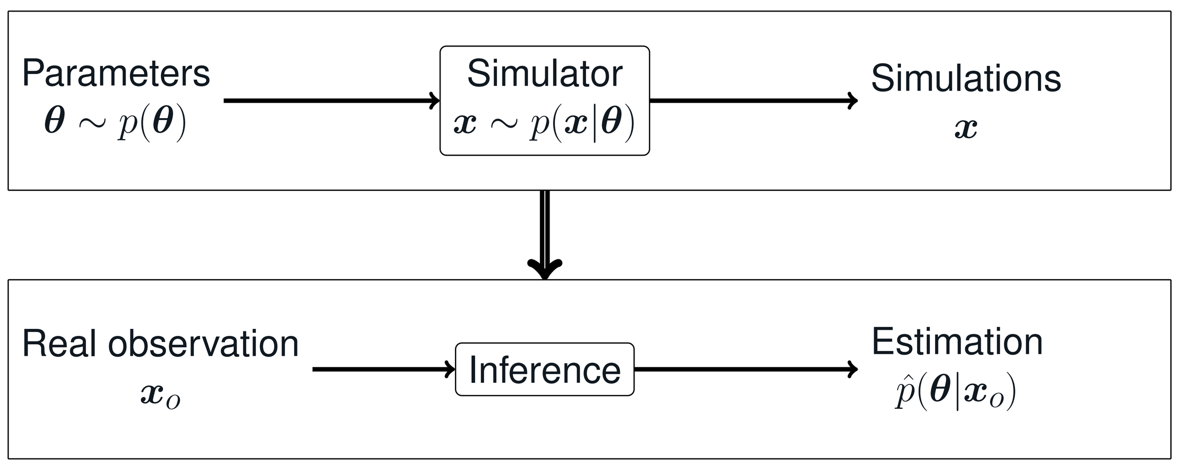
Balancing Simulation-based Inference for Conservative Posteriors

Arnaud Delaunoy*[1], Benjamin Kurt Miller*[2], Patrick Forré[2], Christoph Weniger[2], Gilles Louppe[1]

[1]University of Liège, [2]University of Amsterdam

Simulation-based inference

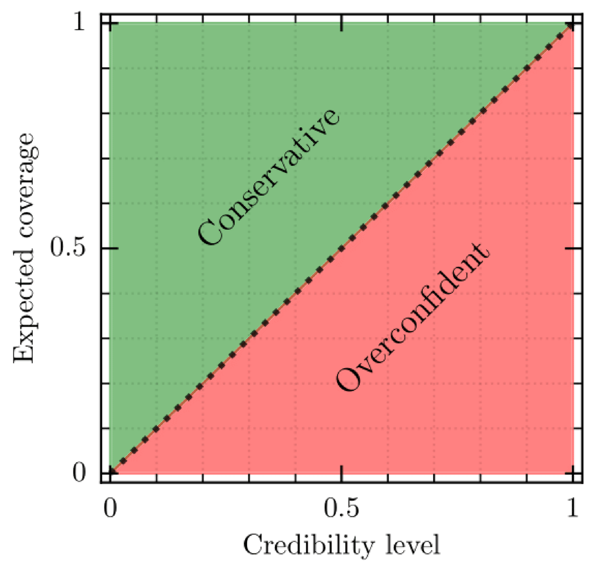


Conservativeness in SBI

Expected coverage probability of the posterior surrogate $\hat{p}(\boldsymbol{\theta} \,|\, \boldsymbol{x})$:

$$1 - \hat{\alpha}[\hat{p}; \alpha] \coloneqq \mathbb{E}_{p(\boldsymbol{\theta}, \boldsymbol{x})}\left[\mathbb{1}(\boldsymbol{\theta} \in \Theta_{\hat{p}(\boldsymbol{\theta} \,|\, \boldsymbol{x})}(1-\alpha))\right].$$

- When $\exists \alpha' : 1 - \hat{\alpha}[\hat{p}; \alpha'] < 1 - \alpha'$, we say that $\hat{p}(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ is *overconfident*.
- Overconfidence is problematic because the surrogate tends to exclude parameter values that are actually plausible at the considered credibility level. On the other hand, extremely *underconfident* surrogates are not informative. Although there is a tradeoff, scientific applications take a cautious approach by favoring underconfidence.
- **We encourage *conservative surrogates at credibility level* α', which have** $1 - \hat{\alpha}[\hat{p}; \alpha'] \geq 1 - \alpha'$.



