Mira★

A Score for Conditional Distribution Accuracy

A route to Bayesian model comparison without evidence

★: *Spotlight Paper*

Sammy Sharief*, Justine Zeghal*,

G. Missael Barco · P. Lemos · Y. Hezaveh · L. Perreault-Levasseur

Co-funded by the European Union

The problem

Conditional Generative Modeling: the task of learning $p(y \mid x)$ from samples $y, x \sim p(y, x)$ has become a **central tool in probabilistic modeling**.

The problem

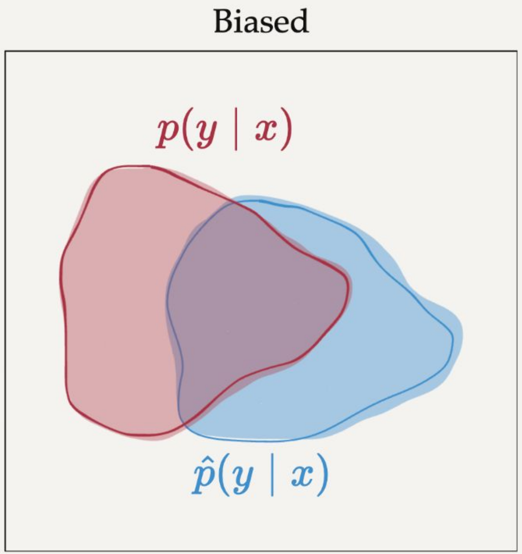
Conditional Generative Modeling: the task of learning $p(y \mid x)$ from samples $y, x \sim p(y, x)$ has become a **central tool in probabilistic modeling**.

But how to validate that the learned distribution $\hat{p}(y \mid x)$ is accurate?

The problem

Conditional Generative Modeling: the task of learning $p(y \mid x)$ from samples $y, x \sim p(y, x)$ has become a **central tool in probabilistic modeling**.

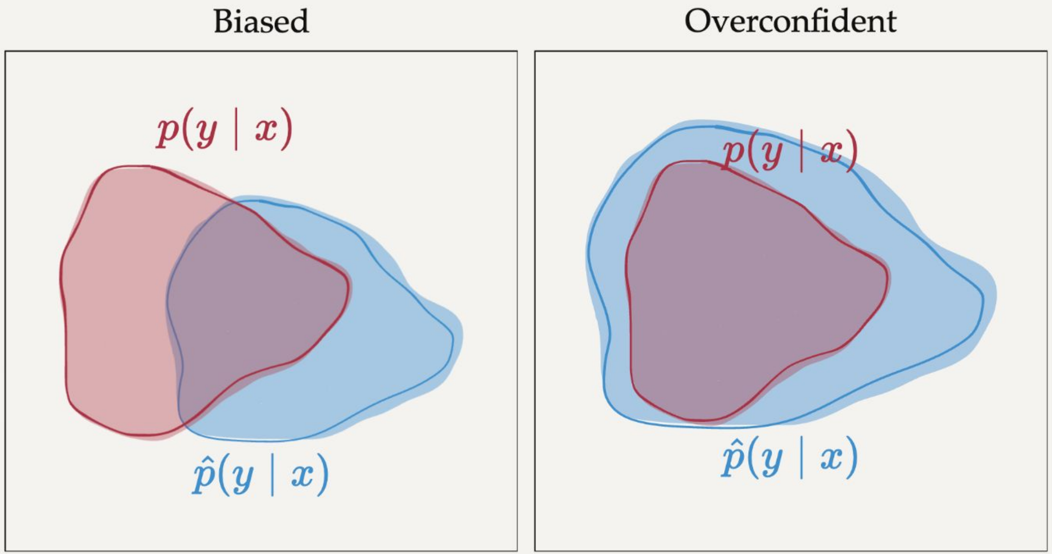
But how to validate that the learned distribution $\hat{p}(y \mid x)$ is accurate?



The problem

Conditional Generative Modeling: the task of learning $p(y \mid x)$ from samples $y, x \sim p(y, x)$ has become a **central tool in probabilistic modeling**.

