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 $\mathbf{pop}_{XY}$ from which we can sample correct pairs (x, y) and learning a score function $s^\theta(x, y) \in \mathbb{R}$ parameterized by θ 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 constrastive 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^θ is a distribution over the same variables parameterized by θ, q^θ is equal to p iff it is the minimizer of the cross entropy between p and q^θ

$$\theta^* \in \arg\min_\theta \mathop{\mathbf{E}}_{z \sim p}\left[-\log q^\theta(z)\right] \qquad \Longleftrightarrow \qquad q^{\theta^*}(z) = p(z) \qquad \forall z$$

assuming the **universality** of q^θ: that is, it is expressive enough to model p so that $p = q^\theta$ for some θ. 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

$$\mathbf{pop}^{q_Y}_{IXY^N}(i, x, y_1 \ldots y_N) := \frac{1}{N}\mathbf{pop}_{XY}(x, y_i)\prod_{j \neq i} q_Y(y_j)$$

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

$$\mathbf{pop}^{q_Y}_{I|XY^N}(i|x, y_1 \ldots y_N) = \frac{\mathbf{pop}_{Y|X}(y_i|x)\prod_{j \neq i} q_Y(y_j)}{\sum_{k=1}^N \mathbf{pop}_{Y|X}(y_k|x)\prod_{j \neq k} q_Y(y_j)} = \frac{\frac{\mathbf{pop}_{Y|X}(y_i|x)}{q_Y(y_i)}}{\sum_{k=1}^N \frac{\mathbf{pop}_{Y|X}(y_k|x)}{q_Y(y_k)}} \quad (1)$$

Let $H^{q_Y}(I|XY^N)$ denote the conditional entropy of $\mathbf{pop}^{q_Y}_{I|XY^N}$. The following observation is made in [8].

Lemma 1.1. Let $q_Y = \mathbf{pop}_Y$. Then $H^{\mathbf{pop}_Y}(I|XY^N) \geq \log N - I(X,Y)$ where $I(X,Y)$ is the mutual information between $(x,y) \sim \mathbf{pop}_{XY}$.

Proof. By (1),

$$\mathop{\mathbf{E}}_{(i,x,y_1\dots y_N)\sim\mathbf{pop}^{\mathbf{pop}_Y}_{IXY^N}}\left[-\log \mathbf{pop}^{\mathbf{pop}_Y}_{I|XY^N}(i|x,y_1\dots y_N)\right]$$

$$= -\underbrace{\mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}}\left[\frac{\mathbf{pop}_{Y|X}(y|x)}{\mathbf{pop}_Y(y)}\right]}_{I(X,Y)} + \underbrace{\mathop{\mathbf{E}}_{(i,x,y_1\dots y_N)\sim\mathbf{pop}^{\mathbf{pop}_Y}_{IXY^N}}\left[\log\sum_{k=1}^{N}\frac{\mathbf{pop}_{Y|X}(y_k|x)}{\mathbf{pop}_Y(y_k)}\right]}_{\geq \log N}$$

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

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

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 $H^{\mathbf{pop}_Y}(I|XY^N) \geq 0$ (Shannon entropy is nonnegative). ☐

1.2 Estimation

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

$$p^\theta_{I|XY^N}(i|x,y_1\dots y_N) := \frac{\exp\left(s^\theta(x,y_i)\right)}{\sum_{j=1}^N \exp\left(s^\theta(x,y_j)\right)} \qquad \forall i \in \{1\dots N\}$$

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

$$\bar{H}^{q_Y}_\theta(I|XY^N) := \mathop{\mathbf{E}}_{(i,x,y_1\dots y_N)\sim\mathbf{pop}^{q_Y}_{IXY^N}}\left[-\log p^\theta_{I|XY^N}(i|x,y_1\dots y_N)\right]$$