Balanced Neural Ratio Estimation

Neural ratio estimation (NRE) trains a classifier $\varpi(y = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{x})$ to discriminate between jointly drawn samples, $p(\boldsymbol{\theta}, \boldsymbol{x})$, and marginal samples, $p(\boldsymbol{x})p(\boldsymbol{\theta})$, i.e. sampling:

$$\pi(\boldsymbol{\theta}, \boldsymbol{x} \,|\, y) \coloneqq \begin{cases} p(\boldsymbol{\theta})p(\boldsymbol{x}) & y = 0 \\ p(\boldsymbol{\theta}, \boldsymbol{x}) & y = 1 \end{cases}$$

with marginals $\pi(y = 0) \coloneqq \pi(y = 1) \coloneqq \frac{1}{2}$.

Data generation for training:

$$y \sim \pi(y) \coloneqq \mathrm{Ber}(y; \tfrac{1}{2}) \begin{cases} y = 0 \rightarrow (\boldsymbol{\theta}, \boldsymbol{x}) \sim p(\boldsymbol{\theta})p(\boldsymbol{x}) \\ y = 1 \rightarrow (\boldsymbol{\theta}, \boldsymbol{x}) \sim p(\boldsymbol{\theta}, \boldsymbol{x}) \end{cases}$$

Inference on $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ and $\boldsymbol{x}_o$:

$$(\boldsymbol{\theta}, \boldsymbol{x}_o) \rightarrow \boxed{\text{Classifier } \varpi(y = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{x}_o)} \rightarrow \hat{p}(\boldsymbol{\theta} \,|\, \boldsymbol{x}_o) = \frac{\varpi(y=1 \,|\, \boldsymbol{\theta}, \boldsymbol{x}_o)}{1 - \varpi(y=1 \,|\, \boldsymbol{\theta}, \boldsymbol{x}_o)} p(\boldsymbol{\theta})$$

Balanced Neural Ratio Estimation (BNRE) regularizes the classifier to be more conservative by minimizing the balancing criterion (using a Lagrange multiplier) which is expressed as

$$B[\varpi] \coloneqq B(\boldsymbol{w}) \coloneqq \left(\mathbb{E}_{p(\boldsymbol{\theta})p(\boldsymbol{x})}\left[\varpi(y = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{x})\right] + \mathbb{E}_{p(\boldsymbol{\theta}, \boldsymbol{x})}\left[\varpi(y = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{x})\right] - 1\right)^2,$$

where $\boldsymbol{w}$ are the classifier weights. This is added to *the main NRE objective, the binary cross entropy*.

Contribution 1: a new view on the balancing criterion

The χ^2 divergence is defined as

$$\chi^2(\pi(y) \,\|\, \varpi(y)) \coloneqq \int \left(\frac{\varpi(y)}{\pi(y)} - 1\right)^2 \pi(y)\, dy.$$

We identify that $B[\varpi] = \chi^2(\pi(y) \,\|\, \varpi(y))$.

- **Enforcing the balancing criterion regularizes the marginal classifier towards the target distribution for y:** $\pi(y)$**.** This new objective aims to be a building block to construct a better understanding of the balancing criterion. It reveals a principle for balancing multi-class classifiers.
- **Why the** χ^2 **divergence?** The Kullback-Leibler divergence $\mathrm{KL}(\pi(y) \,\|\, \varpi(y))$ would be information-theoretically motivated, but it is challenging to optimize due to the $\log$ in the integrand.

Contribution 2: extending balancing beyond NRE

Define a classifier in terms of the variational (unnormalized) posterior approximant $\hat{q}_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})$. We approximate $r(\boldsymbol{\theta}, \boldsymbol{x}) \coloneqq \frac{p(\boldsymbol{\theta}, \boldsymbol{x})}{p(\boldsymbol{\theta})p(\boldsymbol{x})} = \frac{p(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{p(\boldsymbol{\theta})}$ with $\frac{\hat{q}_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{p(\boldsymbol{\theta})}$ which yields the classifier

$$\varpi(y = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{x}; \hat{q}_{\boldsymbol{w}}) \coloneqq \frac{\hat{q}_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})/p(\boldsymbol{\theta})}{1 + \hat{q}_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})/p(\boldsymbol{\theta})}.$$