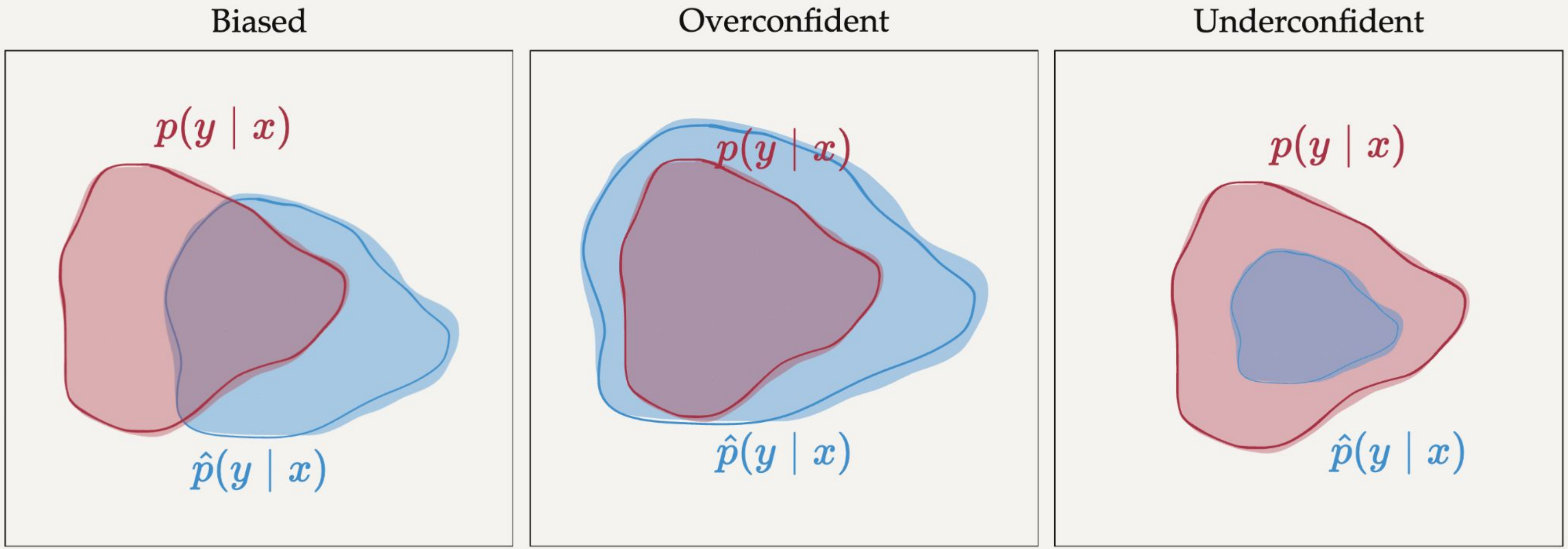
But how to validate that the learned distribution $\hat{p}(y \mid x)$ is accurate?



The problem

Conditional Generative Modeling: the task of learning $p(y \mid x)$ from samples $y, x \sim p(y, x)$ has become a **central tool in probabilistic modeling**.

But how to validate that the learned distribution $\hat{p}(y \mid x)$ is accurate?



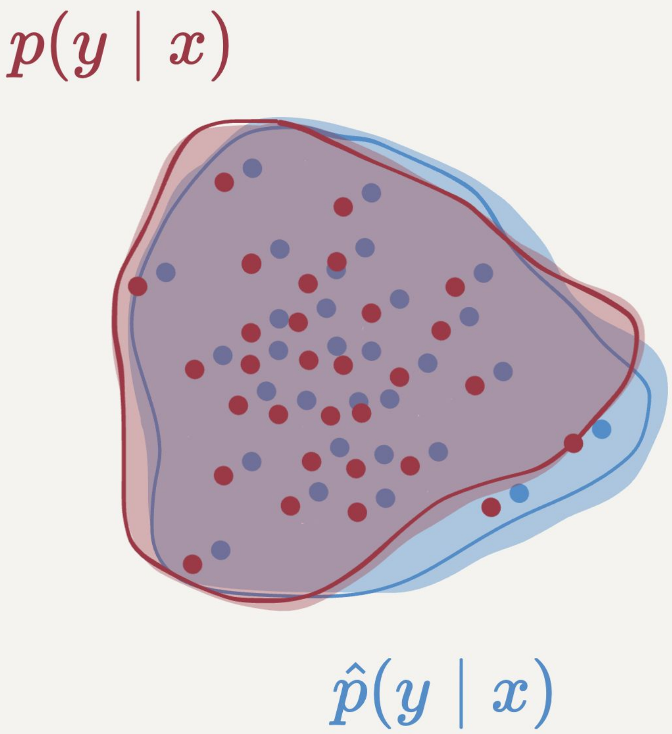
The Solution (inspired by PQMass, Lemos et al. 2025)

Theorem: two distributions are equal if their probability measures are the same over all measurable sets.

The Solution (inspired by PQMass, Lemos et al. 2025)

Theorem: two distributions are equal if their probability measures are the same over all measurable sets.