Note that $\bar{H}^{q_Y}_\theta(I|XY^N) \geq H^{q_Y}(I|XY^N)$ for all θ by the usual property of cross entropy. If $q_Y = \mathbf{pop}_Y$, Corollary 1.2 implies that

$$\begin{aligned}B(\theta) &:= \log N - \bar{H}^{\mathbf{pop}_Y}_\theta(I|XY^N) = \mathop{\mathbf{E}}_{(i,x,y_1\dots y_N)\sim\mathbf{pop}^{\mathbf{pop}_Y}_{IXY^N}}\left[\log\frac{\exp\left(s^\theta(x,y_i)\right)}{\frac{1}{N}\sum_{j=1}^N \exp\left(s^\theta(x,y_j)\right)}\right]\\ &\leq \log N - H^{\mathbf{pop}_Y}(I|XY^N)\\ &\leq \min\{I(X,Y), \log N\}\end{aligned}$$

Thus minimizing $\bar{H}^{\mathbf{pop}_Y}_\theta(I|XY^N)$ over θ 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 $\theta^{q_Y} \in \arg\min_\theta \bar{H}_\theta^{q_Y}(I|XY^N)$. By universality we must have $p_{I|XY^N}^{\theta^{q_Y}} = \mathbf{pop}_{I|XY^N}^{q_Y}$. By (1) this means

$$s^{\theta^{q_Y}}(x,y) = \log \frac{\mathbf{pop}_{Y|X}(y|x)}{q_Y(y)} + \log C_x \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

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

$$\mathbf{pop}_{Y|X}(y|x) = \frac{\exp\left(s^{\theta^{q_Y}}(x,y) + \log q_Y(y)\right)}{\sum_{y'} \exp\left(s^{\theta^{q_Y}}(x,y') + \log q_Y(y')\right)} \tag{2}$$

This is consistent with the "ranking" algorithm in [5]. Note that the additive adjustment is unnecessary if we choose uniformly random q_Y. 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 $\mathbf{pop}_A(1) = 1/N$ and $\mathbf{pop}_A(0) = (N-1)/N$. Given $x \sim \mathbf{pop}_X$ and $a \sim \mathbf{pop}_A$, it defines

$$\mathbf{pop}_{Y|XA}^{q_Y}(y|x,a) := \begin{cases} \mathbf{pop}_{Y|X}(y|x) & \text{if } a = 1 \\ q_Y(y) & \text{if } a = 0 \end{cases}$$

This yields the conditional head probability

$$\mathbf{pop}_{A|XY}^{q_Y}(1|x,y) = \frac{\mathbf{pop}_{Y|X}(y|x)}{\mathbf{pop}_{Y|X}(y|x) + (N-1)q_Y(y)} \tag{3}$$

Given $x \sim \mathbf{pop}_X$ and N iid samples $a_i \sim \mathbf{pop}_A$ and $y_i \sim \mathbf{pop}_{Y|XA}^{q_Y}(\cdot|x,a_i)$ for $i = 1 \ldots N$, the joint conditional probability of the coin flips is given by

$$\mathbf{pop}_{A^N|XY^N}^{q_Y}(a_1 \ldots a_N|x, y_1 \ldots y_N) = \prod_{i=1:a_i=1}^{N} \mathbf{pop}_{A|XY}^{q_Y}(1|x,y_i) \prod_{j=1:a_j=0}^{N} (1 - \mathbf{pop}_{A|XY}^{q_Y}(1|x,y_j))$$

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

$$\begin{aligned} H^{q_Y}(A^N|XY^N) = &\mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}} \left[-\log \mathbf{pop}_{A|XY}^{q_Y}(1|x,y)\right] \\ &+ (N-1) \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ y\sim q_Y}} \left[-\log(1 - \mathbf{pop}_{A|XY}^{q_Y}(1|x,y))\right] \end{aligned} \tag{4}$$

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

Lemma 2.1. Let $\mathrm{KL}(p||q)$ denote the KL divergence between distributions p and q. Then

$$-H^{q_Y}(A^N|XY^N) = \mathrm{KL}\left(\mathbf{pop}_{Y|X}\middle|\middle|\frac{\mathbf{pop}_{Y|X} + (N-1)q_Y}{N}\right) + (N-1)\mathrm{KL}\left(q_Y\middle|\middle|\frac{\mathbf{pop}_{Y|X} + (N-1)q_Y}{N}\right) - \log N - (N-1)\log\left(\frac{N}{N-1}\right)$$

