

How does likelihood maximization work?

How does likelihood maximization lower KL divergence?

Proof in a few lines...

$$\begin{aligned}\min_{\theta} D_{\text{KL}}(p(x) \parallel p_{\theta}(x)) &= \min_{\theta} \mathbb{E}_{p(x)} \log \frac{p(x)}{p_{\theta}(x)} \\ &= \min_{\theta} \mathbb{E}_{p(x)} [\log p(x) - \log p_{\theta}(x)] \\ &= \min_{\theta} \mathbb{E}_{p(x)} [-\log p_{\theta}(x)] \\ &= \max_{\theta} \mathbb{E}_{p(x)} [\log p_{\theta}(x)]\end{aligned}$$

thus maximizing the likelihood brings the two distributions closer together.

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**