30 blue samples in total
30 red samples in total

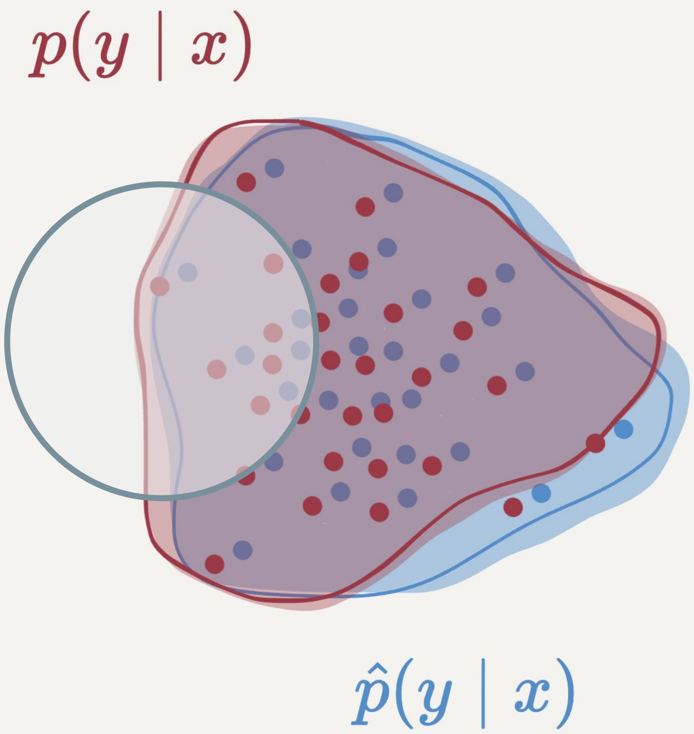


The Solution (inspired by PQMass, Lemos et al. 2025)

Theorem: two distributions are equal if their probability measures are the same over all measurable sets.

30 blue samples in total
30 red samples in total

n = 5 blue samples ✅
k = 6 red samples



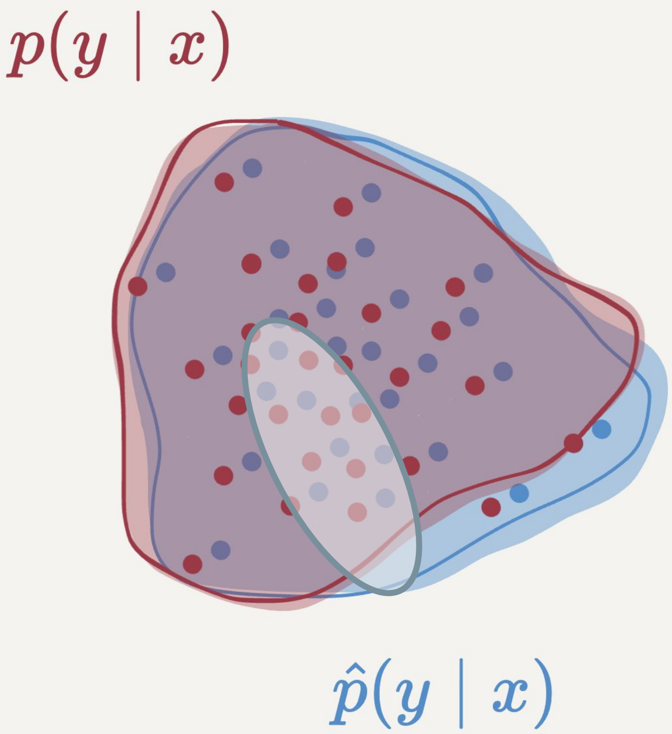
The Solution (inspired by PQMass, Lemos et al. 2025)

Theorem: two distributions are equal if their probability measures are the same over all measurable sets.

30 blue samples in total
30 red samples in total

n = 5 blue samples
k = 6 red samples ✅

n = 8 blue samples
k = 9 red samples ✅



The Solution (inspired by PQMass, Lemos et al. 2025)

> **Theorem**: two distributions are equal if their probability measures are the same over all measurable sets.

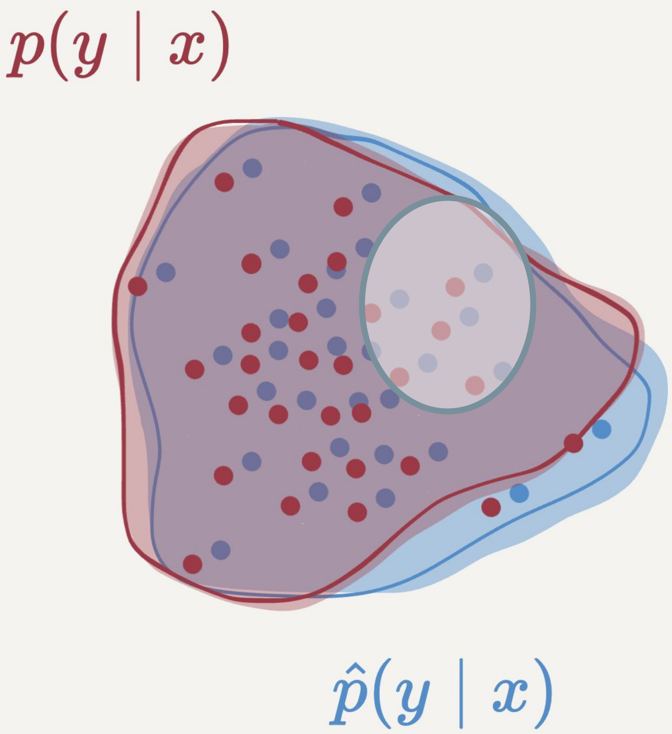
30 blue samples in total
30 red samples in total

n = 5 blue samples
k = 6 red samples ✅

n = 8 blue samples
k = 9 red samples ✅

n = 5 blue samples
k = 5 red samples ✅

⋮



The Solution (inspired by PQMass, Lemos et al. 2025)