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

$$-H^{q_Y}(A^2|XY^2) = 2\mathrm{JSD}\left(\mathbf{pop}_{Y|X}\middle|\middle|q_Y\right) - \log 4$$

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:

$$\mathbf{GAN}(D, q_Y) := \mathop{\mathbf{E}}_{y\sim\mathbf{pop}_Y}[\log D(1|y)] + \mathop{\mathbf{E}}_{y\sim q_Y}[\log(1 - D(1|y))]$$
$$J_{\mathrm{GAN}} := \min_{q_Y} \max_{D} \mathbf{GAN}(D, q_Y)$$

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

$$J_{\mathrm{GAN}} = \min_{q_Y} \mathbf{GAN}(\mathbf{pop}^{q_Y}_{A|Y}, q_Y) = \min_{q_Y} 2\mathrm{JSD}\left(\mathbf{pop}_Y\middle|\middle|q_Y\right) - \log 4 = -\log 4$$

where the minimizer is $q_Y = \mathbf{pop}_Y$. At this equilibrium, we see that the best discriminator is uniform $\mathbf{pop}^{\mathbf{pop}_Y}_{A|Y}(1|y) = 1/2$ and the generator "wins".

2.2 Estimation

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

$$p^\theta_{A|XY}(1|x, y) := \frac{1}{1 + \exp(-s^\theta(x, y))}$$

This is used to define the joint conditional distribution

$$p^\theta_{A^N|XY^N}(a_1 \ldots a_N|x, y_1 \ldots y_N) = \prod_{i=1:a_i=1}^{N} p^\theta_{A|XY}(1|x, y_i) \prod_{j=1:a_j=0}^{N} (1 - p^\theta_{A|XY}(1|x, y_j))$$

The model is again estimated by minimizing the cross (conditional) entropy between $\mathbf{pop}^{q_Y}_{A^N|XY^N}$ and $p^\theta_{A^N|XY^N}$. Similar to (4) this objective can be written in the friendlier form

$$\theta^{q_Y} \in \arg\max_\theta \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}}\left[\log p^\theta_{A|XY}(1|x, y)\right] + (N-1)\mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X\\ y\sim q_Y}}\left[\log(1 - p^\theta_{A|XY}(1|x, y))\right]$$

By universality we must have $p^{\theta^{q_Y}}_{A|XY} = \mathbf{pop}^{q_Y}_{A|XY}$. By (3) this means

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

If $q_Y = \mathbf{pop}_Y$, 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 θ^{q_Y} to recover the underlying conditional distribution

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

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 θ such that

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

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

A Hinge Loss

We want to find θ 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**

$$\operatorname*{arg\,min}_{\theta} \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}} \left[\overbrace{\mathbb{1}\left(\underbrace{s^\theta(x,y) - \max_{y'\neq y} s^\theta(x,y')}_{\text{margin of } (x,y)} \leq 0 \right)}^{\text{zero-one loss on } (x,y)} \right]$$

where $\mathbb{1}(\cdot) \in \{0,1\}$ 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**

$$\operatorname*{arg\,min}_{\theta} \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}} \left[\overbrace{\max\left\{ 0, 1 - \underbrace{\left(s^\theta(x,y) - \max_{y'\neq y} s^\theta(x,y') \right)}_{\text{margin of } (x,y)} \right\}}^{\text{hinge loss on } (x,y)} \right]$$

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 $\mathbf{pop}_{Y|X}$. To this end, we turn the score function into a proper conditional distribution by using the softmax operation:

$$p^\theta_{Y|X}(y|x) := \frac{\exp\left(s^\theta(x,y)\right)}{\sum_{y'} \exp\left(s^\theta(x,y')\right)} \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

Then we find θ that minimizes the cross (conditional) entropy between $\mathbf{pop}_{Y|X}$ and $p^\theta_{Y|X}$:

$$\theta^* \in \operatorname*{arg\,min}_{\theta} \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}} \left[-\log p^\theta_{Y|X}(y|x) \right] \tag{5}$$

