Noise Contrastive Estimation

Karl Stratos

In prediction problems, we’re supposed to predict $y \in \mathcal{Y}$ from $x \in \mathcal{X}$. We do this by assuming a joint population distribution $p_{XY}$ from which we can sample correct pairs $(x, y)$ and learning a score function $s^\theta(x, y) \in \mathbb{R}$ parameterized by $\theta$ such that it assigns a high score to a correct pair and a low score to an incorrect pair. To estimate such a score function, we often use the hinge loss (Appendix A) or the cross-entropy loss (Appendix B).

In noise contrastive estimation (NCE), we choose a “noise” distribution $q_Y$ over $\mathcal{Y}$ and the size of a sample set $N$ and consider the task of distinguishing true samples from fake samples. It underlies many successful methods such as word2vec [7], the generative adversarial networks (GANs) [3], and contrastive predictive coding [8]. It has two popular formulations. 1. **Global**: Infer which of the $N$ samples is true. 2. **Local**: For each individual sample infer if it’s true.

Information theory enables a simple and insightful analysis of NCE. Given any distribution $p$, if $q^\theta$ is a distribution over the same variables parameterized by $\theta$, $q^\theta$ is equal to $p$ iff it is the minimizer of the cross entropy between $p$ and $q^\theta$:

$$\theta^* \in \arg \min_\theta \mathbb{E}_{z \sim p} [-\log q^\theta(z)] \iff q^\theta^*(z) = p(z) \quad \forall z$$

assuming the **universality** of $q^\theta$: that is, it is expressive enough to model $p$ so that $p = q^\theta$ for some $\theta$. While universality should be assumed with a grain of salt (e.g., it might require an exponentially large parameter space), it seems to hold in practice with neural networks and greatly simplifies analysis.

1 **Global NCE**

1.1 **Model**

The global NCE objective assumes a joint distribution

$$p^q_{Y|X,N}(i, x, y_1 \ldots y_N) := \frac{1}{N} p_{XY}(x, y_i) \prod_{j \neq i} q_Y(y_j)$$

That is, we first draw an index $i \in \{1 \ldots N\}$ uniformly at random and for $j = 1 \ldots N$ draw $(x, y_j) \sim p_{XY}$ if $j = i$ but otherwise draw $y_j \sim q_Y$. This yields a conditional distribution over $N$ indices

$$p^q_{i|X,Y,N}(i|x, y_1 \ldots y_N) = \frac{p_{Y|X}(y_i|x) \prod_{j \neq i} q_Y(y_j)}{\sum_{k=1}^N p_{Y|X}(y_k|x) \prod_{j \neq k} q_Y(y_j)} = \frac{p_{Y|X}(y_i|x)}{\sum_{k=1}^N p_{Y|X}(y_k|x) q_Y(y_k)} \quad (1)$$

1
Let $H^q_v(I|XY^N)$ denote the conditional entropy of $p^q_v_{I|XY^N}$. The following observation is made in [8].

**Lemma 1.1.** $H^p_v(I|XY^N) \geq \log N - I(X,Y)$ where $I(X,Y)$ is the mutual information between $(x,y) \sim p_{XY}$.

**Proof.** By (1),

$$E_{(i,x,y_{1:Y}) \sim p^v_{I|XY^N}} \left[ -\log p^v_{I|XY^N}(i|x,y_{1:Y}) \right]$$

$$= -E_{(x,y) \sim p_{XY}} \left[ \frac{p_{Y|X}(y|x)}{p_Y(y)} \right] + \sum_{(i,x,y_{1:Y}) \sim p^v_{I|XY^N}} \left[ \log \sum_{k=1}^{N} \frac{p_{Y|X}(y_k|x)}{p_Y(y_k)} \right] \geq \log N$$

We will not prove the claim that the second term is at least log $N$, but it is intuitive since $p_{Y|X}(y|x) \approx p_Y(y)$ if $y \sim p_Y$ and $p_{Y|X}(y|x) \geq p_Y(y)$ if $y \sim p_{Y|X}(|x)$. A formal proof can be found in [9].