Theorem: two distributions are equal if their probability measures are the same over all measurable sets.

30 blue samples in total

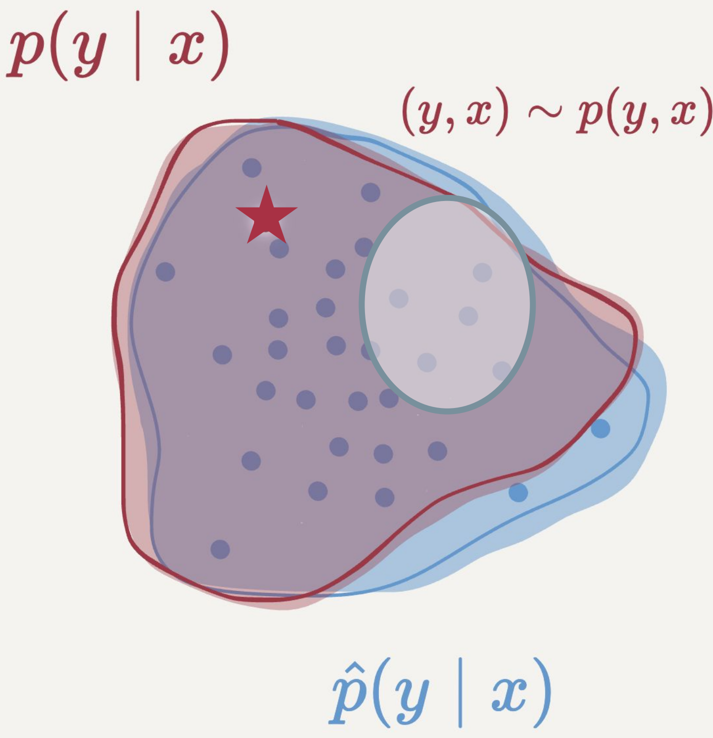
1 red sample in total!!

$n = 5$ blue samples
$k = 6$ red samples ✅

$n = 8$ blue samples
$k = 9$ red samples ✅

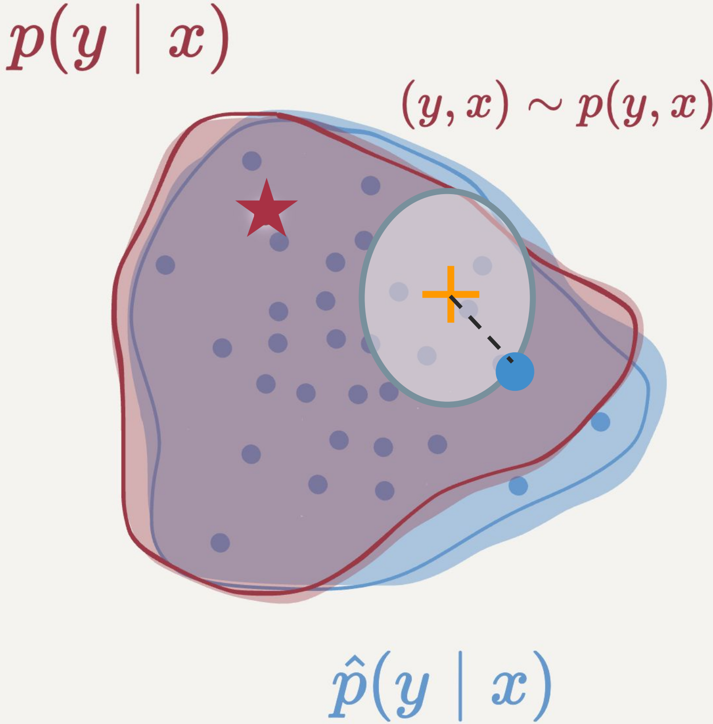
$n = 5$ blue samples
$k = 5$ red samples ✅

⋮



The Solution (inspired by PQMass, Lemos et al. 2025)

