Neural Empirical Bayes: Source Distribution Estimation and its Applications to Simulation-Based Inference

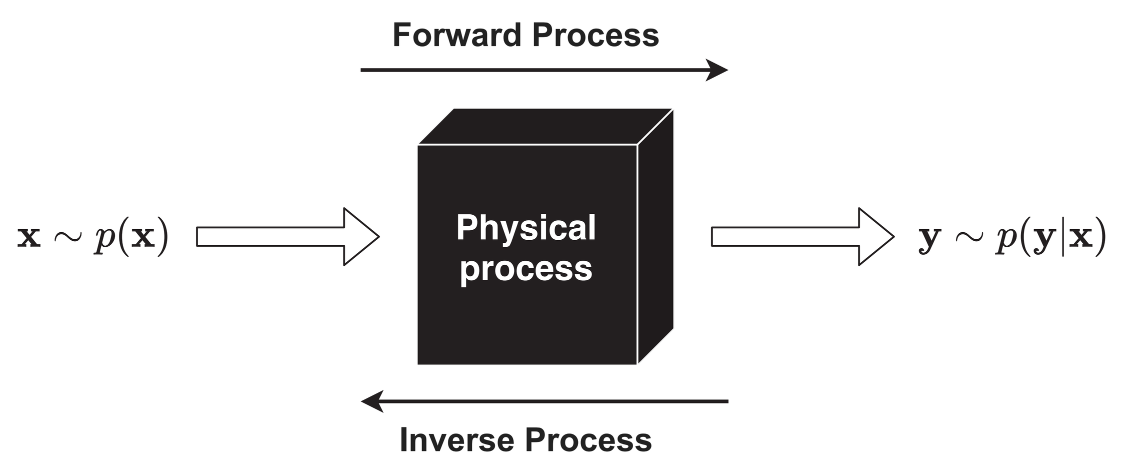
Maxime Vandegar, Michael Kagan, Antoine Wehenkel, Gilles Louppe

{maxime.vandegar, makagan}@slac.stanford.edu, {antoine.wehenkel, g.louppe}@uliege.be