**Corollary 1.2.** $B := \log N - H^p_v(I|XY^N) \leq \min \{I(X,Y), \log N\}$.

**Proof.** The claim that $B \leq I(X,Y)$ follows by rearranging terms in Lemma 1.1. The claim that $B \leq \log N$ follows from the fact that $H^p_v(I|XY^N) \geq 0$ (Shannon entropy is nonnegative).

### 1.2 Estimation

We use a score function $s^\theta(x,y)$ through the softmax function to estimate $p^q_v_{I|XY^N}$

$$p^\theta_{I|XY^N}(i|x, y_{1:Y}) := \frac{\exp \left( s^\theta(x,y_i) \right)}{\sum_{j=1}^{N} \exp \left( s^\theta(x,y_j) \right)} \quad \forall i \in \{1 \ldots N\}$$

We train the model by minimizing the cross (conditional) entropy between $p^q_v_{I|XY^N}$ and $p^\theta_{I|XY^N}$:

$$B^\theta := \sum_{i,x,y_{1:Y}} - \log p^\theta_{I|XY^N}(i|x,y_{1:Y})$$

Note that $B^\theta(I|XY^N) \geq H^q_v(I|XY^N)$ for all $\theta$ by the usual property of cross entropy.

If $q_Y = p_Y$, Corollary 1.2 implies that

$$B(\theta) := \log N - \tilde{B}^\theta(I|XY^N) \leq \log N - H^p_v(I|XY^N) \leq \min \{I(X,Y), \log N\}$$

Thus minimizing $\tilde{B}^\theta(I|XY^N)$ over $\theta$ corresponds to maximizing a parameterized lower bound $B(\theta)$ on $I(X,Y)$, and for this reason global NCE is sometimes called “InfoNCE”. This lower bound cannot be greater than log $N$, which is consistent with the result in [6].
Let $\theta^{q_Y} \in \arg \min_\theta \tilde{H}_\theta^{q_Y} (I|XY^N)$. By universality we must have $p_{I|XY^N}^{\theta^{q_Y}} = p_{I|XY^N}^{0}$. By (1) this means

$$s^{q_Y} (x, y) = \log \frac{p_{Y|X}(y|x)}{q_Y(y)} + \log C_x \quad \forall x \in \mathcal{X}, \ y \in \mathcal{Y}$$

for some constant $C_x > 0$. In particular, we can use the optimal parameter $\theta^{q_Y}$ to recover the underlying conditional distribution

$$p_{Y|X}(y|x) = \frac{\exp \left( s^{q_Y} (x, y) + \log q_Y(y) \right)}{\sum_{y'} \exp \left( s^{q_Y} (x, y') + \log q_Y(y') \right)} \quad (2)$$

This is consistent with the “ranking” algorithm in [5]. A small modification of global NCE gives an unbiased gradient estimator of the cross entropy loss [1, 2] (Appendix C).

## 2 Local NCE

### 2.1 Model

The local NCE objective assumes a biased coin with head probability $1/N$, which we define by $p_A(1) = 1/N$ and $p_A(0) = (N-1)/N$. Given $x \sim p_X$ and $a \sim p_A$, it defines

$$p_{Y|X}^{q}(a|x, y) := \left\{ \begin{array}{ll} p_{Y|X}(y|x) & \text{if } a = 1 \\ q_Y(y) & \text{if } a = 0 \end{array} \right. \quad (3)$$

This yields the conditional head probability

$$p_{A|XY}^{q_Y}(1|x, y) = \frac{p_{Y|X}(y|x)}{p_{Y|X}(y|x) + (N-1)q_Y(y)}$$

Given $x \sim p_X$ and $N$ iid samples $a_i \sim p_A$ and $y_i \sim p_{Y|X}^{q_Y}(1|x, a_i)$ for $i = 1 \ldots N$, the joint conditional probability of the coin flips is given by

$$p_{A^{N}|XY^{N}}^{q_Y}(a_1 \ldots a_N|x, y_1 \ldots y_N) = \prod_{i=1}^{N} p_{A|X}^{q_Y}(1|x, y_i) \prod_{j=1}^{N} (1 - p_{A|X}^{q_Y}(1|x, y_j))$$