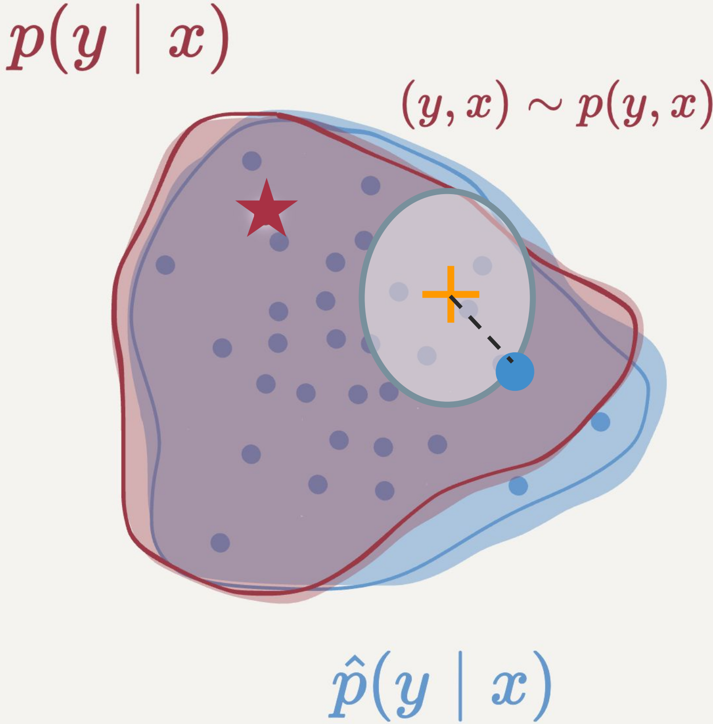
Bayesian approach: what is the probability that the true sample lies inside $(k = 0)$ or outside $(k = 1)$ the region $\mathcal{R}$ given that n samples from $\hat{p}(y \mid x)$ are in $\mathcal{R}$?



The Solution (inspired by PQMass, Lemos et al. 2025)

Bayesian approach: what is the probability that the true sample lies inside $(k = 0)$ or outside $(k = 1)$ the region $\mathcal{R}$ given that n samples from $\hat{p}(y \mid x)$ are in $\mathcal{R}$?

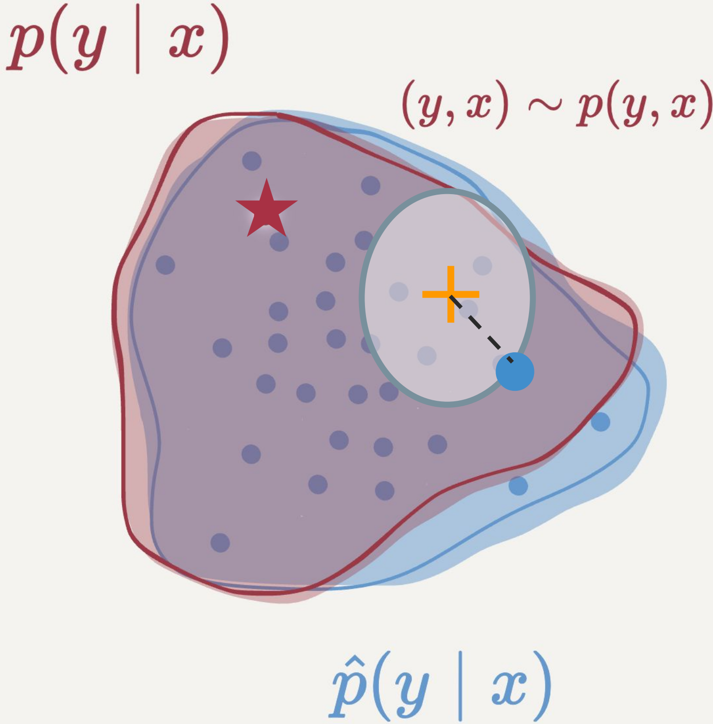
↳ $p(k \mid n)$



The Solution (inspired by PQMass, Lemos et al. 2025)

Bayesian approach: what is the probability that the true sample lies inside $(k = 0)$ or outside $(k = 1)$ the region $\mathcal{R}$ given that n samples from $\hat{p}(y \mid x)$ are in $\mathcal{R}$?

$$\mu = \mathbb{E}_{p(Z)}\left[\mathbb{E}_{p(k,n|Z)}\left[p(k \mid n)\right]\right]$$

