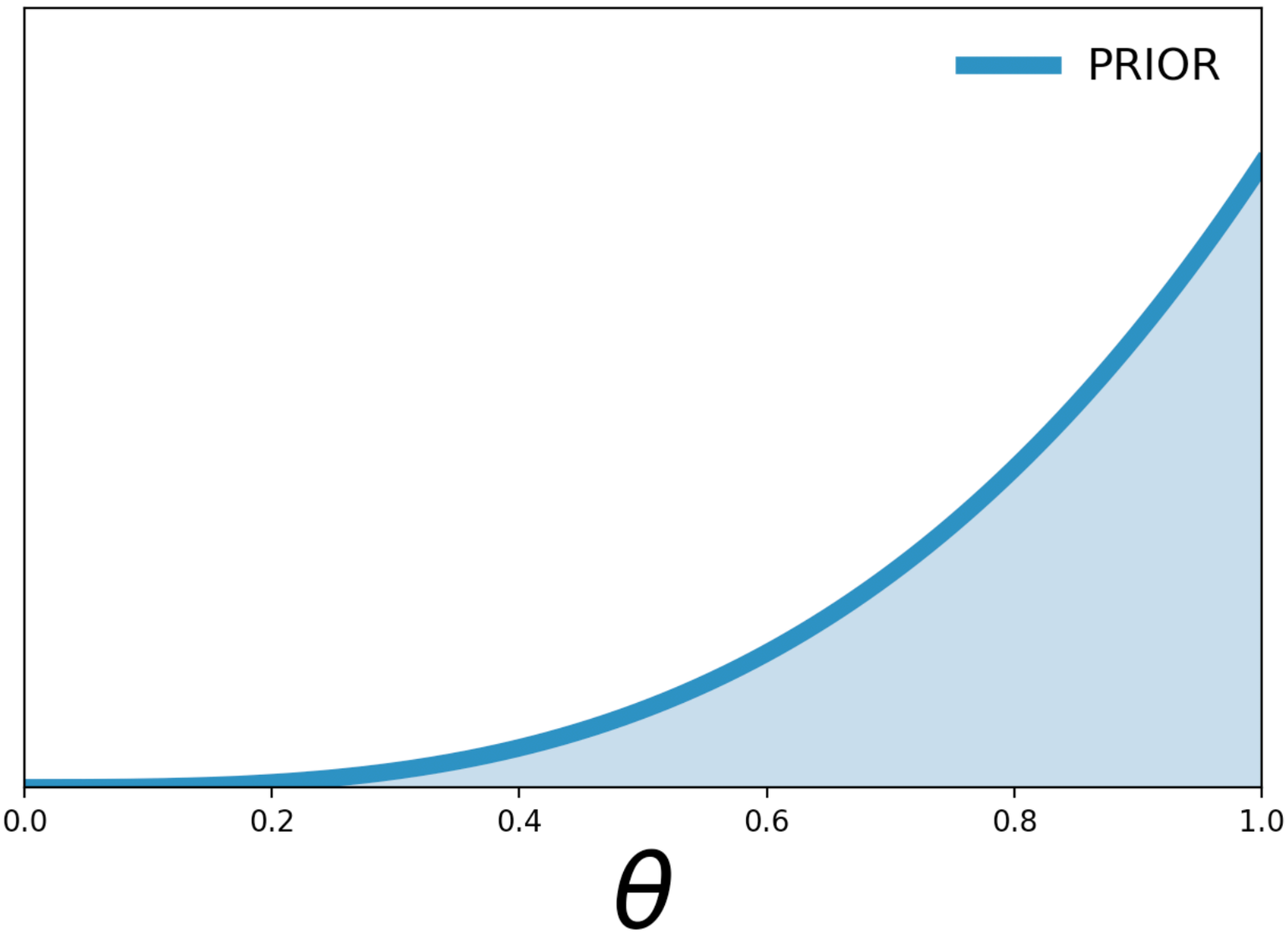
Harvey Dent
aka “Two-Face”
The Dark Knight, 2008

…But what if we have *even stronger* prior information?

Priors for Modeling Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(4, 1)$$

Harvey Dent
aka “Two-Face”
The Dark Knight, 2008



Objective Priors

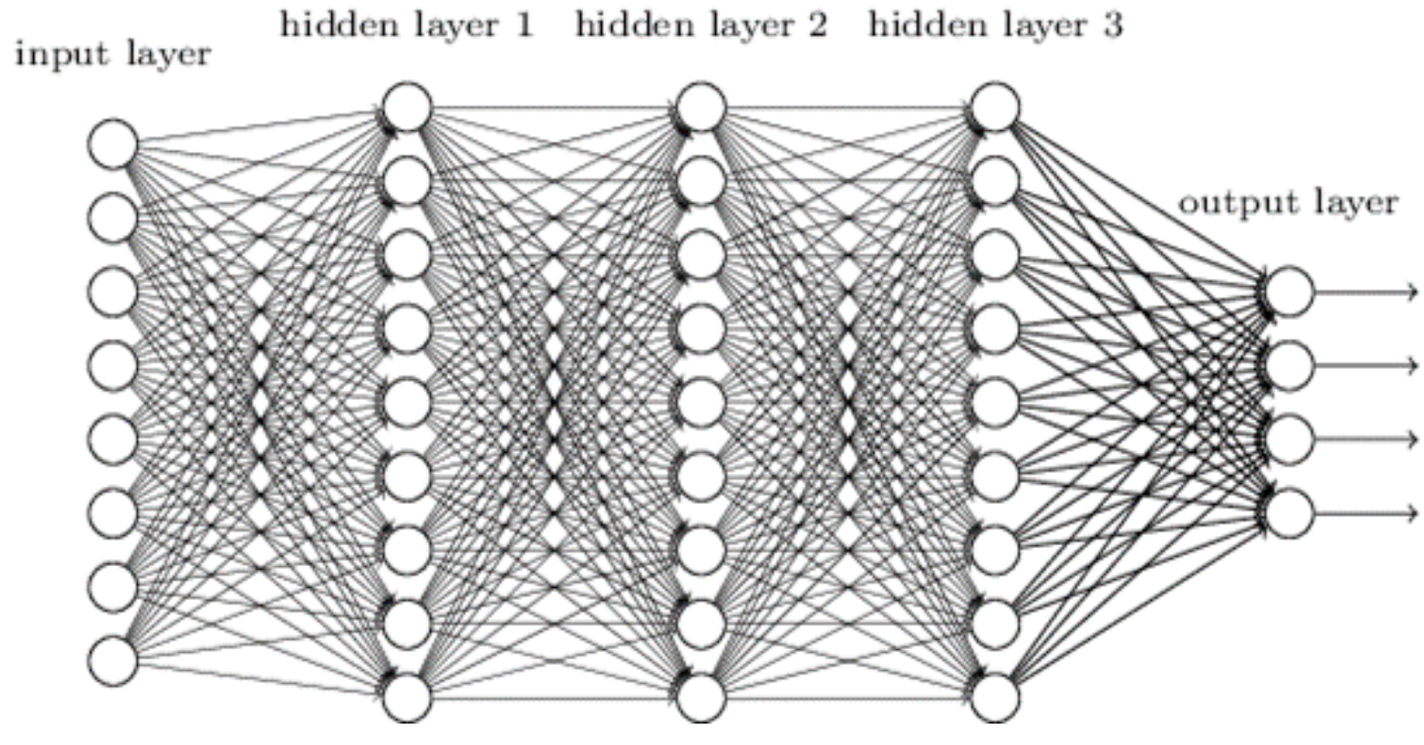
…But what if we have <u>no</u> information?

Say we are completely agnostic about Θ;
What prior represents a *state of ignorance*?

Objective Priors

…But what if we have <u>no</u> information?

Say we are completely agnostic about Θ;
What prior represents a *state of ignorance*?

