Learning Approximately Objective Priors

Galina Boeva

MIPT, 2023

October 3, 2023

Motivation

Main idea

It's difficult to have informative priors, so they turn to uninformative priors, but for many models it's impossible to deduce them. Therefore, the authors propose to approximate these prior.

Definition

Likelihood

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x_i|\theta), \tag{1}$$

where θ are the model parameters and $\mathcal{D}$ is the dataset, which is comporised of N i.i.d. observations $x_i \in \mathcal{X}$. $p(\theta)$ denotes the prior, $p(\theta|\mathcal{D})$ the posterior, and $p(\mathcal{D}) = \int\limits_{\theta} p(\mathcal{D}|\theta)p(\theta)d\theta$ the marginal likelihood.

Reference prior

Reference prior is the distribution that maximizes the mutual information between the parameters θ and the data $\mathcal{D}$:

$$p^*(\theta) = \arg\max_{p(\theta)} \mathbb{I}(\theta, \mathcal{D}) = \arg\max_{p(\theta)} \mathbb{H}[\theta] - \mathbb{H}[\theta|\mathcal{D}] \tag{2}$$

Definition

Mutual information

Mutual information in terms of a Kullback-Leibler divergence:

$$I(\theta, \mathcal{D}) = \int_{\mathcal{D}} p(\mathcal{D}) KLD[p(\theta|\mathcal{D})||p(\theta)] d\mathcal{D}. \tag{3}$$

$$I(\theta, \mathcal{D}) = -\int_{\theta} p(\theta) \log \frac{p(\theta)}{f(\theta)} d\theta, f(\theta) = \exp\Big\{ \int_{\mathcal{D}} p(\mathcal{D}|\theta) \log p(\theta|\mathcal{D}) d\mathcal{D} \Big\}. \tag{4}$$

Jeffreys Prior

$$\pi(\theta) \propto \sqrt{det \mathcal{F}[\theta]}. \tag{5}$$

Bernstein Von Mises theorem

$$p(\theta|\mathcal{D}) \approx N(\theta_{MLE}, \mathcal{F}^{-1}(\theta)) \tag{6}$$

Information lower bound

Argmax λ

$$\boldsymbol{\lambda}_* = \arg\max_{\boldsymbol{\lambda}} \mathbb{I}(\boldsymbol{\theta}, \mathcal{D}) = \arg\max_{\boldsymbol{\lambda}} \int_{\boldsymbol{\theta}} p_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \int_{\mathcal{D}} p(\mathcal{D}|\boldsymbol{\theta}) \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{p_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) p(\mathcal{D})} d\mathcal{D} d\boldsymbol{\theta}$$

$$= \arg\max_{\boldsymbol{\lambda}} \int_{\boldsymbol{\theta}} p_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \int_{\mathcal{D}} p(\mathcal{D}|\boldsymbol{\theta}) \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{p(\mathcal{D})} d\mathcal{D} d\boldsymbol{\theta} =$$

$$= \arg\max_{\boldsymbol{\lambda}} \mathbb{E}_{\boldsymbol{\theta}_{\boldsymbol{\lambda}}} [-\mathbb{H}_{\mathcal{D}|\boldsymbol{\theta}}[\mathcal{D}] - \mathbb{E}_{\mathcal{D}|\boldsymbol{\theta}}[\log p(\mathcal{D})]] \tag{7}$$

Bounding $\log p(\mathcal{D})$

Using variational Renyi bound:

$$\log p(\mathcal{D}) \leq \frac{1}{1-\alpha} \log \mathbb{E}_{\boldsymbol{\theta}}(p(\mathcal{D}|\boldsymbol{\theta})^{(1-\alpha)}), \alpha \leq 0. \tag{8}$$