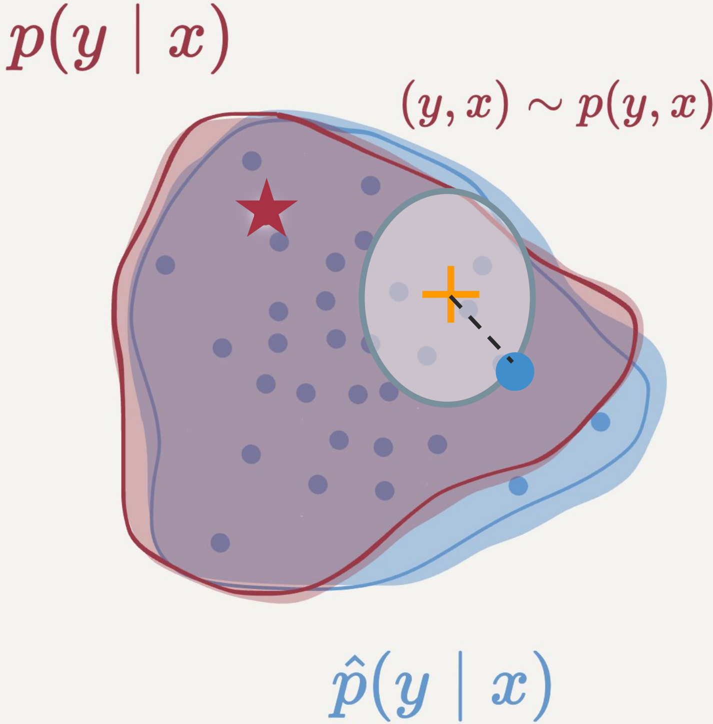


The Solution (inspired by PQMass, Lemos et al. 2025)

Bayesian approach: what is the probability that the true sample lies inside $(k=0)$ or outside $(k=1)$ the region $\mathcal{R}$ given that n samples from $\hat{p}(y \mid x)$ are in $\mathcal{R}$?

$$\mu = \mathbb{E}_{p(Z)}\left[\mathbb{E}_{p(k,n|Z)}\left[p(k \mid n)\right]\right]$$

Theorem: If $\hat{p}(y \mid x) = p(y \mid x)$ then $p(k \mid n) \sim Beta(2,1)$



The Solution (inspired by PQMass, Lemos et al. 2025)

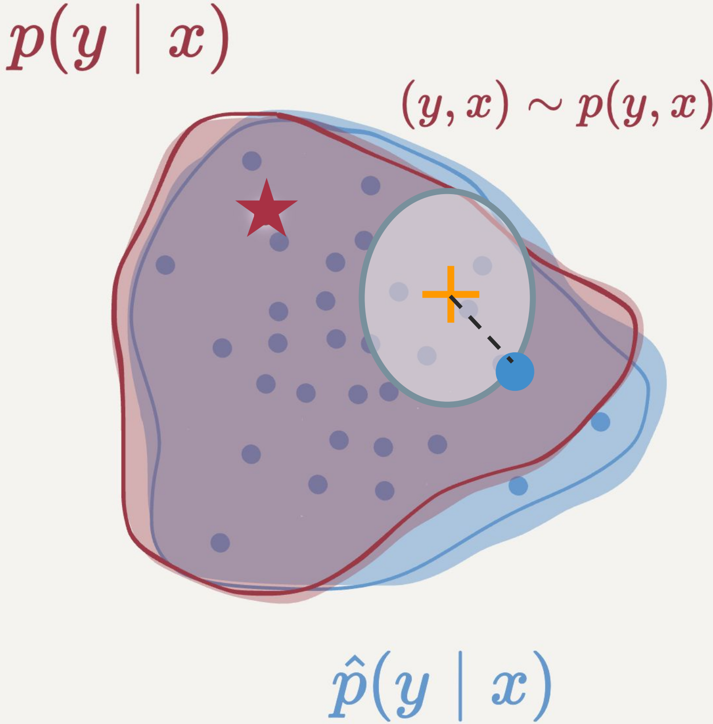
Bayesian approach: what is the probability that the true sample lies inside $(k = 0)$ or outside $(k = 1)$ the region $\mathcal{R}$ given that n samples from $\hat{p}(y \mid x)$ are in $\mathcal{R}$?