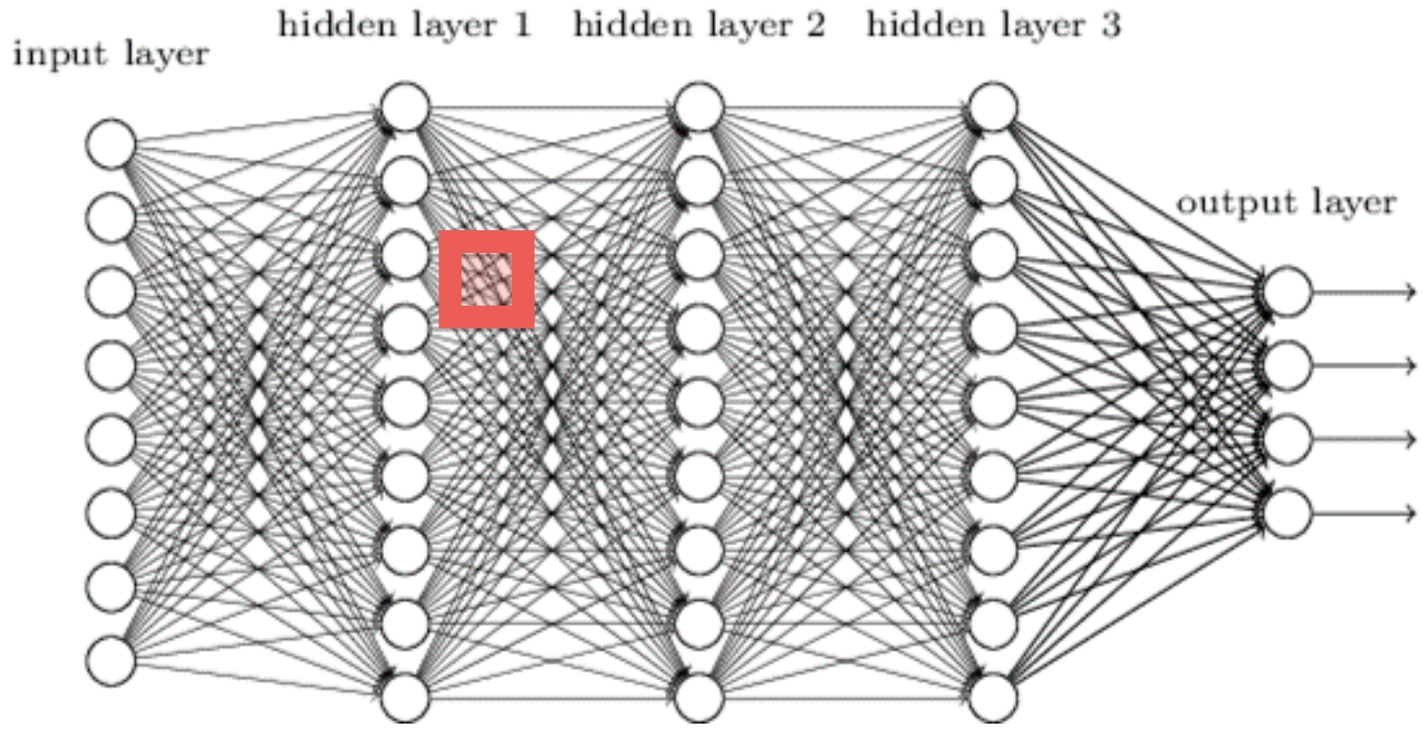


Objective Priors

…But what if we have <u>no</u> information?

Say we are completely agnostic about Θ;
What prior represents a *state of ignorance*?

What prior beliefs do I have about the 341st weight?

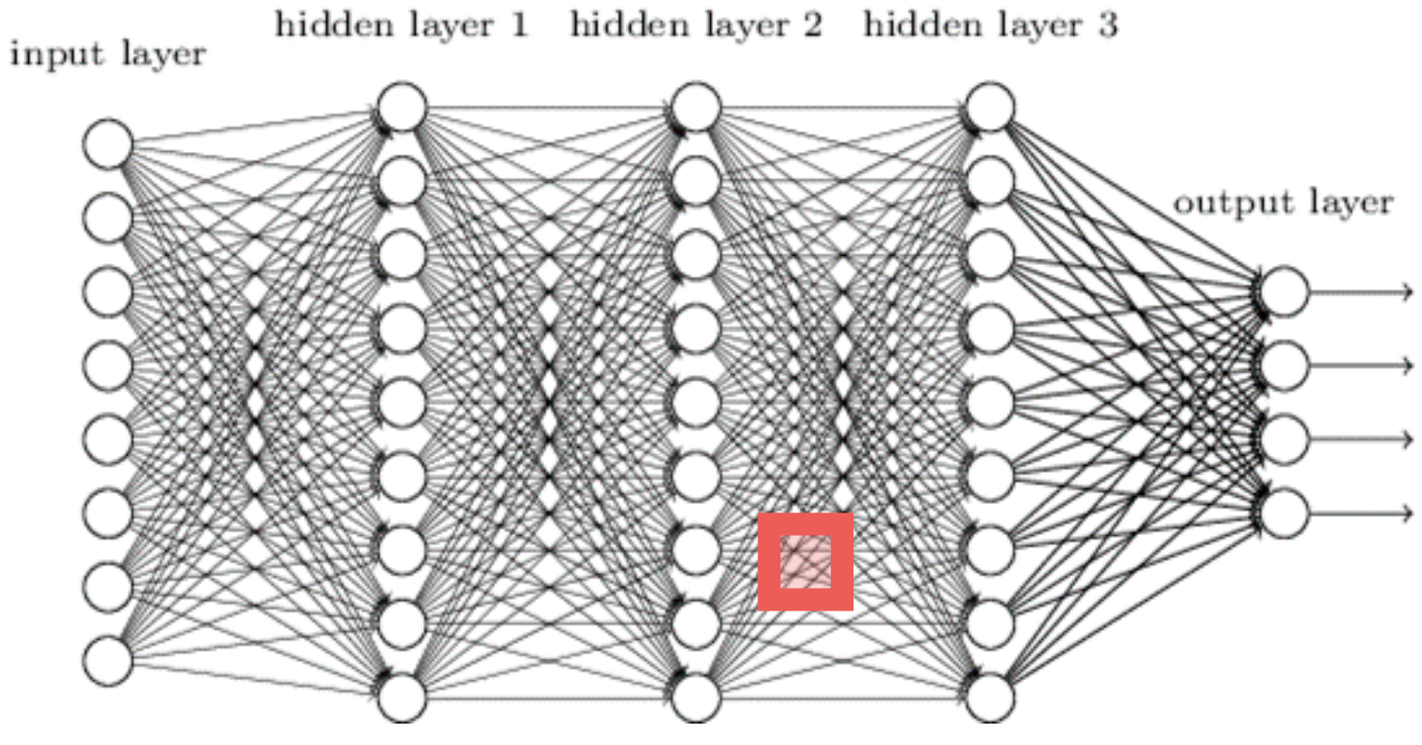


Objective Priors

…But what if we have <u>no</u> information?

Say we are completely agnostic about Θ;
What prior represents a *state of ignorance*?