By universality we must have $p^{\theta^*}_{Y|X} = \mathbf{pop}_{Y|X}$. This means

$$\frac{\exp\left(s^{\theta^*}(x,y)\right)}{\sum_{y'}\exp\left(s^{\theta^*}(x,y')\right)} = \frac{\mathbf{pop}_{XY}(x,y)}{\sum_{y'}\mathbf{pop}_{XY}(x,y')} \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

and it follows that $\exp\left(s^{\theta^*}(x,y)\right) = C_x \mathbf{pop}_{XY}(x,y)$ for some $C_x > 0$. Hence

$$s^{\theta^*}(x,y) = \log \mathbf{pop}_{XY}(x,y) + \log C_x \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

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 $t \in \mathcal{X}$ and define the loss function of $\theta \in \mathbb{R}^{|\mathcal{X}|}$ by

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

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

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

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

Lemma C.1.

$$\mathop{\mathbf{E}}_{\underline{n}\sim q^m}\left[\widehat{Z}_{q,\underline{n}}(\theta)\right] = Z(\theta)$$

Proof.

$$\begin{aligned}
\mathop{\mathbf{E}}_{\underline{n}\sim q^m}\left[\widehat{Z}_{q,\underline{n}}(\theta)\right] &= \exp(\theta_t) + \mathop{\mathbf{E}}_{\underline{n}\sim q^m}\left[\frac{1}{m}\sum_{i=1}^m \frac{\exp(\theta_{n_i})}{q(n_i)}\right] \\
&= \exp(\theta_t) + \mathop{\mathbf{E}}_{n\sim q}\left[\frac{\exp(\theta_n)}{q(n)}\right] \\
&= \exp(\theta_t) + \sum_{n\in\mathcal{X}\backslash\{t\}} q(n)\frac{\exp(\theta_n)}{q(n)} \\
&= \sum_{x\in\mathcal{X}}\exp(\theta_x) \\
&= Z(\theta)
\end{aligned}$$

□

It is convenient to define $\phi_{q,\underline{n}}(\theta) \in \mathbb{R}^{m+1}$ where

$$[\phi_{q,\underline{n}}(\theta)]_i = \begin{cases} \theta_{n_i} - \log(mq(n_i)) & \text{if } i < m+1 \\ \theta_t & \text{otherwise} \end{cases}$$

We can now write $\widehat{L}_{q,\underline{n}}(\theta) = -\log p_{\phi_{q,\underline{n}}(\theta)}(m+1)$ where

$$p_{\phi_{q,\underline{n}}(\theta)}(i) := \frac{\exp([\phi_{q,\underline{n}}(\theta)]_i)}{\sum_{j=1}^{m+1} \exp([\phi_{q,\underline{n}}(\theta)]_j)} \qquad \forall i \in \{1 \dots m+1\}$$

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:

$$\nabla L(\theta) = \mathop{\mathbf{E}}_{x \sim p_\theta} [\mathbb{1}_x] - \mathbb{1}_t \tag{6}$$

$$\nabla \mathop{\mathbf{E}}_{\underline{n} \sim q^m} \left[\widehat{L}_{q,\underline{n}}(\theta)\right] = \mathop{\mathbf{E}}_{\substack{\underline{n} \sim q^n \\ i \sim p_{\phi_{q,\underline{n}}(\theta)}}} \left[\nabla [\phi_{q,\underline{n}}(\theta)]_i\right] - \mathbb{1}_t \tag{7}$$

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

Lemma C.2. $\nabla L(\theta) = \nabla \mathop{\mathbf{E}}_{\underline{n} \sim q^m} \left[\widehat{L}_{q,\underline{n}}(\theta)\right]$ 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

$$p_\theta(l) = \mathop{\mathbf{E}}_{\substack{\underline{n} \sim q^n \\ i \sim p_{\phi_{q,\underline{n}}(\theta)}}} \left[\frac{\partial [\phi_{q,\underline{n}}(\theta)]_i}{\partial \theta_l}\right] = \mathop{\mathbf{E}}_{\underline{n} \sim q^n} \left[\sum_{i=1}^{m+1} \frac{\exp([\phi_{q,\underline{n}}(\theta)]_i)}{\widehat{Z}_{q,\underline{n}}(\theta)} \frac{\partial [\phi_{q,\underline{n}}(\theta)]_i}{\partial \theta_l}\right] \tag{8}$$

