# TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Noisy Channel RDAs

# Noisy-Channel RDAs

In the image compression case study, training was based on a differentiable loss

$$\Phi^* = \operatorname*{argmin}_{\Phi} \Big(\; E_y\;\; -\ln p_\Phi(z_\Phi(y))\Big) + \lambda E_{y,\epsilon}\; \mathrm{Dist}(y, y_\Phi(z_\Phi(y)+\epsilon))$$

In a rate-distortion auto-encoder we will replace the conceptually dubious differential entropy rate term with a conceptually legitimate mutual information (channel capacity) rate term.

## Mutual Information as a Channel Rate

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \Big( \; E_y \; - \ln p_\Phi(z_\Phi(y))\Big) + \lambda E_{y,\epsilon} \, \mathrm{Dist}(y, y_\Phi(z_\Phi(y) + \epsilon))$$

is replaced by

$$\tilde{z} \; = \; z_\Phi(y) + f_\Phi(y, \epsilon) \;\; \epsilon \text{ is fixed (parameter independent) noise}$$

$$\Phi^* \; = \; \underset{\Phi}{\operatorname{argmin}} \;\; I_\Phi(y, \tilde{z}) + \lambda E_{y,\epsilon} \, \mathrm{Dist}(y, y_\Phi(\tilde{z}))$$

By the channel capacity theorem $I(y, \tilde{z})$ is the **rate** of information transfer from $y$ to $\tilde{z}$.

## Mutual Information as a Channel Rate

$$\tilde{z} = z_\Phi(y) + f_\Phi(y, \epsilon) \quad \epsilon \text{ is fixed (parameter independent) noise}$$

$$\Phi^* = \operatorname*{argmin}_\Phi \; I_\Phi(y, \tilde{z}) + \lambda E_{y,\epsilon} \operatorname{Dist}(y, y_\Phi(\tilde{z}))$$

Taking the distribution on $\epsilon$ to be parameter independent is called the "reparameterization trick" and allows SGD.

$$\nabla_\Phi \, E_{y,\epsilon} \operatorname{Dist}(y, y_\Phi(z_\Phi(y) + f_\Phi(y, \epsilon)))$$

$$= E_{y,\epsilon} \, \nabla_\Phi \operatorname{Dist}(y, y_\Phi(z_\Phi(y) + f_\Phi(y, \epsilon)))$$

## Mutual Information as a Channel Rate

$$\tilde{z} = z_\Phi(y) + f_\Phi(y, \epsilon) \quad \epsilon \text{ is fixed (parameter independent) noise}$$

$$\Phi^* = \operatorname*{argmin}_\Phi \; I_\Phi(y, \tilde{z}) + \lambda E_{y,\epsilon} \,\mathrm{Dist}(y, y_\Phi(\tilde{z}))$$

Typically $f_\Phi(y, \epsilon)$ is simple, such as $\sigma_\Phi(y) \odot \epsilon$, so that $p_\Phi(\tilde{z}|y)$ is easily computed.

# Mutual Information Replaces Cross Entropy

$$I_\Phi(y, \tilde{z}) \;=\; E_{y,\epsilon}\;\ln\frac{\mathrm{pop}(y)p_\Phi(\tilde{z}|y)}{\mathrm{pop}(y)p_{\mathrm{pop},\Phi}(\tilde{z})}$$

$$=\; E_{y,\epsilon}\;\ln\frac{p_\Phi(\tilde{z}|y)}{p_{\mathrm{pop},\Phi}(\tilde{z})}$$

where $\quad p_{\mathrm{pop},\Phi}(\tilde{z}) \;=\; E_{y\sim\mathrm{pop}}\;\; p_\Phi(\tilde{z}|y)$

## A Variational Bound

$$p_{\mathrm{pop},\Phi}(\tilde{z}) = E_{y\sim \mathrm{pop}}\;\; p_\Phi(\tilde{z}|y)$$

We cannot compute $p_{\mathrm{pop},\Phi}(\tilde{z})$.

Instead we will use a variational bound involving a computable model $q_\Phi(\tilde{z})$

# A Variational Bound

$$
\begin{aligned}
I(y, \tilde{z}) &= E_{y,\epsilon}\ \ln \frac{p_\Phi(\tilde{z}|y)}{p_{\mathrm{pop},\Phi}(\tilde{z})} \\
&= E_{y,\epsilon}\ \ln \frac{p_\Phi(\tilde{z}|y)}{q_\Phi(\tilde{z})} + E_{y,\epsilon}\ \ln \frac{q_\Phi(\tilde{z})}{p_{\mathrm{pop},\Phi}(\tilde{z})} \\
&= E_{y,\epsilon}\ \ln \frac{p_\Phi(\tilde{z}|y)}{q_\Phi(\tilde{z})} - KL(p_{\mathrm{pop},\Phi}(\tilde{z}), q_\Phi(\tilde{z})) \\
&\leq E_{y,\epsilon}\ \ln \frac{p_\Phi(\tilde{z}|y)}{q_\Phi(\tilde{z})}
\end{aligned}
$$

# A Fundamental Equation for the Continuous Case

$$\tilde{z} = z_\Phi(y) + f_\Phi(y, \epsilon)$$

$$\Phi^* = \operatorname*{argmin}_\Phi E_{y,\epsilon} \; \ln \frac{p_\Phi(\tilde{z}|y)}{q_\Phi(\tilde{z})} + \lambda \mathrm{Dist}(y, y_\Phi(\tilde{z}))$$

**END**