What prior beliefs do I have about the 833rd weight? etc…

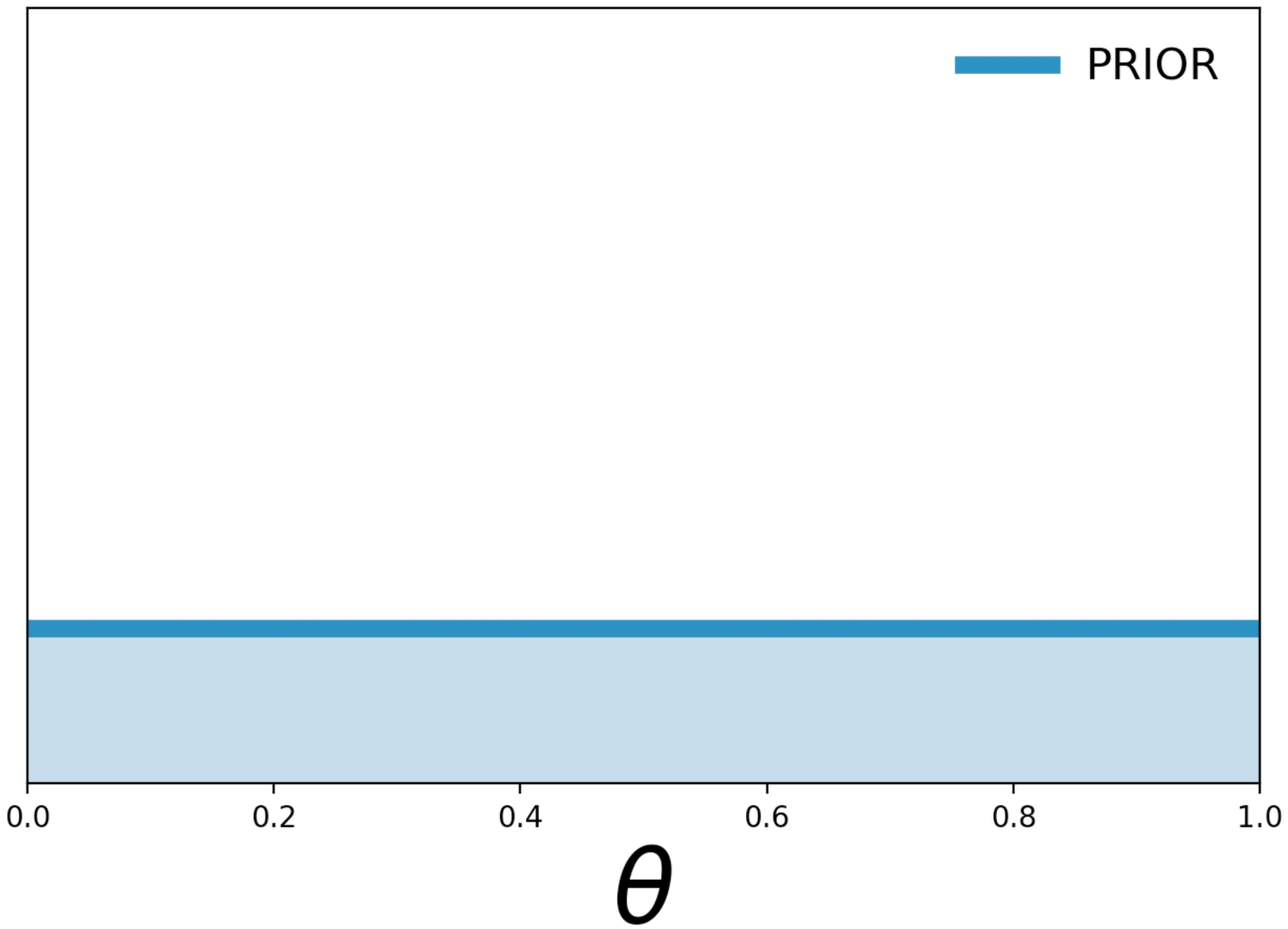


Objective Priors for Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(\alpha_0, \beta_0)$$

Objective Priors for Coin Tosses

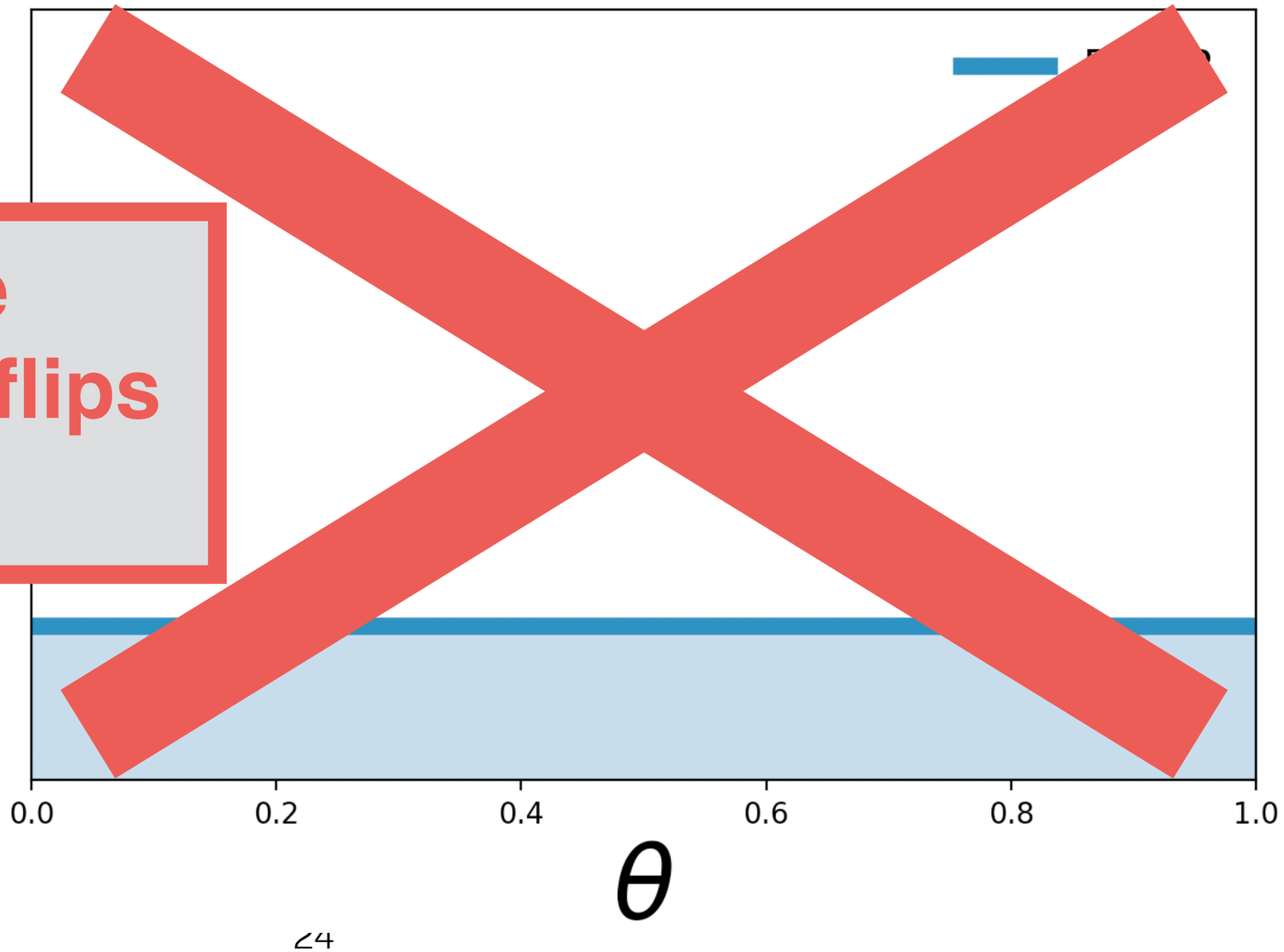
$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(1, 1)$$



Objective Priors for Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(1, 1)$$