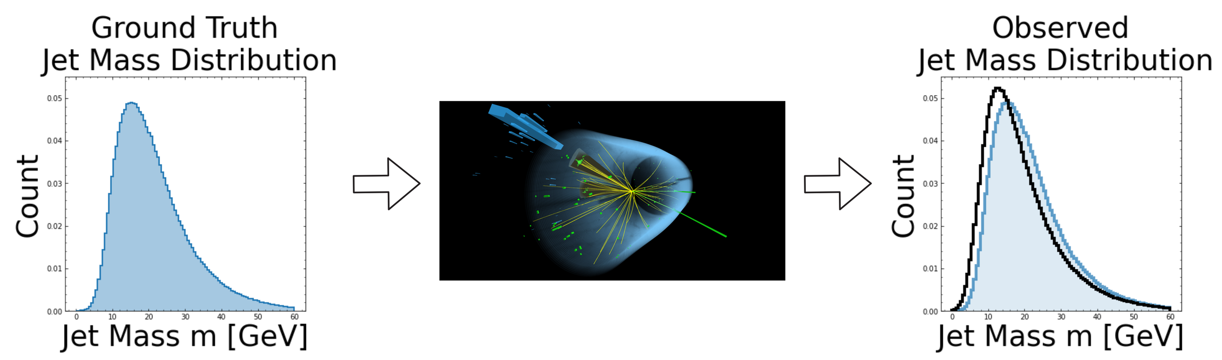
arXiv:2011.05836

Motivation

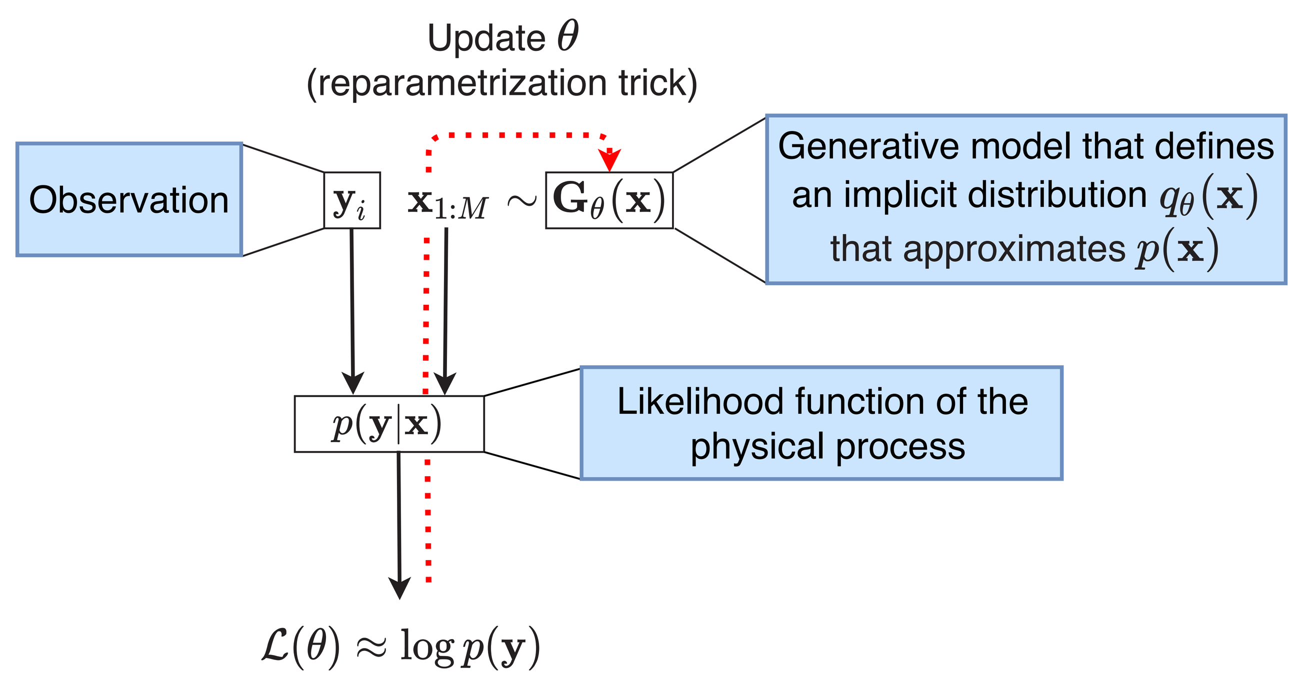
Aim to learn the *source distribution* $p(\mathbf{x})$ that has generated observations $\{\mathbf{y}_i\}_{i=1}^N$.



Applications:



Neural Empirical Bayes (NEB)



- **Goal:** build a model $q_\theta(\mathbf{x})$ of the source distribution, trained by maximizing the total log-marginal likelihood:

$$\log q_\theta(\mathbf{y}_i) = \log \int p(\mathbf{y}_i|\mathbf{x}) q_\theta(\mathbf{x}) d\mathbf{x}$$
$$= \log \mathbb{E}_{q_\theta}\left[p(\mathbf{y}_i|\mathbf{x})\right].$$

- **Ill-posed inverse problem:** relevant inductive bias can be embedded in the structure of $q_\theta(\mathbf{x})$.
- **Likelihood-free inference:** $p(\mathbf{y}|\mathbf{x})$ not known in closed-form, a surrogate model can be built with a density estimator (e.g. Normalizing Flow).

Monte Carlo estimators

Biased estimator

$$\mathcal{L}_K(\theta) = \log \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{y}|\mathbf{x}_k), \ \mathbf{x}_k = \mathbf{G}_\theta(\epsilon_k)$$
$$\epsilon_k \sim p(\epsilon).$$

- $\mathbb{E}\left[\mathcal{L}_{K+1}(\theta)\right] \geq \mathbb{E}\left[\mathcal{L}_K(\theta)\right]$.
- $\lim_{K \to \infty} \mathcal{L}_K(\theta) = \log p(\mathbf{y})$.