$$\mu = \mathbb{E}_{p(Z)}[\mathbb{E}_{p(k,n|Z)}[p(k \mid n)]]$$

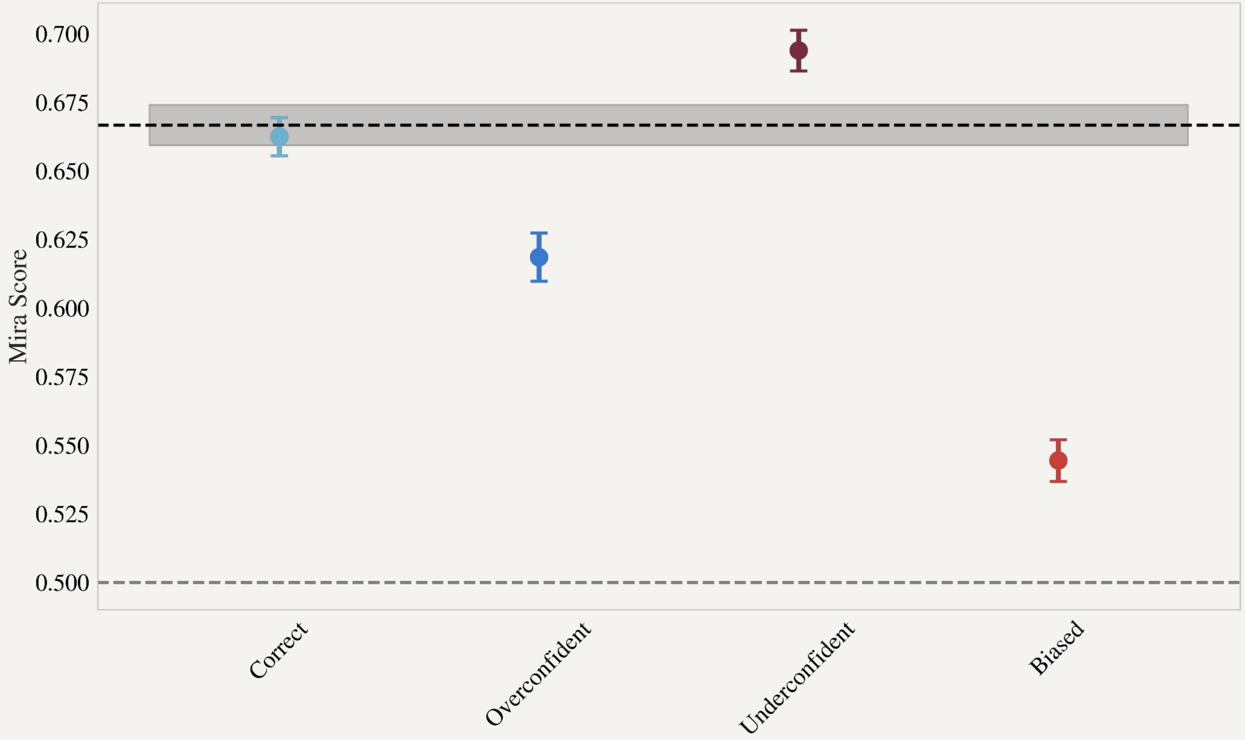
Theorem: If $\hat{p}(y \mid x) = p(y \mid x)$ then $p(k \mid n) \sim Beta(2, 1)$

Lemma: If $\hat{p}(y \mid x) = p(y \mid x)$ then

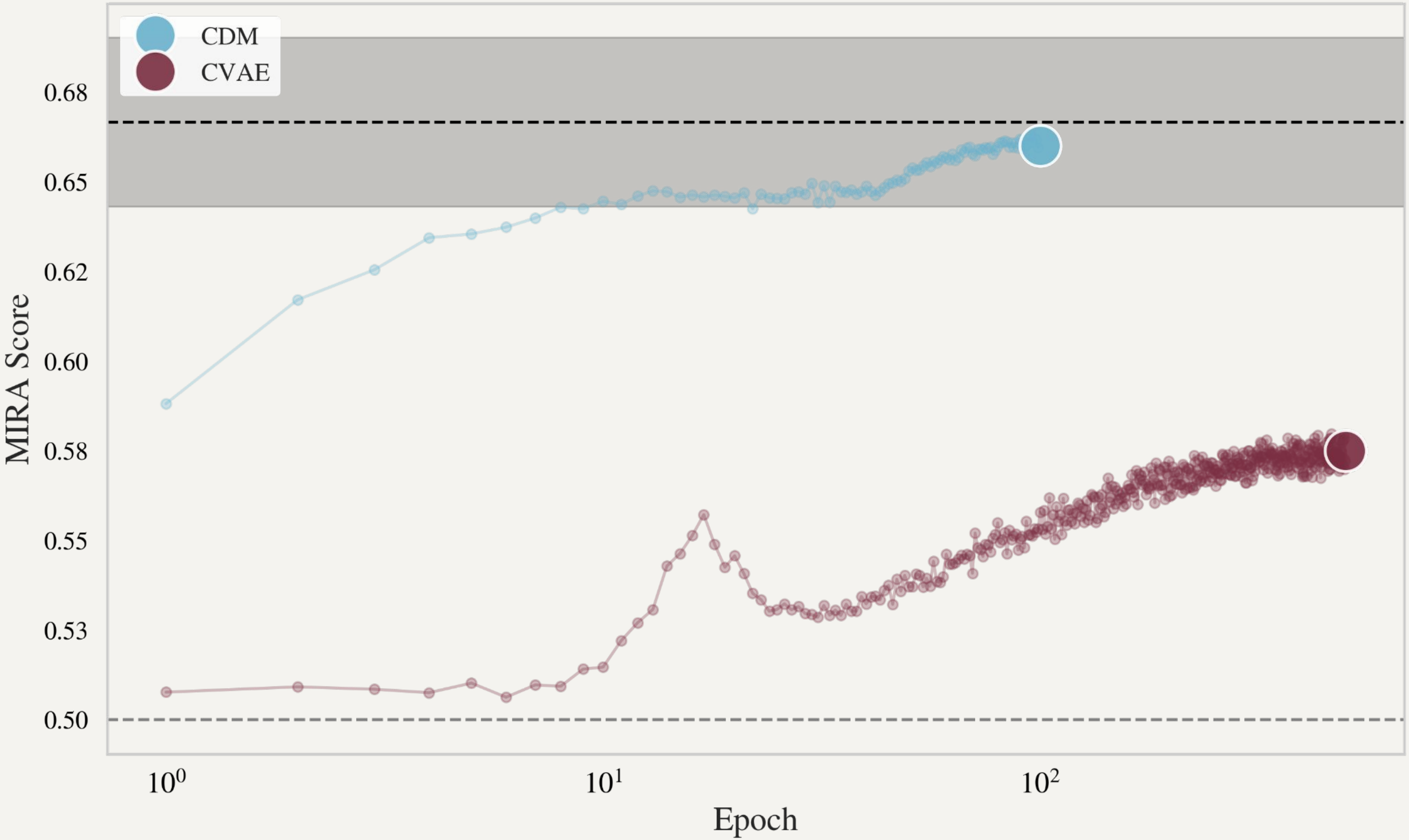
$$\mu_{\text{Mira}} = \mathbb{E}_{p(Z)}[\mathbb{E}_{p(k,n|Z)}[p(k \mid n)]] = 2/3$$