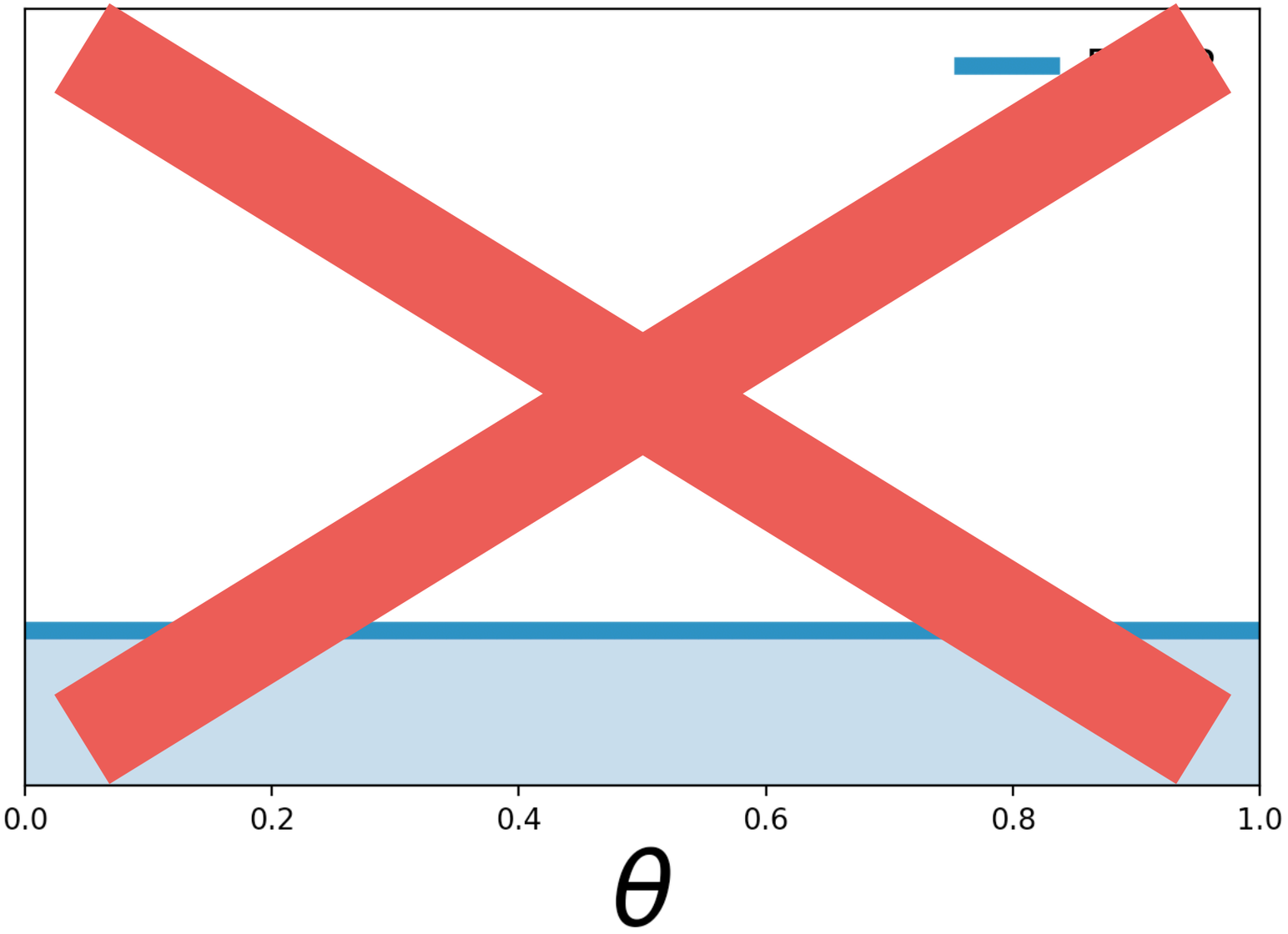
Implies that we've already seen two flips
one heads, one tails.



Objective Priors for Coin Tosses

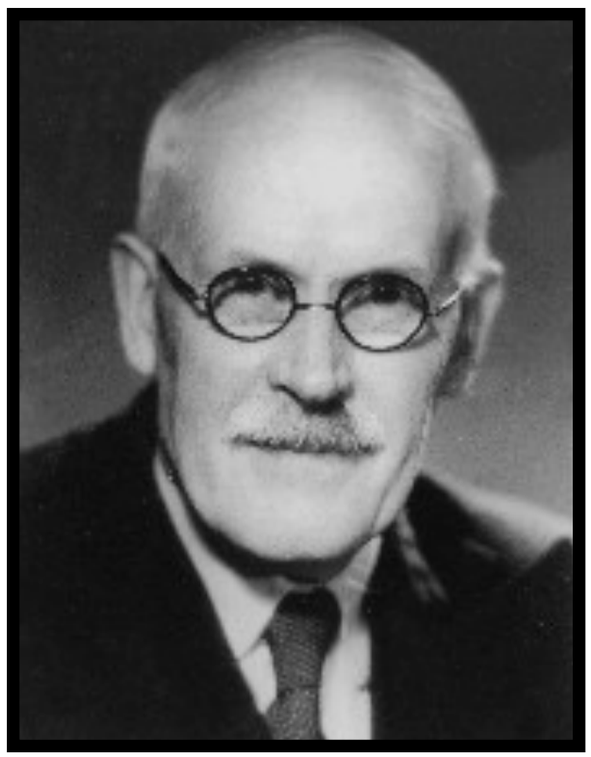
$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\ \texttt{Beta}(1, 1)$$

Pierre-Simon Laplace
(1814)