Unbiased estimator

$$\hat{\mathcal{L}}_K(\theta) = \mathcal{L}_K(\theta) + \eta(\theta),$$

where $\eta(\theta)$ is a random variable defined as

$$\eta(\theta) = \sum_{j=0}^{J} \frac{\mathcal{L}_{K+j+1}(\theta) - \mathcal{L}_{K+j}(\theta)}{P(\mathcal{J} \geq j)},$$

with $J \sim P(J)$ such that $P(\mathcal{J} \geq j) > 0, \forall j > 0$.

Variational estimators

ELBO estimator

If the source distribution model $\mathbf{G}_\theta(\cdot)$ allows explicit evaluation of $q_\theta(\mathbf{x})$, variational estimators can be used.

$$\mathcal{L}^{\text{ELBO}}(\theta, \psi) = \log p(\mathbf{y}|\mathbf{x}) - q_\psi(\mathbf{x}|\mathbf{y}) + q_\theta(\mathbf{x}),$$
$$\mathbf{x} \sim q_\psi(\mathbf{x}|\mathbf{y}).$$

- $\mathbb{E}\left[\mathcal{L}^{\text{ELBO}}(\theta, \psi)\right] \leq \log p(\mathbf{y})$.

IWAEs estimator

$$\mathcal{L}_K^{\text{IW}}(\theta, \psi) = \log \frac{1}{K}\sum_{k=1}^{K} \frac{p(\mathbf{y}|\mathbf{x}_k) q_\theta(\mathbf{x}_k)}{q_\psi(\mathbf{x}_k|\mathbf{y})}, \ \mathbf{x}_k \sim q_\psi(\mathbf{x}|\mathbf{y}).$$

- $\mathbb{E}\left[\mathcal{L}_{K+1}^{\text{IW}}(\theta, \psi)\right] \geq \mathbb{E}\left[\mathcal{L}_K^{\text{IW}}(\theta, \psi)\right]$.
- $\lim_{K \to \infty} \mathcal{L}_K^{\text{IW}}(\theta, \psi) = \log p(\mathbf{y})$.

