(B) Unsupervised Data Augmentation (UDA)

UDA : Target model $p_\theta(y \mid x)$

$p_L(x)$ // Labeled data dist

$p_U(x)$ // Unlabeled data dist.

Perfect model f^*

Supervised Augmentation : $\hat{x} \sim q(\hat{x} \mid x)$

UDA : input x :

$$\left.\begin{array}{l} p_\theta(y \mid x) \\ p_\theta(y \mid x, \underbrace{\epsilon}_{\text{noise}}) \end{array}\right\} \xrightarrow[\text{Divergence}]{\text{minimize}} \mathcal{D}\left(p_\theta(y \mid x) \,\|\, p_\theta(y \mid x, \epsilon)\right)$$

$$\hat{x} = \overset{\text{quality??}}{q}(x, \epsilon)$$

Total Objective

$$\min_\theta \mathcal{J}(\theta) = \begin{cases} \overbrace{\mathbb{E}_{x_1 \sim P_L(x)}\left[-\log p_\theta(f^*(x_1) \mid x_1)\right]}^{\text{supervised}} \\ + \lambda \underbrace{\mathbb{E}_{x_2 \sim P_U(x)} \mathbb{E}_{\hat{x} \sim q(\hat{x} \mid x_2)}\left[CE\left(\underset{\text{fixed copy (no grad)}}{p_{\tilde{\theta}}}(y \mid x_2) \,\|\, \underset{\text{updated network}}{p_\theta}(y \mid \hat{x})\right)\right]}_{\text{unsupervised}} \end{cases}$$

Sharpening Prediction : indicator

$$\frac{1}{|B|} \sum_{x \in B} \overbrace{I\left(\max_{y'} p_{\tilde{\theta}}(y' \mid x) > \beta\right)}^{} CE\left(p_{\tilde{\theta}}^{\text{sharp}}(y \mid x) \,\|\, p_\theta(y \mid \hat{x})\right)$$

$$p_{\tilde{\theta}}^{\text{sharp}}(y \mid x) = \frac{\exp(z_y / \tau)}{\sum_{y'} \exp(z_{y'} / \tau)}$$

z_y → logit label for y