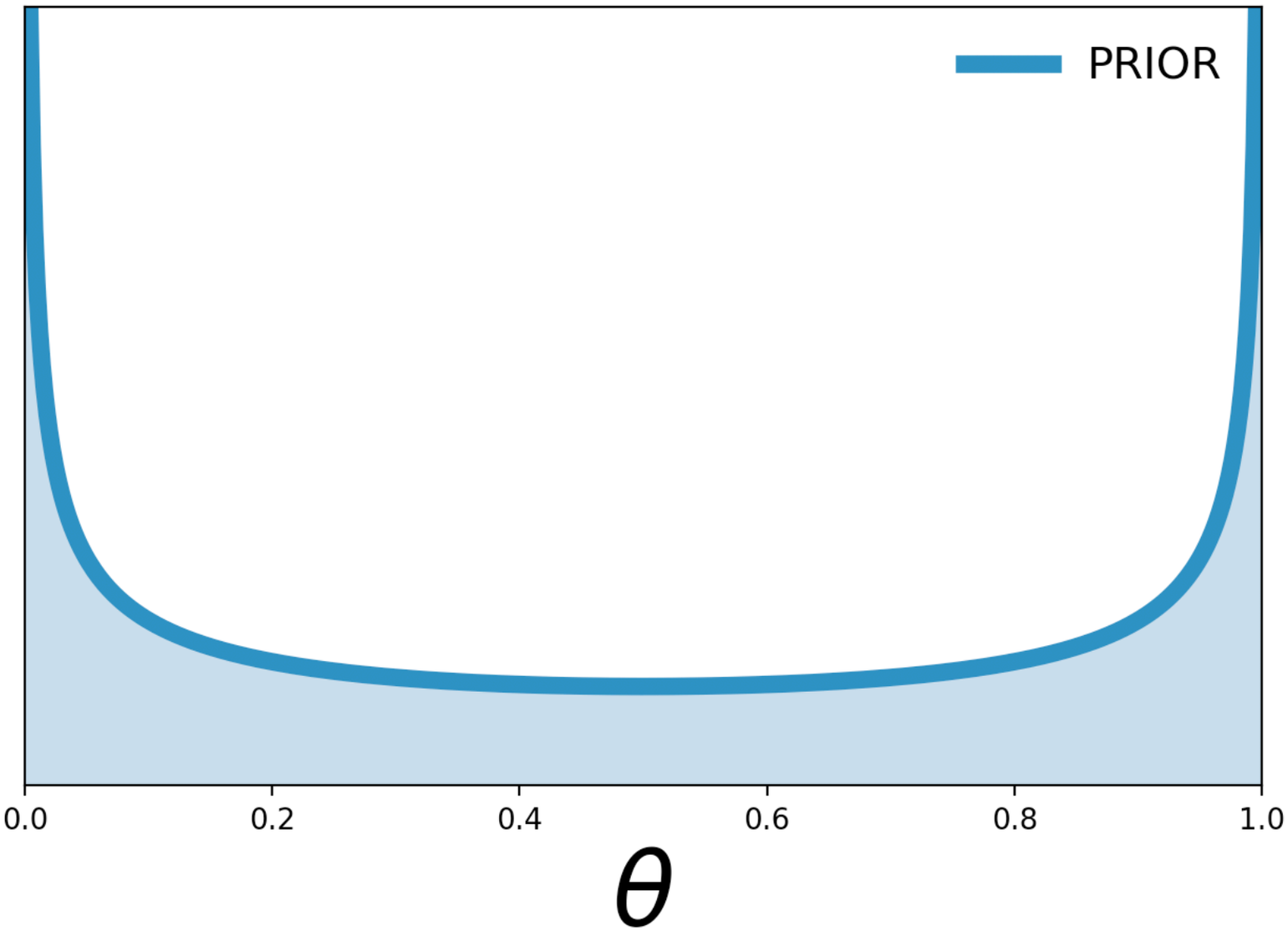


Objective Priors for Coin Tosses

$$p(\theta|\mathbf{X}) \propto \texttt{Binomial}(N, \theta)\, \texttt{Beta}(0.5, 0.5)$$



Sir Harold Jeffreys
(1946)



Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\ \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\ \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathbf{X}} p(\mathbf{X})\ \mathrm{KLD}[p(\boldsymbol{\theta}|\mathbf{X})||p(\boldsymbol{\theta})]\ d\mathbf{X}$$

Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\ \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathbf{X}} p(\mathbf{X})\ \mathrm{KLD}[p(\boldsymbol{\theta}|\mathbf{X})||p(\boldsymbol{\theta})]\ d\mathbf{X}$$

Maximize **divergence between posterior and prior**

Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\ \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathbf{X}} p(\mathbf{X})\ \mathrm{KLD}[p(\boldsymbol{\theta}|\mathbf{X})||p(\boldsymbol{\theta})]\ d\mathbf{X}$$

Averaged over marginal likelihood

Maximize **divergence between posterior and prior**

Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\ \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathbf{X}} p(\mathbf{X})\ \mathrm{KLD}[p(\boldsymbol{\theta}|\mathbf{X})||p(\boldsymbol{\theta})]\ d\mathbf{X}$$

More Examples

$$p(\theta) \propto 1$$

Gaussian Mean

$$p(\theta) \propto \frac{1}{\sigma}$$

Gaussian Scale

$$p(\theta) \propto \frac{1}{\sqrt{\lambda}}$$

Poisson Rate

Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\ \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathbf{X}} p(\mathbf{X})\ \mathrm{KLD}[p(\boldsymbol{\theta}|\mathbf{X})||p(\boldsymbol{\theta})]\ d\mathbf{X}$$

Properties

- ☐ **Invariant** to model reparametrization.
- ☐ **Credible intervals** nearly match **confidence intervals**.*

* Under certain conditions

Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\ \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathbf{X}} p(\mathbf{X})\ \mathrm{KLD}[p(\boldsymbol{\theta}|\mathbf{X})||p(\boldsymbol{\theta})]\ d\mathbf{X}$$

Problems / Obstacles

- [ ]
- [ ]

Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \, \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

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

Problems / Obstacles

- [ ] **Calculus of variations problem:** Solve for function
- [ ]

Objective Priors

[Bernardo, 1979]

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \, \mathbb{I}[\boldsymbol{\theta}, \mathbf{X}]$$

$$= \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathbf{X}} p(\mathbf{X}) \, \mathrm{KLD}\left[p(\boldsymbol{\theta}|\mathbf{X}) \,|p(\boldsymbol{\theta})\right] \, d\mathbf{X}$$

Problems / Obstacles

- [ ] **Calculus of variations problem:** Solve for function
- [ ] **Need posterior in closed-form**

CONTRIBUTION

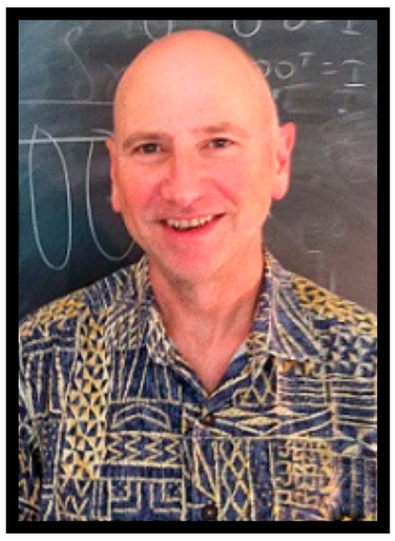
Approximating Objective Priors

Can we approximate objective priors so as to expand their use in machine learning?

Previous Work

Markov Chain Monte Carlo [Lafferty & Wasserman, 2001]**:**
Use MCMC to draw samples from the objective prior.

$$p^*(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} \delta[\hat{\theta}_s]$$



Iterative Markov Chain Monte Carlo Computation of Reference Priors and Minimax Risk

John Lafferty
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lafferty@cs.cmu.edu

Larry Wasserman
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
larry@stat.cmu.edu

John Lafferty Larry Wasserman

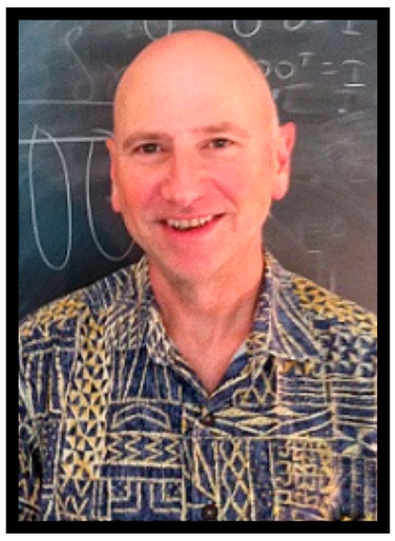
UAI 2001

Previous Work

Markov Chain Monte Carlo [Lafferty & Wasserman, 2001]**:**
Use MCMC to draw samples from the objective prior.

$$p^*(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} \delta[\hat{\theta}_s]$$

Drawback: Have to perform a second modeling step in order to evaluate the prior's density.



Iterative Markov Chain Monte Carlo Computation of Reference Priors and Minimax Risk

John Lafferty
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lafferty@cs.cmu.edu

Larry Wasserman
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
larry@stat.cmu.edu

John Lafferty Larry Wasserman

UAI 2001

A Parametric Approach

We propose optimizing a **parametric approximation**.

$$p^*(\theta) \approx p(\theta; \underline{\lambda}^*)$$

Optimize w.r.t. $\boldsymbol{\lambda}$ instead of function itself.

A Parametric Approach

We propose optimizing a **parametric approximation**.

$$p^*(\theta) \approx p(\theta; \underline{\lambda^*})$$

Optimize w.r.t. $\boldsymbol{\lambda}$ instead of function itself.

Objective is a **lower bound** on the mutual information.

$$\mathbb{I}[\boldsymbol{\theta}, \mathbf{X}] \geq \mathbb{E}_{\boldsymbol{\theta}_{\boldsymbol{\lambda}}, \mathbf{X}} \left[\log p(\mathbf{X}|\boldsymbol{\theta}) \; - \; \max_{\boldsymbol{\theta}_{\boldsymbol{\lambda}}} \; \log p(\mathbf{X}|\boldsymbol{\theta}) \right]$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} \mathrm{KLD} \left[p(\mathbf{X}|\hat{\boldsymbol{\theta}}_s) \; || \; p(\mathbf{X}|\hat{\boldsymbol{\theta}}_{\max}) \right]$$

Samples depend on $\boldsymbol{\lambda}$.

A Parametric Approach

BENEFITS

A Parametric Approach

BENEFITS

- **Comparison to Lafferty & Wasserman's [2001] MCMC**: No need to perform a second modeling step to evaluate the prior.