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

  $$\frac{\partial [\phi_{q,\underline{n}}(\theta)]_i}{\partial \theta_t} = \begin{cases} 1 & \text{if } i = m+1 \\ 0 & \text{otherwise} \end{cases}$$

  Therefore the last term of (8) is

  $$\mathop{\mathbf{E}}_{\underline{n} \sim q^n} \left[\frac{\exp(\theta_t)}{\widehat{Z}_{q,\underline{n}}(\theta)}\right] = \frac{\exp(\theta_t)}{\mathop{\mathbf{E}}_{\underline{n} \sim q^n} \left[\widehat{Z}_{q,\underline{n}}(\theta)\right]} = \frac{\exp(\theta_t)}{Z(\theta)} = p_\theta(t)$$

  Note that this holds for any choice of q.

- $l \neq t$: In this case we have

  $$\frac{\partial [\phi_{q,\underline{n}}(\theta)]_i}{\partial \theta_l} = \begin{cases} [[n_i = l]] & \text{if } i < m+1 \\ 0 & \text{otherwise} \end{cases}$$

  Therefore the last term of (8) is

  $$\mathop{\mathbf{E}}_{\underline{n} \sim q^n} \left[\frac{1}{\widehat{Z}_{q,\underline{n}}(\theta)} \sum_{i=1}^{m} \frac{\exp(\theta_{n_i})}{mq(n_i)} [[n_i = l]]\right] \overset{*}{=} \frac{\mathop{\mathbf{E}}_{n \sim q} \left[\frac{\exp(\theta_n)}{q(n)} [[n = l]]\right]}{\mathop{\mathbf{E}}_{\underline{n} \sim q^n} \left[\widehat{Z}_{q,\underline{n}}(\theta)\right]} = \frac{\exp(\theta_l)}{Z(\theta)} = p_\theta(l)$$

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

□

Define a distribution q^*_θ over $\mathcal{X}\backslash\{t\}$ by

$$q^*_\theta(n) = \frac{\exp(\theta_n)}{\sum_{x\in\mathcal{X}\backslash\{t\}} \exp(\theta_x)}$$

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

$$\widehat{L}_{q^*_\theta,\underline{n}}(\theta) = -\log \frac{\exp(\theta_t)}{\exp(\theta_t) + \frac{1}{m}\sum_{i=1}^{m} \frac{\exp(\theta_{n_i})}{q^*_\theta(n_i)}} = -\log \frac{\exp(\theta_t)}{\exp(\theta_t) + \sum_{x\in\mathcal{X}\backslash\{t\}} \exp(\theta_x)} = L(\theta)$$

Getting q^*_θ requires computing a normalization term $\sum_{x\in\mathcal{X}\backslash\{t\}} \exp(\theta_x)$ for each target $t \in \mathcal{X}$. As a more efficient alternative in practice, we can approximate this distribution by p_θ 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

$$\begin{aligned}
&- H^{q_Y}(A^N|XY^N) \\
&= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ a_i\sim\mathbf{pop}_A,\, y_i\sim\mathbf{pop}^{q_Y}_{Y|XA}(\cdot|x,a_i)}} \left[\log \mathbf{pop}^{q_Y}_{A^N|XY^N}(a_1 \ldots a_N|x, y_1 \ldots y_N)\right] \\
&= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ a_i\sim\mathbf{pop}_A,\, y_i\sim\mathbf{pop}^{q_Y}_{Y|XA}(\cdot|x,a_i)}} \left[\sum_{i=1}^{N} [[a_i = 1]] \log \mathbf{pop}^{q_Y}_{A|XY}(1|x, y_i) + [[a_i = 0]] \log(1 - \mathbf{pop}^{q_Y}_{A|XY}(1|x, y_i))\right] \\
&= N \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ a\sim\mathbf{pop}_A,\, y\sim\mathbf{pop}^{q_Y}_{Y|XA}(\cdot|x,a)}} \left[[[a = 1]] \log \mathbf{pop}^{q_Y}_{A|XY}(1|x, y) + [[a = 0]] \log(1 - \mathbf{pop}^{q_Y}_{A|XY}(1|x, y))\right]
\end{aligned}$$

