How does likelihood maximization work?

Many models are trained on likelihood maximization …

but why do we expect this to learn distributions?

$$\max_{\theta} \mathbb{E}_{\mathbf{x}\sim p_x} p_\theta(\mathbf{x}) = \min_{\theta} \; \mathbb{E}_{\mathbf{x}\sim p_x} - \log p_\theta(\mathbf{x})$$

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…