A Parametric Approach

BENEFITS

- **Comparison to Lafferty & Wasserman's [2001] MCMC**: No need to perform a second modeling step to evaluate the prior.
- **Conjugate Objective Priors**: Can pick a prior family that is conjugate and then find the member that is nearest to a objective prior.

A Parametric Approach

BENEFITS

- **Comparison to Lafferty & Wasserman's [2001] MCMC**: No need to perform a second modeling step to evaluate the prior.
- **Conjugate Objective Priors**: Can pick a prior family that is conjugate and then find the member that is nearest to a objective prior.
- **Proper Objective Priors**: If concerned b/c the true prior is improper, can use an approximation that is proper.

A Parametric Approach

BENEFITS

- **Comparison to Lafferty & Wasserman's [2001] MCMC**: No need to perform a second modeling step to evaluate the prior.
- **Conjugate Objective Priors**: Can pick a prior family that is conjugate and then find the member that is nearest to a objective prior.
- **Proper Objective Priors**: If concerned b/c the true prior is improper, can use an approximation that is proper.

Gaussian mean:

$$p^*(\boldsymbol{\mu}) \propto 1$$

A Parametric Approach

BENEFITS

- **Comparison to Lafferty & Wasserman's [2001] MCMC**: No need to perform a second modeling step to evaluate the prior.
- **Conjugate Objective Priors**: Can pick a prior family that is conjugate and then find the member that is nearest to a objective prior.
- **Proper Objective Priors**: If concerned b/c the true prior is improper, can use an approximation that is proper.

Gaussian mean:

$$p^*(\boldsymbol{\mu}) \propto 1$$

Parametric Approximation:

$$\mathrm{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{\sigma}_\lambda)$$

A Parametric Approach

BENEFITS

- **Comparison to Lafferty & Wasserman's [2001] MCMC**: No need to perform a second modeling step to evaluate the prior.
- **Conjugate Objective Priors**: Can pick a prior family that is conjugate and then find the member that is nearest to a objective prior.
- **Proper Objective Priors**: If concerned b/c the true prior is improper, can use an approximation that is proper.

Gaussian mean:	**Parametric Approximation**:	**Optimization Objective**:
$p^*(\boldsymbol{\mu}) \propto 1$	$\mathrm{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{\sigma}_\lambda)$	$\mathcal{J}(\boldsymbol{\lambda}) = \|\|\boldsymbol{\sigma}_\lambda\|\|_2^2$