Let $H^{q_Y}(A^N|XY^N)$ denote the conditional entropy of $p_{A^{N}|XY^{N}}^{q_Y}$. We write it in the friendlier form (see Appendix D for details)

$$H^{q_Y}(A^N|XY^N) = \mathbb{E}_{(x,y) \sim p_{XY}} \left[ -\log p_{A|X}^{q_Y}(1|x, y) \right] + (N-1) \mathbb{E}_{y \sim q_Y} \left[ -\log (1 - p_{A|X}^{q_Y}(1|x, y)) \right] \quad (4)$$

The following lemma can be easily shown by plugging in (3) into (4) (again see Appendix D for details).

**Lemma 2.1.** Let $KL(p||q)$ denote the KL divergence between distributions $p$ and $q$. Then

$$-H^{q_Y}(A^N|XY^N) = KL \left( p_{Y|X} \left\| \frac{p_{Y|X} + (N-1)q_Y}{N} \right\| \right) + (N-1)KL \left( q_Y \left\| \frac{p_{Y|X} + (N-1)q_Y}{N} \right\| \right)$$

$$- \log N - (N-1) \log \left( \frac{N}{N-1} \right)$$

3
Corollary 2.2. Let $JSD(p||q) = \frac{1}{2}KL(p||\frac{p+q}{2}) + \frac{1}{2}KL(q||\frac{p+q}{2}$ denote the Jensen-Shannon divergence. With $N = 2$ we have from Lemma 2.1

$$-H^{q_Y}(A^2|XY^2) = 2JSD\left(\left\| p_{Y|X} \right\| q_Y \right) - \log 4$$

To make the connection to GANs [3] clear, let $|\mathcal{X}| = 1$ and eliminate the dependence on $X$. Recall the adversarial objective of GANs and its equilibrium:

$$GAN(D, q_Y) := E_{y \sim p_Y} \left[ \log D(1|y) \right] + E_{y \sim q_Y} \left[ \log(1 - D(1|y)) \right]$$

$$J_{GAN} := \min q_Y \max D_{GAN}(D, q_Y)$$

where $D : \mathcal{Y} \rightarrow [0, 1]$ is a discriminator and $q_Y$ is viewed as a generator. It can be verified that setting $D(1|y) = p_{A|Y}(1|y) = p_Y(y)/(p_Y(y) + q_Y(y))$ (3) maximizes $GAN(D, q_Y)$ for any $q_Y$. But $GAN(p_{A|Y}^{q_Y}, q_Y) = -H^{q_Y}(A^2|Y^2)$, thus by Corollary 2.2

$$J_{GAN} = \min q_Y GAN(p_{A|Y}^{q_Y}, q_Y) = \min q_Y 2JSD\left(\left\| p_{Y|X} \right\| q_Y \right) - \log 4 = -\log 4$$

where the minimizer is $q_Y = p_Y$. At this equilibrium, we see that the best discriminator is uniform $p_{A|Y}^{q_Y}(1|y) = 1/2$ and the generator “wins”.

2.2 Estimation

We use a score function $s^\theta(x, y)$ through the sigmoid function to estimate $p_{A|X}^{q_Y}$

$$p_{A|X}^{q_Y}(1|x, y) := \frac{1}{1 + \exp(-s^\theta(x, y))}$$

This is used to define the joint conditional distribution

$$p_{A|X,Y}^\theta(a_1 \ldots a_N|x, y_1 \ldots y_N) = \prod_{i=1}^N p_{A|X}^\theta(1|x, y_i) \prod_{j=1}^N (1 - p_{A|X}^\theta(1|x, y_j))$$

The model is again estimated by minimizing the cross (conditional) entropy between $p_{A|X,Y}^\theta$ and $p_{A|X,Y}^\theta$. Similar to (4) this objective can be written in the friendlier form

$$\theta^{q_Y} \in \arg \max_{\theta} E_{(x,y) \sim p_{XY}} \left[ \log p_{A|X,Y}^\theta(1|x, y) \right] + (N-1) E_{y \sim q_Y} \left[ \log(1 - p_{A|X,Y}^\theta(1|x, y)) \right]$$