Use the tower rule $\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]]$ on each term of the expectation. For the first term,

$$\begin{aligned}
N \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ a\sim\mathbf{pop}_A,\, y\sim\mathbf{pop}^{q_Y}_{Y|XA}(\cdot|x,a)}} \left[[[a = 1]] \log \mathbf{pop}^{q_Y}_{A|XY}(1|x, y)\right] &= N \left(\frac{1}{N} \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ y\sim\mathbf{pop}_{Y|X}(\cdot|x)}} \left[\log \mathbf{pop}^{q_Y}_{A|XY}(1|x, y)\right]\right) \\
&= \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}} \left[\log \mathbf{pop}^{q_Y}_{A|XY}(1|x, y)\right]
\end{aligned}$$

For the second term,

$$\begin{aligned}
N \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ a\sim\mathbf{pop}_A,\, y\sim\mathbf{pop}^{q_Y}_{Y|XA}(\cdot|x,a)}} \left[[[a = 0]] \log(1 - \mathbf{pop}^{q_Y}_{A|XY}(1|x, y))\right] &= N \left(\frac{N-1}{N} \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ y\sim q_Y}} \left[\log(1 - \mathbf{pop}^{q_Y}_{A|XY}(1|x, y))\right]\right) \\
&= (N-1) \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ y\sim q_Y}} \left[\log(1 - \mathbf{pop}^{q_Y}_{A|XY}(1|x, y))\right]
\end{aligned}$$

To get Lemma 2.1, first note that $(\mathbf{pop}_{Y|X}(\cdot|x) + (N-1)q_Y)/N$ is a proper conditional distribution over $\mathcal{Y}$. The first term of $-H^{q_Y}(A^N|XY^N)$ is

$$\begin{aligned}
\mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}}\left[\log \mathbf{pop}^{q_Y}_{A|XY}(1|x,y)\right] &= \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}}\left[\log \frac{\mathbf{pop}_{Y|X}(y|x)}{\mathbf{pop}_{Y|X}(y|x) + (N-1)q_Y(y)}\right] \\
&= \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}_{XY}}\left[\log \frac{\frac{\mathbf{pop}_{Y|X}(y|x)}{N}}{\frac{\mathbf{pop}_{Y|X}(y|x)+(N-1)q_Y(y)}{N}}\right] \\
&= \mathrm{KL}\left(\mathbf{pop}_{Y|X}\middle|\middle| \frac{\mathbf{pop}_{Y|X} + (N-1)q_Y}{N}\right) - \log N
\end{aligned}$$

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

$$\begin{aligned}
(N-1)\mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ y\sim q_Y}}\left[\log(1-\mathbf{pop}^{q_Y}_{A|XY}(1|x,y))\right] &= (N-1)\mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ y\sim q_Y}}\left[\log \frac{(N-1)q_Y(y)}{\mathbf{pop}_{Y|X}(y|x) + (N-1)q_Y(y)}\right] \\
&= (N-1)\mathrm{KL}\left(q_Y\middle|\middle| \frac{\mathbf{pop}_{Y|X} + (N-1)q_Y}{N}\right) - (N-1)\log\frac{N}{N-1}
\end{aligned}$$

References

[1] Bengio, Y. and Senécal, J.-S. (2008). Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, **19**(4), 713–722.

[2] Blanc, G. and Rendle, S. (2018). Adaptive sampled softmax with kernel based sampling. In *International Conference on Machine Learning*, pages 590–599.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[4] Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

[5] Ma, Z. and Collins, M. (2018). Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. *arXiv preprint arXiv:1809.01812*.

[6] McAllester, D. and Stratos, K. (2020). Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR.

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26.

[8] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[9] Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. (2019). On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180.

[10] Rawat, A. S., Chen, J., Yu, F. X. X., Suresh, A. T., and Kumar, S. (2019). Sampled softmax with random fourier features. In *Advances in Neural Information Processing Systems*, pages 13857–13867.

[11] Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.