A Parametric Approach

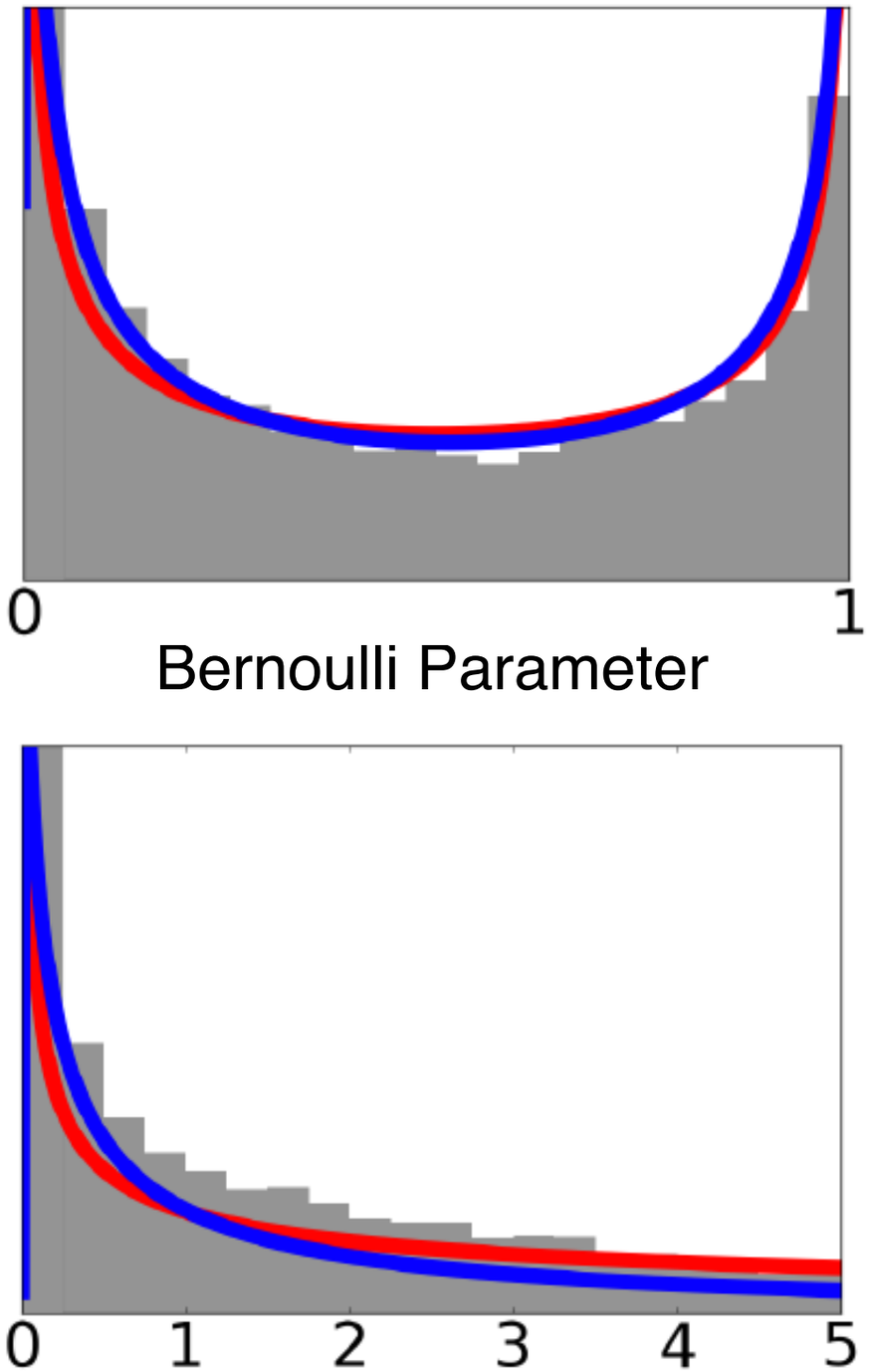
BENEFITS

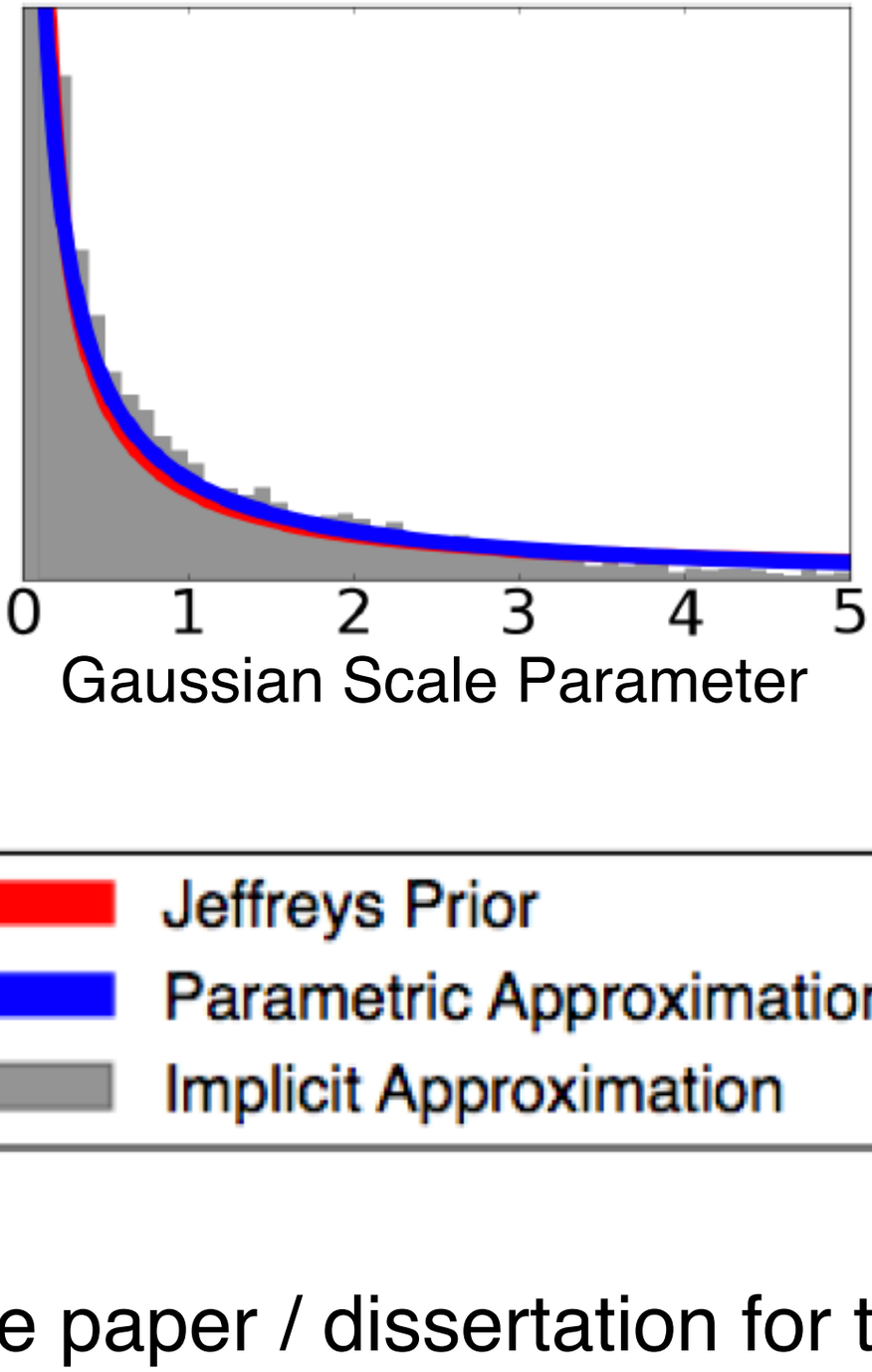
- ☐ **Comparison to Lafferty & Wasserman's [2001] MCMC**: No need to perform a second modeling step to evaluate the prior.
- ☐ **Conjugate Objective Priors**: Can pick a prior family that is conjugate and then find the member that is nearest to a objective prior.
- ☐ **Proper Objective Priors**: If concerned b/c the true prior is improper, can use an approximation that is proper.

Gaussian mean:	**Parametric Approximation**:	**Optimization Objective**:
$p^*(\boldsymbol{\mu}) \propto 1$	$\mathrm{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{\sigma}_\lambda)$	$\mathcal{J}(\boldsymbol{\lambda}) = \|\|\boldsymbol{\sigma}_\lambda\|\|_2^2$

RESULT: $p^* \propto 1 \approx \mathrm{N}(\cdot, \infty)$

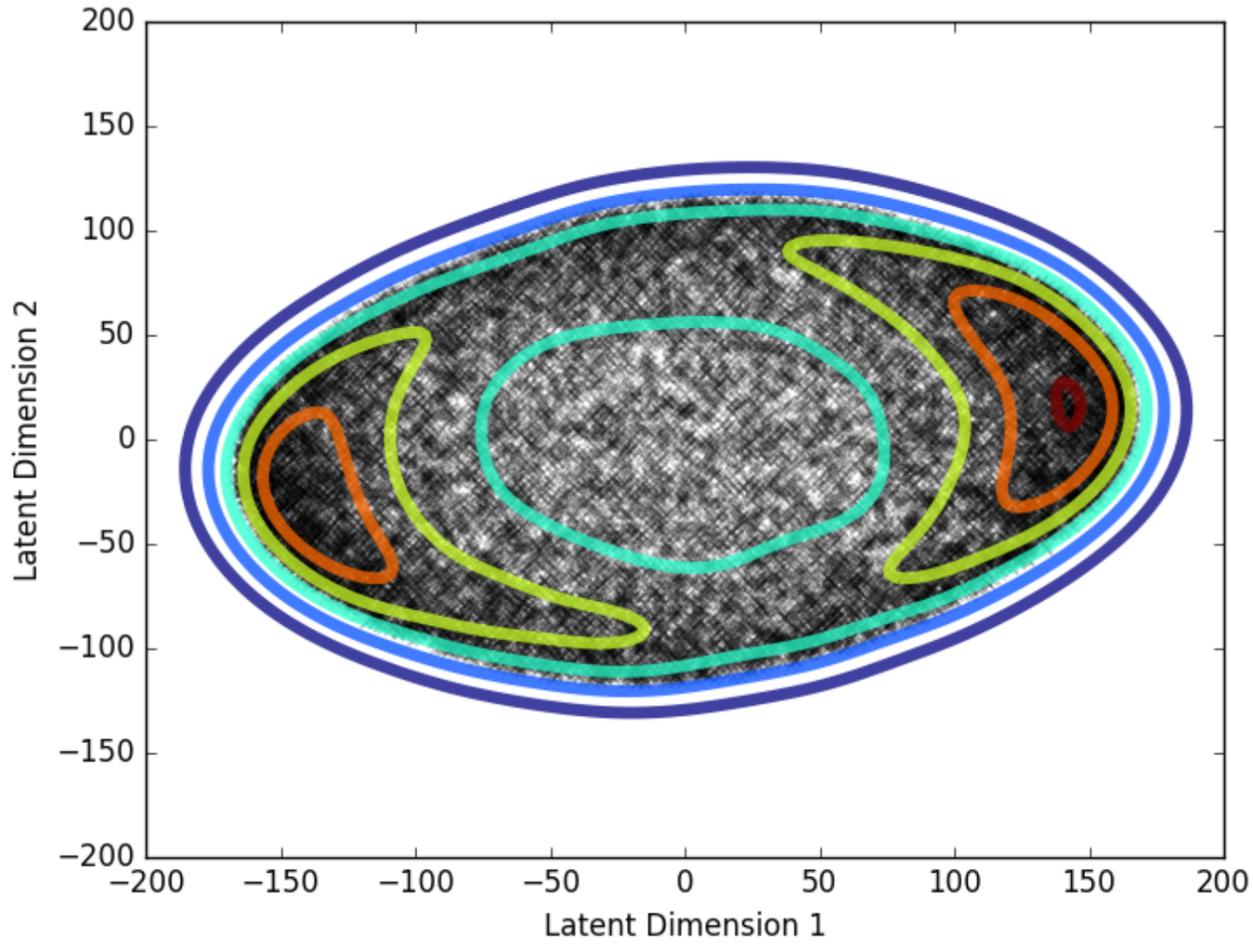
Recovering Jeffreys Priors





See paper / dissertation for two-sample tests comparing approximation methods.