By universality we must have $p_{A|X,Y}^{\theta^{q_Y}} = p_{A|X,Y}^{q_Y}$. By (3) this means

$$s^{\theta^{q_Y}}(x, y) = \log \frac{p_{Y|X}(y|x)}{q_Y(y)} - \log(N - 1) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

If $q_Y = p_Y$, the optimal score of $(x, y)$ is the pointwise mutual information (PMI) minus the log of the number of negative examples: this gives the analysis of the skip-gram.
objective of word2vec in [4]. We can use the optimal parameter $\theta^{iv}$ to recover the underlying conditional distribution

$$p_{Y|X}(y|x) = \exp\left(s^{\theta^{iv}}(x, y) + \log q_Y(y) + \log(N - 1)\right)$$

This is consistent with the “binary” algorithm in [5]. Note that unlike (2) this calculation doesn’t require normalization. This implies that the score function must self-normalized (Assumption 2.2 in [5]), that is we must be able to at least find $\theta$ such that

$$\sum_y \exp\left(s^\theta(x, y) + \log q_Y(y) + \log(N - 1)\right) = 1 \quad \forall x \in \mathcal{X}$$

This is a strong assumption when $|\mathcal{X}|$ is larger than the number of variables in $\theta$, so universality cannot be taken for granted in this case.
A Hinge Loss

We want to find \( \theta \) that maximizes the probability of the event that \( s^\theta(x, y) > s^\theta(x, y') \) for all \( y' \neq y \). This is equivalent to minimizing the zero-one loss

\[
\arg \min_\theta \mathbb{E}_{(x,y) \sim p_{XY}} \begin{cases} 
\text{zero-one loss on } (x,y) \\
\mathbb{I} \left( s^\theta(x, y) - \max_{y' \neq y} s^\theta(x, y') \leq 0 \right) \\
\text{margin of } (x,y)
\end{cases}
\]

where \( \mathbb{I}(\cdot) \in \{0,1\} \) is the indicator function. The indicator function is difficult to optimize for a number of reasons (e.g., it has zero gradient almost everywhere wrt the margin), so we instead define the hinge loss

\[
\arg \min_\theta \mathbb{E}_{(x,y) \sim p_{XY}} \begin{cases} 
\text{hinge loss on } (x,y) \\
\max \left\{ 0, 1 - \left( s^\theta(x, y) - \max_{y' \neq y} s^\theta(x, y') \right) \right\} \\
\text{margin of } (x,y)
\end{cases}
\]

Note that for any fixed \((x,y)\), the hinge loss on \((x,y)\) is a convex upper bound on the zero-one loss on \((x,y)\) where the convexity is wrt the margin of \((x,y)\).

In some applications, it’s neither necessary nor useful to exactly maximize over the negative space \( \{y' \in \mathcal{Y} : y' \neq y\} \) to compute the margin. This is because the search is intractable and/or exact maximization has some undesirable quality (e.g., it’s in fact an alternative viable prediction). In this case, maximization is replaced by sampling [11].

B Cross-Entropy Loss

We frame the problem as conditional density estimation of \( p_{Y|X} \). To this end, we turn the score function into a proper conditional distribution by using the softmax operation:

\[
p_{Y|X}^\theta(y|x) := \frac{\exp \left( s^\theta(x, y) \right)}{\sum_{y'} \exp \left( s^\theta(x, y') \right)} \quad \forall x \in \mathcal{X}, \ y \in \mathcal{Y}
\]

Then we find \( \theta \) that minimizes the cross (conditional) entropy between \( p_{Y|X} \) and \( p_{Y|X}^\theta \):

\[
\theta^* \in \arg \min_\theta \mathbb{E}_{(x,y) \sim p_{XY}} \left[ -\log p_{Y|X}^\theta(y|x) \right]
\]
By universality we must have $p^*_{Y|X} = p_{Y|X}$. This means

$$\frac{\exp(s^* (x, y))}{\sum_{y'} \exp(s^* (x, y'))} = \frac{p_{XY}(x, y)}{\sum_{y'} p_{XY}(x, y')} \quad \forall x \in X, \ y \in Y$$

and it follows that $\exp(s^* (x, y)) = C_x p_{XY}(x, y)$ for some $C_x > 0$. Hence

$$s^* (x, y) = \log p_{XY}(x, y) + \log C_x \quad \forall x \in X, \ y \in Y$$

That is, the optimal score of $(x, y)$ is the log probability of $(x, y)$ shifted by some constant dependent on $x$.