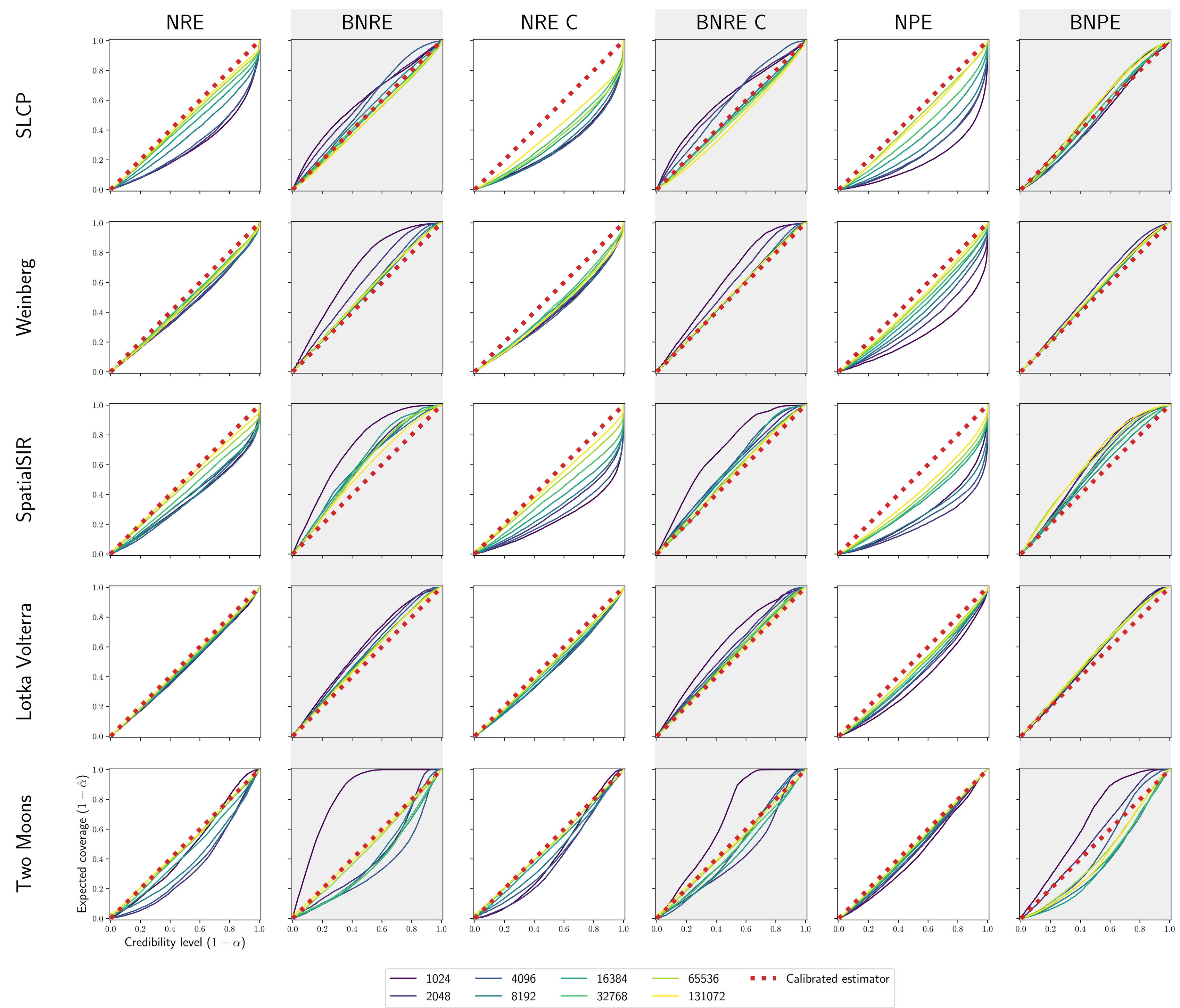
The balancing criterion can be expressed

$$B(\boldsymbol{w}) \coloneqq \left(\int (\pi(\boldsymbol{\theta}, \boldsymbol{x} \,|\, y = 0) + \pi(\boldsymbol{\theta}, \boldsymbol{x} \,|\, y = 1))\varpi(y = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{x}; \hat{q}_{\boldsymbol{w}}) d\boldsymbol{\theta}\, d\boldsymbol{x} - 1\right)^2$$

$$= \left(\int (p(\boldsymbol{\theta})p(\boldsymbol{x}) + p(\boldsymbol{\theta}, \boldsymbol{x}))\frac{\hat{q}_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})/p(\boldsymbol{\theta})}{1 + \hat{q}_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})/p(\boldsymbol{\theta})} d\boldsymbol{\theta}\, d\boldsymbol{x} - 1\right)^2$$

