

# How does likelihood maximization work?

Many models are trained on some form of likelihood maximization ...

but why do we expect this to learn distributions?

$$D_{\text{KL}}( p(x) \parallel p_{\theta}(x) )$$

How does likelihood maximization lower KL divergence?

Proof in a few lines...

# How does a Normalizing Flow normalize?

Twist, expand and contract space to make things look normal.

- Needs to do so in a **smooth** and **invertible** manner => **Diffeomorphism**