C Gradient Estimation

Without loss of generality we consider the following simplified setting. Fix some target $t \in X$ and define the loss function of $\theta \in \mathbb{R}^{|X|}$ by

$$L(\theta) := -\log \frac{\exp(\theta_t)}{\sum_{x \in X} \exp(\theta_x)} = \log Z(\theta) - \theta_t$$

where $Z(\theta) := \sum_{x \in X} \exp(\theta_x)$. Now, let $q$ be any full-support distribution over $X \setminus \{t\}$. For any $n = (n_1 \ldots n_m) \in (X \setminus \{t\})^m$ we define

$$\tilde{L}_{q,n}(\theta) := -\log \frac{\exp(\theta_t)}{\exp(\theta_t) + \frac{1}{m} \sum_{i=1}^{m} \frac{\exp(\theta_n)}{q(n)}} = \log \tilde{Z}_{q,n}(\theta) - \theta_t$$

where $\tilde{Z}_{q,n}(\theta) := \exp(\theta_t) + \frac{1}{m} \sum_{i=1}^{m} \frac{\exp(\theta_n)}{q(n)}$.

Lemma C.1.

$$E_{n \sim q^m} \left[ \tilde{L}_{q,n}(\theta) \right] = Z(\theta)$$

Proof.

$$E_{n \sim q^m} \left[ \tilde{L}_{q,n}(\theta) \right] = \exp(\theta_t) + E_{n \sim q} \left[ \frac{1}{m} \sum_{i=1}^{m} \frac{\exp(\theta_n)}{q(n)} \right]$$

$$= \exp(\theta_t) + E_{n \sim q} \left[ \frac{\exp(\theta_n)}{q(n)} \right]$$

$$= \exp(\theta_t) + \sum_{n \in X \setminus \{t\}} q(n) \frac{\exp(\theta_n)}{q(n)}$$

$$= \sum_{x \in X} \exp(\theta_x)$$

$$= Z(\theta)$$

\[\square\]
It is convenient to define $\phi_{q,\mathcal{L}}(\theta) \in \mathbb{R}^{m+1}$ where
\[
[\phi_{q,\mathcal{L}}(\theta)]_i = \begin{cases} 
\theta_{n_i} - \log(mq(n_i)) & \text{if } i < m + 1 \\
\theta_i & \text{otherwise}
\end{cases}
\]

We can now write $\hat{L}_{q,\mathcal{L}}(\theta) = -\log p_{\phi_{q,\mathcal{L}}}(\theta)(m + 1)$ where
\[
p_{\phi_{q,\mathcal{L}}}(\theta)(i) := \frac{\exp([\phi_{q,\mathcal{L}}(\theta)]_i)}{\sum_{j=1}^{m+1} \exp([\phi_{q,\mathcal{L}}(\theta)]_j)} \quad \forall i \in \{1 \ldots m + 1\}
\]

Let $p_\theta(x) := \exp(\theta_x)/\sum_{x \in \mathcal{X}} \exp(\theta_x)$ denote the full softmax. The following gradient expressions are easy to verify:

\[
\nabla L(\theta) = \mathbb{E}_{x \sim p_\theta} [\mathbb{I}_x] - \mathbb{I}_t \tag{6}
\]

\[
\nabla \mathbb{E}_{i \sim p_{\phi_{q,\mathcal{L}}}} [\hat{L}_{q,\mathcal{L}}(\theta)] = \mathbb{E}_{i \sim q^n} \left[\nabla [\phi_{q,\mathcal{L}}(\theta)]_i \right] - \mathbb{I}_t \tag{7}
\]

where $\mathbb{I}_x \in \{0,1\}^{|\mathcal{X}|}$ denotes a one-hot vector with 1 at index $x$.