Information lower bound

Lower bound for λ

$$\mathbb{I}(\theta, \mathcal{D}) \geq \mathbb{E}_{\boldsymbol{\theta}_{\boldsymbol{\lambda}}} \Big[- \mathbb{H}_{\mathcal{D}|\theta}[\mathcal{D}] - \frac{1}{1-\alpha} \log \mathbb{E}_{\boldsymbol{\theta}}[p(\mathcal{D}|\theta)^{1-\alpha}] \Big]. \tag{9}$$

$$\mathbb{I}(\theta, \mathcal{D}) \geq \mathbb{J}_{\mathcal{RP}}(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\theta}_{\boldsymbol{\lambda}}} \Big[- \mathbb{H}_{\mathcal{D}|\theta}[\mathcal{D}] - \mathbb{E}_{\mathcal{D}|\theta}[\max_s \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)] \Big]. \tag{10}$$

$$\mathbb{J}_{\mathcal{RP}}(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\theta}_{\boldsymbol{\lambda}}} \mathbb{E}_{\mathcal{D}|\boldsymbol{\theta}} \Big[\log p(\mathcal{D}|\boldsymbol{\theta}) - \max_s \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s) \Big]. \tag{11}$$

$$\hat{\mathbb{J}}_{\mathcal{RP}}(\boldsymbol{\lambda}) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{H}[p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)||p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)] - \mathbb{H}_{\mathcal{D}|\hat{\boldsymbol{\theta}}_s}[\mathcal{D}] =$$

$$= \frac{1}{S} \sum_{s=1}^{S} KLD[p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)||p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)] \tag{12}$$

Particle descent

Stein Variational Gradient Descent

$$\bar{\theta}_j^{t+1} = \bar{\theta}_j^t + \eta\phi[\theta]$$

$$\phi[\theta] = \frac{1}{K}\sum_{k=1}^{K} \kappa(\bar{\theta}_k^t, \bar{\theta}_j^t)\nabla_{\bar{\theta}_k} \log p(\bar{\theta}_k^t) + \nabla_{\bar{\theta}_k}\kappa(\bar{\theta}_k^t, \bar{\theta}_j^t).$$

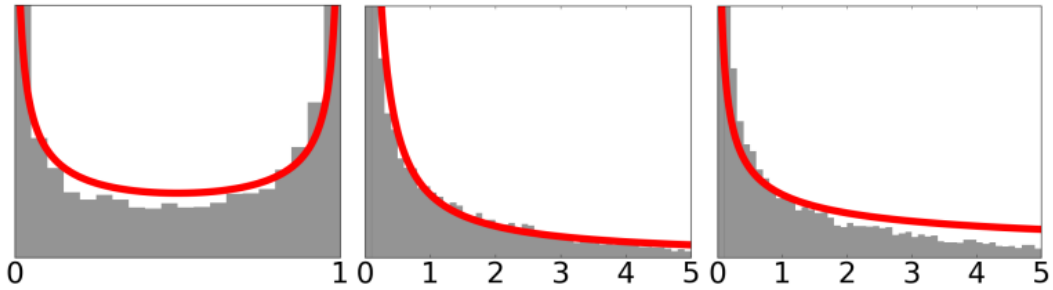
A-SVGD for RP Approximations

$$\nabla\bar{\theta} \log f(\bar{\theta}) = \nabla_{\bar{\theta}} \int_{\mathcal{D}} p(\mathcal{D}|\bar{\theta}) \log p(\bar{\theta}|\mathcal{D}) d\mathcal{D} =$$

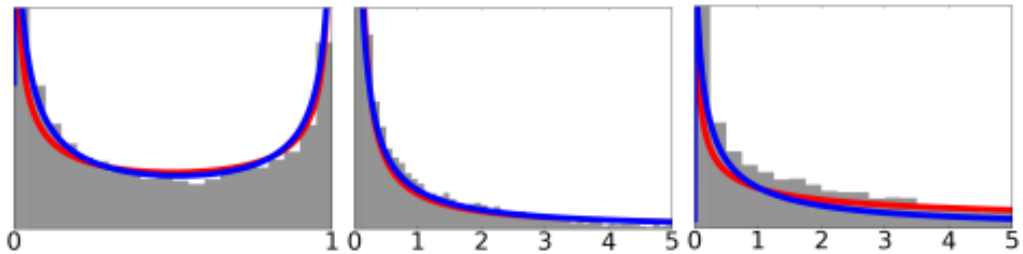
$$= \int_{\mathcal{D}} p(\mathcal{D}|\bar{\theta}) \log \frac{p(\mathcal{D}|\bar{\theta})1_\Theta}{p(\mathcal{D})} d\mathcal{D} = -\nabla_{\bar{\theta}}\mathbb{H}_{\mathcal{D}|\bar{\theta}}[\mathcal{D}] - \nabla_{\bar{\theta}}\mathbb{E}_{\mathcal{D}|\bar{\theta}}[\log p(\mathcal{D})]$$