- **We propose BNPE** which regularizes NPE's maximum likelihood-based objective with the balance criterion to train a normalized density estimator $q_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})$. We have $\hat{q}_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \coloneqq q_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})$.
- **We propose BNRE-C** which regularizes NRE-C's multi-class, classifier-based objective with the (binary) balance criterion to train a ratio estimator. How do we define the binary classifier since NRE-C normally only defines a multi-class classifier? We do it in terms of the (unnormalized) density estimator $\hat{q}_{\boldsymbol{w}}(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \coloneqq \exp \circ h_{\boldsymbol{w}}(\boldsymbol{\theta}, \boldsymbol{x})p(\boldsymbol{\theta})$ where $h_{\boldsymbol{w}}$ is a neural network. Using the above definition, the corresponding (binary) classifier is $\varpi(y = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{x}; \hat{q}_{\boldsymbol{w}}) \coloneqq \frac{\exp \circ h_{\boldsymbol{w}}(\boldsymbol{\theta}, \boldsymbol{x})p(\boldsymbol{\theta})/p(\boldsymbol{\theta})}{1 + \exp \circ h_{\boldsymbol{w}}(\boldsymbol{\theta}, \boldsymbol{x})p(\boldsymbol{\theta})/p(\boldsymbol{\theta})} = \sigma \circ h_{\boldsymbol{w}}(\boldsymbol{\theta}, \boldsymbol{x})$. This regularizing classifier is the same binary classifier as in BNRE.

Results



BNPE and BNRE-C produce more conservative posteriors than NPE and NRE-C, respectively.

What if the posterior surrogate is imbalanced?

We observed that BNPE can be harder to balance than classifier-based algorithms. We show that this can be mitigated with a proper initialization scheme.

- **Hypothesis**: In order to be balanced, BNPE needs to learn the prior as the equivalent classifier is a function of both the approximate posterior and the prior.
- **Solution**: Initialize the normalizing flow in a balanced state (close to the prior).

How?

- Use the prior as base distribution or add a transformation that maps the base distribution to the prior at the end of the normalizing flow.
- Initialize all the transformations to an identity function.

Solving NPE "leakage" Several NPE papers point out that the variational posterior can "leak" mass outside the prior, i.e. put estimated posterior density in a region with zero prior density. *The bijection from the support of the variational posterior to the prior support solves leakage.* It is generally applicable when such a bijection can be constructed (holds for topologically isomorphic supports).

Going further

- The χ^2 divergence, equivalent to the balancing criterion, naturally leads to the following regularization term for K classes classification

$$\chi^2(\tilde{\pi}(y) \,\|\, \tilde{\varpi}(y)) = \frac{1}{K+1}\sum_{i=0}^{K}\left(\int \tilde{\varpi}(y = i \,|\, \Theta, \boldsymbol{x}) \left(\sum_{j=0}^{K} \tilde{\pi}(\Theta, \boldsymbol{x} \,|\, y = j)\right) d\Theta\, d\boldsymbol{x} - 1\right)^2.$$

The effect of this regularizer remains to be studied!

- We extend balancing to algorithms that provide an approximate posterior density however some methods do not fall into this framework (score-based methods, GANs, ...). Future work could reformulate our regularizer to apply to these works. It would require defining a purely sample-based (don't evaluate the estimated posterior density) version of the balancing criterion.

Take-home messages

- The balancing criterion can be expressed as the χ^2 divergence between the marginal classifier and target marginal distribution over classes. This provides a new perspective on balancing and serves as a building block for further development.
- The balancing criterion can be extended to algorithms that provide an approximate posterior density. This broadens the applicability of balancing, enabling more conservative algorithms.
- Empirically, balancing makes posteriors more conservative. Although, it may require a hyperparameter search to find the Lagrange multiplier that yields a conservative posterior estimate.