Finding the Variational Autoencoder's Objective Prior



p(**z**)

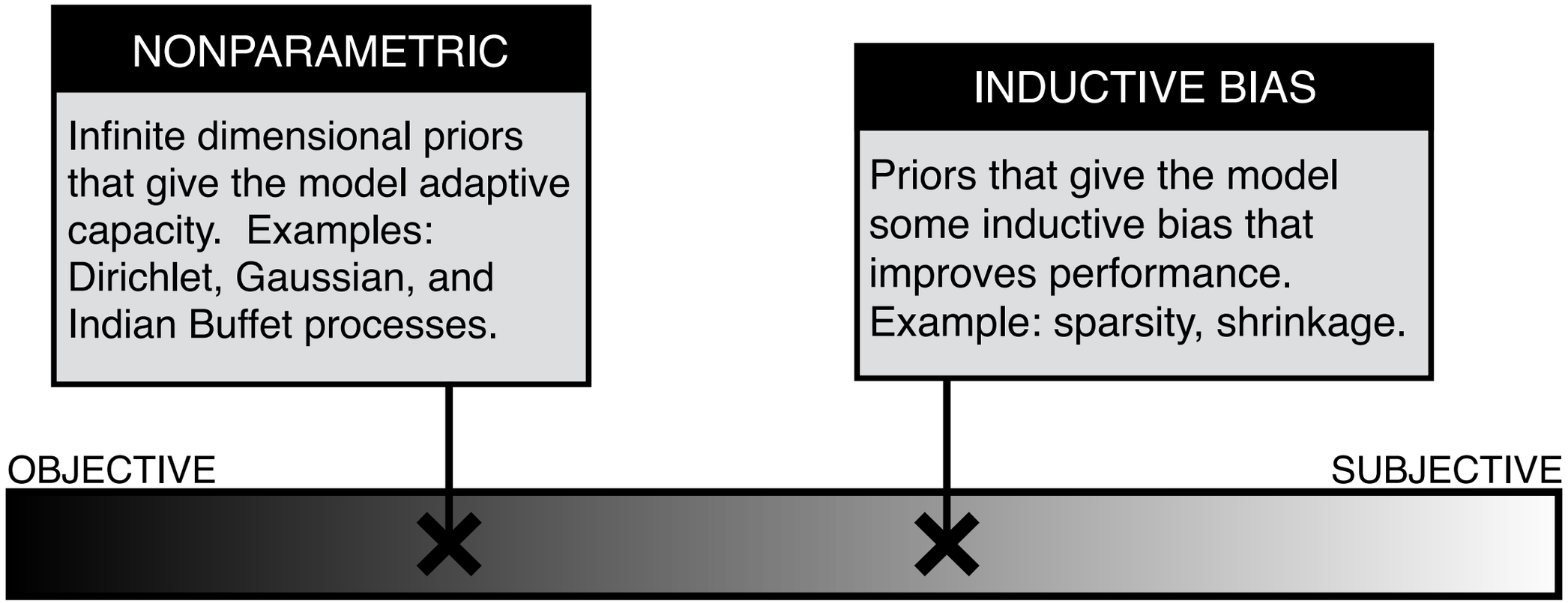
Conclusions

Priors for Machine Learning

A SPECTRUM OF PRIORS

Priors for Machine Learning

A SPECTRUM OF PRIORS



Priors for Machine Learning

A SPECTRUM OF PRIORS

NONPARAMETRIC

Infinite dimensional priors that give the model adaptive capacity. Examples: Dirichlet, Gaussian, and Indian Buffet processes.

INDUCTIVE BIAS

Priors that give the model some inductive bias that improves performance. Example: sparsity, shrinkage.

OBJECTIVE

SUBJECTIVE

SUBJECTIVE

Priors determined by intuition, previous experiments, or expert knowledge.

Priors for Machine Learning

A SPECTRUM OF PRIORS

NONPARAMETRIC

Infinite dimensional priors that give the model adaptive capacity. Examples: Dirichlet, Gaussian, and Indian Buffet processes.

INDUCTIVE BIAS

Priors that give the model some inductive bias that improves performance. Example: sparsity, shrinkage.

OBJECTIVE ← → SUBJECTIVE

Published as a conference paper at ICLR 2019

FUNCTIONAL VARIATIONAL BAYESIAN NEURAL NETWORKS

Shengyang Sun*†, **Guodong Zhang***†, **Jiaxin Shi***‡, **Roger Grosse**†
†University of Toronto, †Vector Institute, ‡Tsinghua University
{ssy, gdzhang, rgrosse}@cs.toronto.edu, shijx15@mails.tsinghua.edu.cn

SUBJECTIVE

Priors determined by intuition, previous experiments, or expert knowledge.

Priors for Machine Learning

A SPECTRUM OF PRIORS

NONPARAMETRIC

Infinite dimensional priors that give the model adaptive capacity. Examples: Dirichlet, Gaussian, and Indian Buffet processes.

INDUCTIVE BIAS

Priors that give the model some inductive bias that improves performance. Example: sparsity, shrinkage.

OBJECTIVE

SUBJECTIVE

THIS TALK

OBJECTIVE

Priors that 'follow the data,' perhaps having frequentist properties. Examples: diffuse, Jeffreys, reference, empirical Bayes.

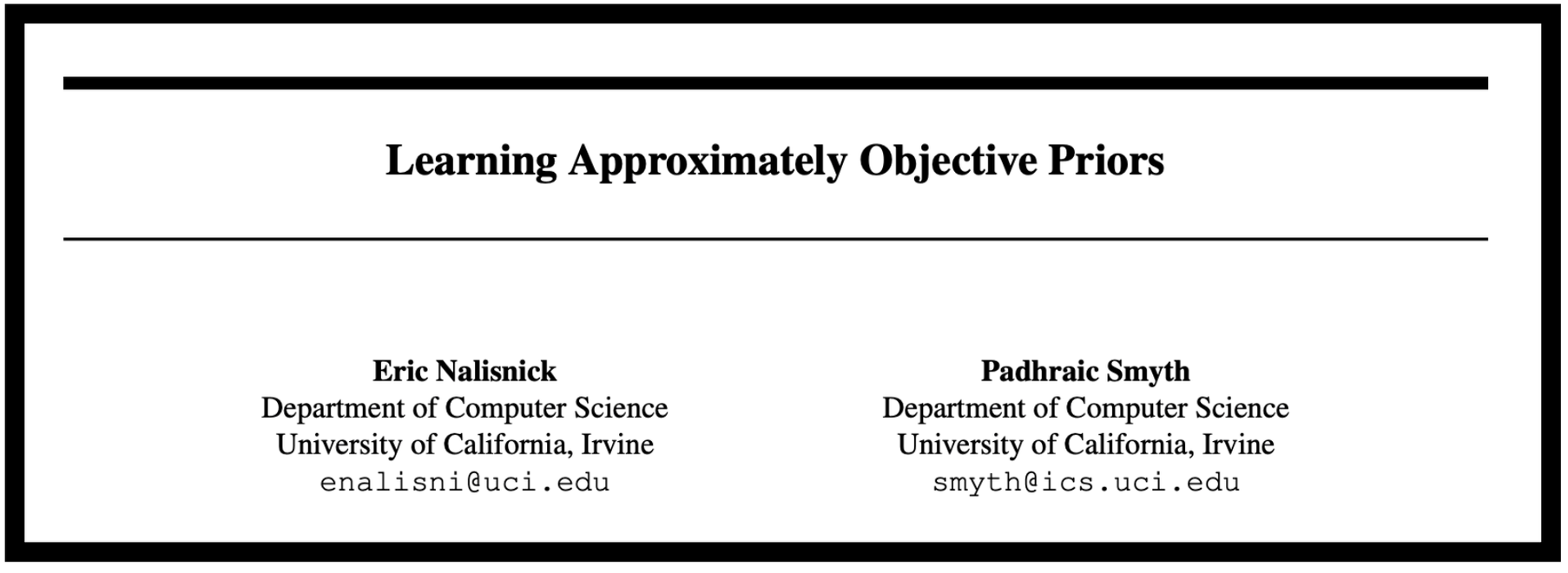
SUBJECTIVE

Priors determined by intuition, previous experiments, or expert knowledge.

Thank you. Questions?

In collaboration with…

Padhraic Smyth



Learning Approximately Objective Priors

Eric Nalisnick
Department of Computer Science
University of California, Irvine
enalisni@uci.edu

Padhraic Smyth
Department of Computer Science
University of California, Irvine
smyth@ics.uci.edu

UAI 2017