How does likelihood maximization work?

Many models are trained on likelihood maximization …

but why do we expect this to learn distributions?

$$D_{\mathbf{KL}}(p(x) \,||\, p_\theta(x))$$

How does likelihood maximization lower KL divergence?

Proof in a few lines…

How does likelihood maximization work?

How does likelihood maximization lower KL divergence?

Proof in a few lines…

$$\min_\theta D_{\mathbf{KL}}(\, p(x) \,||\, p_\theta(x)\,) = \min_\theta \mathbb{E}_{p(x)} \log \frac{p(x)}{p_\theta(x)}$$

$$= \min_\theta \mathbb{E}_{p(x)}[\log p(x) - \log p_\theta(x)]$$

$$= \min_\theta \mathbb{E}_{p(x)}[-\log p_\theta(x)]$$

$$= \max_\theta \mathbb{E}_{p(x)}[\log p_\theta(x)]$$

thus maximizing the likelihood brings the two distributions closer together.