**Lemma C.2.** $\nabla L(\theta) = \nabla \mathbb{E}_{i \sim q^n} [\hat{L}_{q,\mathcal{L}}(\theta)]$ iff $q(x) \propto \exp(\theta_x)$ for all $x \in \mathcal{X}$.

**Proof.** From (6) and (7) it is clear that the statement is equivalent to

\[
p_\theta(l) = \mathbb{E}_{i \sim q^n} \left[\frac{\partial[\phi_{q,\mathcal{L}}(\theta)]_i}{\partial \theta_l} \right] = \mathbb{E}_{i \sim q^n} \left[\sum_{i=1}^{m+1} \exp([\phi_{q,\mathcal{L}}(\theta)]_i) \frac{\partial[\phi_{q,\mathcal{L}}(\theta)]_i}{\partial \theta_l} \right] \tag{8}
\]

for all $l \in \mathcal{X}$, iff $q(x) \propto \exp(\theta_x)$ for all $x \in \mathcal{X}$.

- $l = t$: In this case we have
  \[
  \frac{\partial[\phi_{q,\mathcal{L}}(\theta)]_i}{\partial \theta_l} = \begin{cases} 
  1 & \text{if } i = m + 1 \\
  0 & \text{otherwise}
  \end{cases}
  \]
  Therefore the last term of (8) is
  \[
  \mathbb{E}_{i \sim q^n} \left[\frac{\exp(\theta_1)}{\mathbb{E}_{i \sim q^n} [\hat{L}_{q,\mathcal{L}}(\theta)]} \right] = \frac{\exp(\theta_1)}{\mathbb{E}_{i \sim q^n} [\hat{L}_{q,\mathcal{L}}(\theta)]} = \frac{\exp(\theta_t)}{Z(\theta)} = p_\theta(t)
  \]
  Note that this holds for any choice of $q$.

- $l \neq t$: In this case we have
  \[
  \frac{\partial[\phi_{q,\mathcal{L}}(\theta)]_i}{\partial \theta_l} = \begin{cases} 
  [\lfloor n_i = l \rfloor] & \text{if } i < m + 1 \\
  0 & \text{otherwise}
  \end{cases}
  \]
  Therefore the last term of (8) is
  \[
  \mathbb{E}_{i \sim q^n} \left[\frac{1}{\mathbb{E}_{i \sim q^n} [\hat{L}_{q,\mathcal{L}}(\theta)]} \sum_{i=1}^{m+1} \exp(\theta_{n_i}) [\lfloor n_i = l \rfloor] \right] = \frac{\mathbb{E}_{i \sim q^n} \left[\frac{\exp(\theta_1)}{q(n)} [\lfloor n = l \rfloor] \right]}{\mathbb{E}_{i \sim q^n} [\hat{L}_{q,\mathcal{L}}(\theta)]} = \frac{\exp(\theta_t)}{Z(\theta)} = p_\theta(l)
  \]

8
where the equality with * holds iff \( \hat{L}_{q,n}(\theta) = \exp(\theta_t) + \frac{1}{m} \sum_{i=1}^{m} \exp(\theta_{n_i}) \) is constant for all \( n \in (\mathcal{X} \setminus \{t\})^m \). This implies that \( q(x) \propto \exp(\theta_x) \) for all \( x \in \mathcal{X} \).

\[ \tag*{\Box} \]

Define a distribution \( q_\theta^* \) over \( \mathcal{X} \setminus \{t\} \) by

\[
q_\theta^*(a) = \frac{\exp(\theta_i)}{\sum_{x \in \mathcal{X} \setminus \{t\}} \exp(\theta_x)}
\]

We see that indeed for any \( n \in (\mathcal{X} \setminus \{t\})^m \),

\[
\hat{L}_{q_\theta^*,n}(\theta) = -\log \frac{\exp(\theta_t)}{\exp(\theta_t) + \frac{1}{m} \sum_{i=1}^{m} \exp(\theta_{n_i})} = -\log \frac{\exp(\theta_t)}{\exp(\theta_t) + \sum_{x \in \mathcal{X} \setminus \{t\}} \exp(\theta_x)} = L(\theta)
\]

Getting \( q_\theta^* \) requires computing a normalization term \( \sum_{x \in \mathcal{X} \setminus \{t\}} \exp(\theta_x) \) for each target \( t \in \mathcal{X} \). As a more efficient alternative in practice, we can approximate this distribution by \( p_\theta \) and exclude sampled targets. The bias of the gradient estimator using an approximate \( \hat{q}_\theta \neq q_\theta^* \) is analyzed in \([10]\).