Source estimation

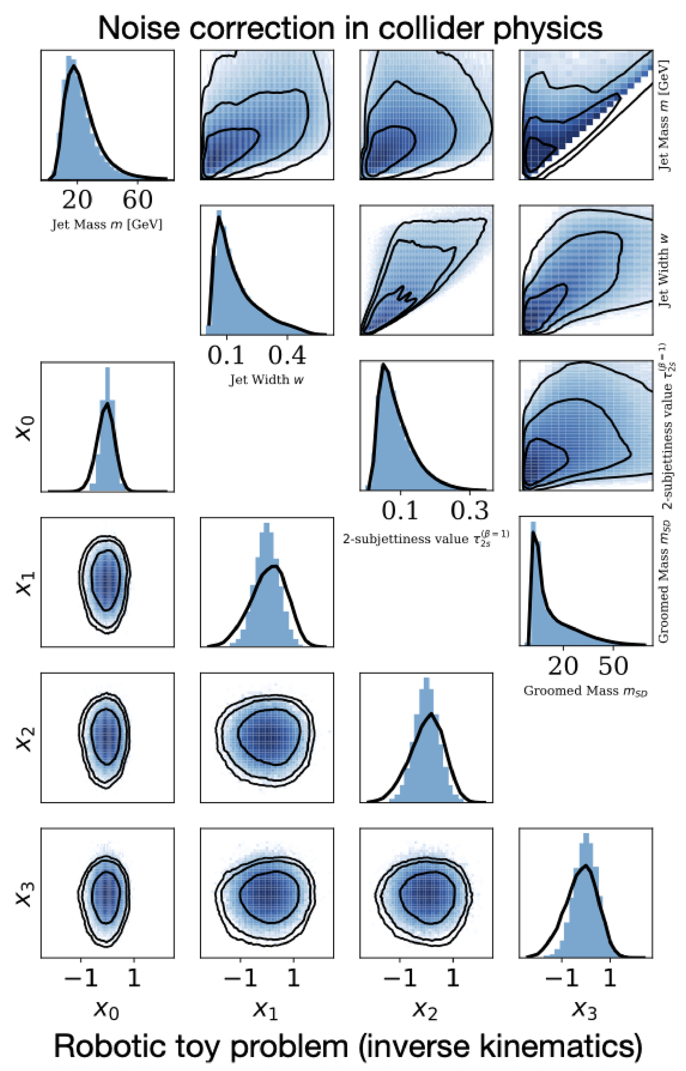


Figure 1: Source distributions $p(\mathbf{x})$ in blue against the estimated source distributions $q_\theta(\mathbf{x})$ in black.

- **De-biasing** can give substantial improvements when the number of Monte Carlo samples K is small. Little impact when K is large.
- **Source distribution** $q_\theta(\mathbf{x})$ learned with NEB can approximate $p(\mathbf{x})$ well in toy and collider physics experiments.

  Regenerated distributions in $\mathbf{y}$-space are indistinguishable from the observed distributions.
- **Inductive bias** in the density estimator, e.g. introducing symmetries / bounds / smoothness, leads to considerable improvements.

$\mathbf{y}$-space	K	$\mathcal{L}_K$	$\hat{\mathcal{L}}_K$
SLCP	10	$0.82_{\pm 0.01}$	$0.65_{\pm 0.04}$
	1024	$0.53_{\pm 0.01}$	$0.52_{\pm 0.01}$
Two-moons	10	$0.69_{\pm 0.02}$	$0.56_{\pm 0.02}$
	1024	$0.52_{\pm 0.01}$	$0.53_{\pm 0.01}$
Inverse kinematics	10	$0.80_{\pm 0.13}$	$0.67_{\pm 0.08}$
	1024	$0.66_{\pm 0.03}$	$0.62_{\pm 0.03}$

	$\mathcal{L}^{\text{ELBO}}$	$\mathcal{L}_{128}^{\text{IW}}$	$\mathcal{L}_{1024}$
$\mathbf{x}$-space	$0.99_{\pm 0.02}$	$0.63_{\pm 0.06}$	$0.57_{\pm 0.05}$
$\mathbf{y}$-space	$0.87_{\pm 0.08}$	$0.51_{\pm 0.01}$	$0.50_{\pm 0.01}$

$\mathbf{x}$-space	$\mathcal{L}^{\text{ELBO}}$	$\mathcal{L}_{128}^{\text{IW}}$	$\mathcal{L}_{1024}$
Unconstrained	$0.75_{\pm 0.00}$	$0.75_{\pm 0.00}$	$0.55_{\pm 0.02}$
Constrained	$0.51_{\pm 0.01}$	$0.50_{\pm 0.01}$	$0.51_{\pm 0.02}$

Metric: the ROC AUC of a classifier trained to discriminate between $p(\mathbf{x})$ and $q_\theta(\mathbf{x})$ ($\mathbf{x}$-space) as well as between $p(\mathbf{y})$ and $\int p(\mathbf{y}|\mathbf{x})q_\theta(\mathbf{x})d\mathbf{x}$ ($\mathbf{y}$-space). The closer to 0.5, the better.

Posterior inference

Once learned, $q_\theta(\mathbf{x})$ enables posterior inference with $p(\mathbf{y}|\mathbf{x})$. Using rejection sampling is efficient as $\mathbf{G}_\theta(\cdot)$ allows efficient parallel sampling.