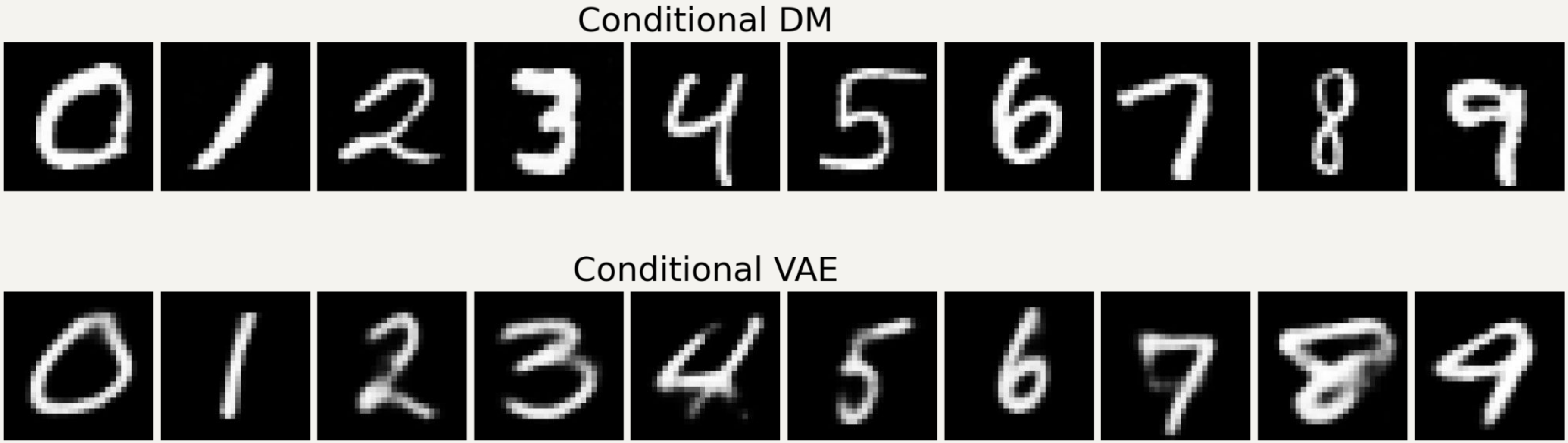
Detection of Various Pathological Scenarios



Evaluation of Conditional Generative Models



Evaluation of Conditional Generative Models



Takeaway message

Mira: Sample-based score for evaluating conditional distribution using only pairs of the true data generating process.

Takeaway message

Mira: Sample-based score for evaluating conditional distribution using only pairs of the true data generating process.

Usage of Mira:

- Evaluating the quality of conditional generative models
- Bayesian model comparison by switching the comparison in posterior space

Takeaway message

Mira: Sample-based score for evaluating conditional distribution using only pairs of the true data generating process.

Usage of Mira:

- Evaluating the quality of conditional generative models
- Bayesian model comparison by switching the comparison in posterior space

Benefits over existing validation scores/tests:

- Works in high-dimensions
- It is a scalar value (unlike for instance coverage tests)
- Works with few samples
- Does not rely on training a model (e.g. Local Classifier Two Sample Test, L-C2ST)
- We demonstrate that Mira detects miscalibration even when other methods fail.

Thank you for your attention!

The code

```
pip install mira_score
```