How does likelihood maximization work?

Many models are trained on some form of likelihood maximization …

but why do we expect this to learn distributions?

How does likelihood maximization work?

Many models are trained on some form of likelihood maximization …

but why do we expect this to learn distributions?

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

How does likelihood maximization lower KL divergence?

Proof in a few lines…