**D Detailed Derivations**

To get (4), note that

\[- H^{\varphi_Y}(A^N|X^Y^N) = E_{\mathcal{X},\mathcal{Y}, a_1, \ldots, a_{N-1}, y_1, \ldots, y_N} \left[ \log p_{A^N|X^N}(a_1 \ldots a_N|X, y_1, \ldots, y_N) \right] \]

\[
= E_{\mathcal{X},\mathcal{Y}, a_1, \ldots, a_{N-1}, y_1, \ldots, y_N} \left[ \sum_{a_N=1} \log p_{A|X,Y}(1|X, y_1, \ldots, y_N) + \sum_{a_N=0} \log (1 - p_{A|X,Y}(1|X, y_1, \ldots, y_N)) \right] \]

\[
= E_{\mathcal{X},\mathcal{Y}, a_1, \ldots, a_{N-1}, y_1, \ldots, y_N} \left[ \sum_{a_N=1} \log p_{A|X,Y}(1|X, y_1, \ldots, y_N) + \sum_{a_N=0} \log (1 - p_{A|X,Y}(1|X, y_1, \ldots, y_N)) \right] \]

Use the tower rule \( E[X] = E[E[X|Y]] \) on each term of the expectation. For the first term,

\[
N E_{\mathcal{X},\mathcal{Y}, a_1, \ldots, a_{N-1}, y_1, \ldots, y_N} \left[ \log p_{A|X,Y}(1|X, y_1, \ldots, y_N) \right]
= N \left( \frac{1}{N} \sum_{x \sim p_X} \log p_{A|X,Y}(1|X, y) \right) \]

\[
= E_{\mathcal{X},\mathcal{Y}} \left[ \log p_{A|X,Y}(1|X, y) \right]
\]

For the second term,

\[
N E_{\mathcal{X},\mathcal{Y}, a_1, \ldots, a_{N-1}, y_1, \ldots, y_N} \left[ \log (1 - p_{A|X,Y}(1|X, y_1, \ldots, y_N)) \right]
= N \left( \frac{N - 1}{N} \sum_{x \sim p_X} \log (1 - p_{A|X,Y}(1|X, y)) \right)
\]

\[
= (N - 1) E_{\mathcal{X},\mathcal{Y}} \left[ \log (1 - p_{A|X,Y}(1|X, y)) \right]
\]
To get Lemma 2.1, first note that \( (p_{Y|X}(\cdot|x) + (N-1)q_Y)/N \) is a proper conditional distribution over \( Y \). The first term of \( -H^q_Y(A^N|XY^N) \) is

\[
E_{(x,y) \sim p_{XY}} \left[ \log p^q_{A|XY}(1|x,y) \right] = E_{(x,y) \sim p_{XY}} \left[ \log \frac{p_{Y|X}(y|x)}{p_{Y|X}(y|x) + (N-1)q_Y(y)} \right] = KL \left( p_{Y|X} \left\| \frac{p_{Y|X}(y|x) + (N-1)q_Y(y)}{N} \right\| \right) - \log N
\]

The second term of \( -H^q_Y(A^N|XY^N) \) is similarly

\[
(N-1) E_{y \sim q_Y} \left[ \log (1 - p^q_{A|XY}(1|x,y)) \right] = (N-1) E_{y \sim q_Y} \left[ \log \frac{(N-1)q_Y(y)}{p_{Y|X}(y|x) + (N-1)q_Y(y)} \right] = (N-1)KL \left( q_Y \left\| \frac{p_{Y|X} + (N-1)q_Y}{N} \right\| - (N-1) \log \frac{N}{N-1} \right)
\]

References