$$= \nabla_{\bar{\theta}} \frac{1}{S} \sum_s KLD[p(\mathcal{D}|\bar{\theta})||p(\mathcal{D}|\hat{\theta}_s)].$$

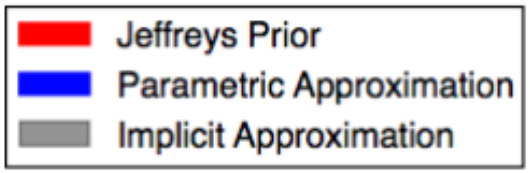
Empirical results: Approximation



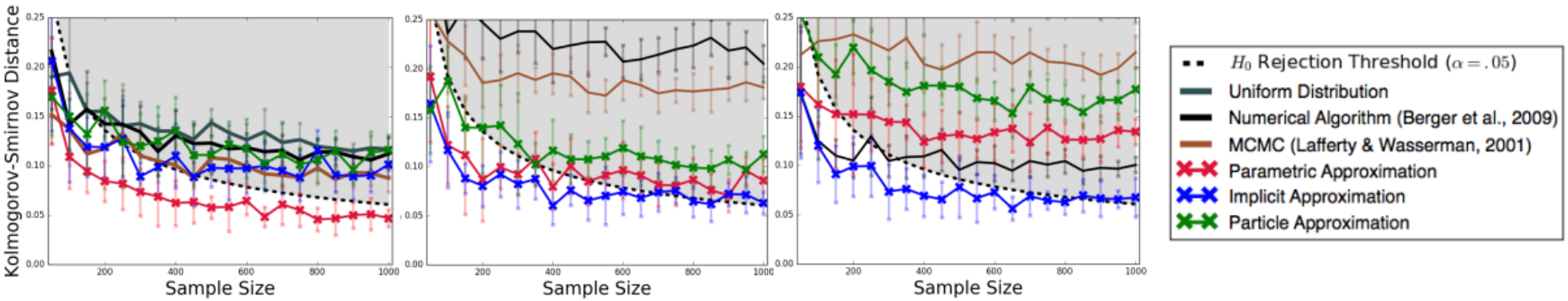
(a) Bernoulli (b) Gaussian Scale (c) Poisson



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



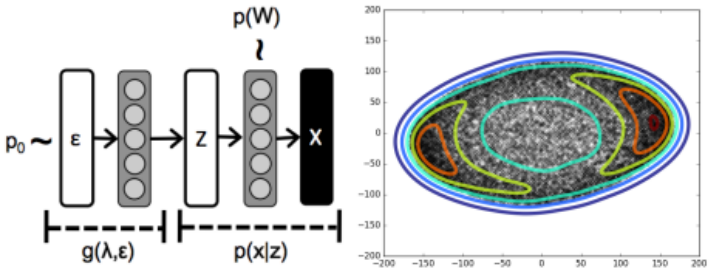
Empirical results: Approximation Quality



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

Figure: Quantifying the Approximation Quality.

Empirical results: VAE



(a) Training Configuration (b) Approximation

Figure: Learning the Variational Autoencoder's Reference Prior. (a) computational pipeline from the implicit prior through the VAE decoder; (b) RP approximation (contours are generated via kernel density estimation on 10, 000 samples).

Literature

1. **Main article** Learning Approximately Objective Priors.