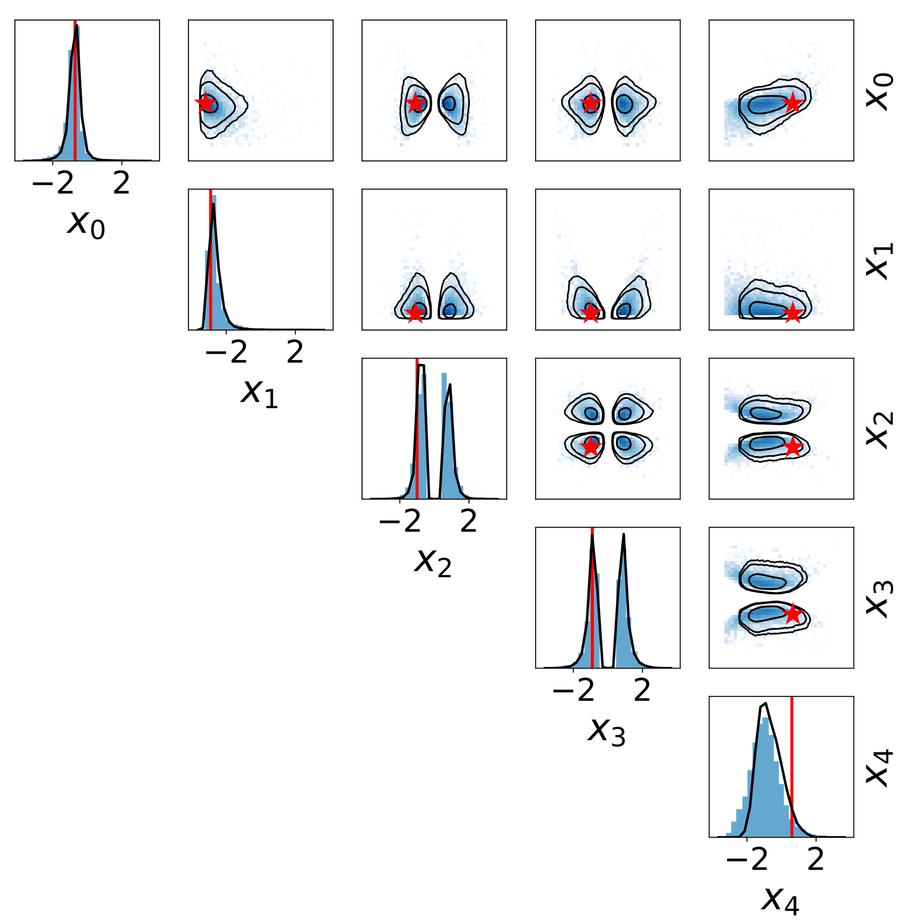


Figure 2: The exact posterior distribution in blue against its approximation in black obtained with $q_\theta(\mathbf{x})$ and a surrogate of $p(\mathbf{y}|\mathbf{x})$ using rejection sampling. The generating source data $\mathbf{x}^*$ are shown in red.

Take home messages

Neural Empirical Bayes learns data-informed priors / source distributions that successfully recover ground truths in a variety of experiments. NEB further enables posterior inference.

Source estimation and posterior inference can be performed efficiently without running a simulator at inference time.

The biased Monte Carlo estimator performs well while being computationally efficient.

Inductive bias helps mitigate the ill-posed nature of problems, and is easily introduced in the models.

References

[1] Thomas Mc Cauley. Display of a event observed in the CMS detector in which a Higgs boson decays to bottom quarks. CMS Collection., Oct 2018.

[2] Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. International Conference on Learning Representations, 2014.

[3] Yuri Burda, Roger B. Grosse, and Ruslan Salakhut- dinov. Importance weighted autoencoders. CoRR, abs/1509.00519, 2016.

[4] Yixin Wang, Andrew C. Miller, and David M. Blei. Comment: Variational autoencoders as empirical bayes. Statist. Sci., 34(2):229–233, 05 2019.