At t_i

$H_i \longrightarrow$ N.N $\longrightarrow C_i$ $\Rightarrow$ ~~Ct~~ Minimize the lower bound of $I(x_{i+1}, C_i)$

*CPC paper

I.B.

- There exists a structured dist: $p(H_i, X_{i+1})$
- Relevant info in $H_i \equiv I(H_i; X_{i+1})$
- $\hat{H}_i$ = minimal sufficient statistics,
  = simplest mapping of H that captures $I(H_i, X_{i+1})$
- Best representation found by: minimizing

$$\mathcal{L}[p(\hat{H}|H)] = \underbrace{I(H;\hat{H})}_{\text{compressive}} - \beta\underbrace{(I(\hat{H}; X_{i+1}))}_{\text{"predictive info preserving"}}$$

- Data processing inequality:

$$I(H, X_{i+1}) \geq I(\hat{H}, X_{i+1}) \geq I(\hat{X}_{i+1}, X_{i+1})$$

- Rate distortion theory:
  How much can you compress input & maintain a distortion less/eq. D on reconstruction?

- Distortion Measure

$$d(\mathcal{H},\hat{\mathcal{H}}) = KL[p(x_{i+1}|\mathcal{H}_i) \| p(x_{i+1}\|\hat{\mathcal{H}})]$$

- ~~When $\hat{\mathcal{H}}$ optimized sufficient statistic:~~

$$\mathbb{E}[d(\mathcal{H},\hat{\mathcal{H}})] = I(\mathcal{H}; x_{i+1}|\hat{\mathcal{H}})$$

↑ Residual info in history about x_{i+1} not captured by $\hat{\mathcal{H}}$.

↑ Want to reduce this value.

- Rewrite the loss:

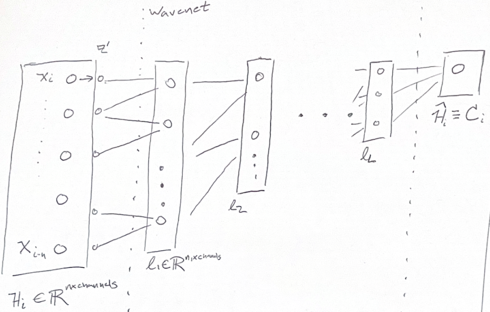
$$\mathcal{L} = \underbrace{I(\mathcal{H};\hat{\mathcal{H}})}_{\text{compression}} + \beta\underbrace{(I(\mathcal{H}; x_{i+1}\|\hat{\mathcal{H}}))}_{\text{Residual Information}}.$$

- IB Distortion

$$D_N = I(\mathcal{H}_i; x_{i+1}|\hat{x}_{i+1})$$

- Represantational Complexity of Output

$$R_N = I(\mathcal{H}_i; \hat{x}_{i+1})$$



IB Goals;

1. minimize $I(H_i; C_i)$
2. maximize $I(C_i; X_{i+1})$ *

   OR

   minimize $I(H_i; X_{i+1} | C_i)$

CPC Training:

Maximize LB of $I(X_{i+1}; C_i)$*

Dimensionality

$H_i = \mathbb{R}^{n \times channels}$

$Z' = \mathbb{R}^{n \times latent\_dim}$

$l_i = \mathbb{R}^{(f(\{dilation_j, kernel_j\}|_{j=0}^{i}, features_i)}$

$C_i = \mathbb{R}^{context\_dim}$