Formal Limitations on the Measurement of Mutual Information: Supplementary Material

David McAllester
Toyota Technological Institute at Chicago

Karl Stratos
Rutgers University

1 PROOF OF THEOREM 2.1

For any distribution r_X over X, we can write

$$\begin{aligned} D_{\mathrm{KL}}(p_X||q_X) &= \mathop{\mathbb{E}}_{x\sim p_X}\left[\ln\frac{r_X(x)}{q_X(x)}\right] + D_{\mathrm{KL}}(p_X||r_X) \\ &\geq \mathop{\mathbb{E}}_{x\sim p_X}\left[\ln\frac{r_X(x)}{q_X(x)}\right] \end{aligned} \tag{1}$$

Let $f : \mathcal{X} \to \mathbb{R}$ be a bounded function and define

$$r_X(x) = \frac{q_X(x)e^{f(x)}}{\mathop{\mathbb{E}}_{x\sim q_X}\left[e^{f(x)}\right]} \qquad \forall x \in \mathcal{X}$$

which is a valid distribution over X. Plugging this into the lower bound in (1), we have

$$\mathop{\mathbb{E}}_{x\sim p_X}\left[\ln\frac{r_X(x)}{q_X(x)}\right] = \mathop{\mathbb{E}}_{x\sim p_X}\left[f(x)\right] - \ln \mathop{\mathbb{E}}_{x\sim q_X}\left[e^{f(x)}\right] \tag{2}$$

By (1), the supremum of (2) over the choice of f is precisely the KL divergence between p_X and q_X. It can be easily verified that an optimal f is given by

$$f(x) = \ln\frac{p_X(x)}{q_X(x)} \qquad \forall x \in \mathcal{X}$$

Since (2) is invariant to translation of f, without loss of generality we can assume that the range of f is bounded in $[0, F_{\max}]$ for some constant $F_{\max}$.

2 MUTUAL INFORMATION AS THE SUPREMUM OVER BINNINGS

We now show that the mutual information $I(X, Y; p_{XY})$ for X and Y continuous can be expressed as the supremum of $I(C(X), C'(Y); p_{XY})$ over discrete binnings of the continuous space. We first consider the case where $X, Y \in \mathbb{R}$ and where the mutual information can be written as a Riemann integral over densities.

$$\begin{aligned} &I(X, Y; p_{XY}) \\ &= \int p_{XY}(x, y) \ln\frac{p_{XY}(x, y)}{p_X(x)p_Y(y)}\, dx\, dy \\ &= \lim_{\epsilon\to 0} \sum_{i,j\in\mathbb{Z}} p_{XY}(i\epsilon, j\epsilon)\ \ln\frac{p_{XY}(i\epsilon, j\epsilon)}{p_X(i\epsilon)p_Y(j\epsilon)}\epsilon^2 \end{aligned}$$

where $\mathbb{Z}$ is the set of all integers. For each $i \in \mathbb{Z}$, define the half-open interval $C_{i,\epsilon} := [i\epsilon, (i + 1)\epsilon)$. The probability of the interval is approximately $\epsilon p_X(i\epsilon)$ under p_X (similarly for p_Y and p_{XY}). Therefore we can write the last expression as

$$\begin{aligned} &\lim_{\epsilon\to 0} \sum_{i,j\in\mathbb{Z}} p_{XY}(C_{i,\epsilon} \times C_{j,\epsilon})\ \ln\frac{p_{XY}(C_{i,\epsilon} \times C_{j,\epsilon})}{p_X(C_{i,\epsilon})p_Y(C_{j,\epsilon})} \\ &= \lim_{\epsilon\to 0} \sum_{i,j\in\mathbb{Z}} p_{I_\epsilon J_\epsilon}(i, j)\ \ln\frac{p_{I_\epsilon J_\epsilon}(i, j)}{p_{I_\epsilon}(i)p_{J_\epsilon}(j)} \\ &= \lim_{\epsilon\to 0} I(I_\epsilon, J_\epsilon; p_{I_\epsilon J_\epsilon}) \end{aligned}$$

where (I_ϵ, J_ϵ) denote the indices (i, j) such that $x \in C_{i,\epsilon}$ and $y \in C_{j,\epsilon}$ for $(x, y) \sim p_{XY}$.

This proof immediately generalizes to higher dimensions where the mutual information can be expressed as a Riemann integral. We believe that this statement remains true for arbitrary measures on product spaces where the mutual information is finite. However the proof for this extremely general case appears to be nontrivial.

3 PAC-BAYESIAN BOUNDS

The PAC-Bayesian bounds apply to "broad basin" losses and loss estimates such as the following:

$$\begin{aligned} H_\sigma(S, q_X^\theta) &= \mathop{\mathbb{E}}_{x\sim p_X}\left[\mathop{\mathbb{E}}_{\epsilon\sim N(0,\sigma I)}\left[-\ln q_X^{\theta+\epsilon}(x)\right]\right] \\ \widehat{H}_\sigma(S, q_X^\theta) &= \frac{1}{|S|}\sum_{x\in S} \mathop{\mathbb{E}}_{\epsilon\sim N(0,\sigma I)}\left[-\ln q_X^{\theta+\epsilon}(x)\right] \end{aligned}$$

Under mild smoothness conditions on $q_X^\theta(x)$ as a function of θ we have

$$\lim_{\sigma\to 0} H_\sigma(p_X, q_X^\theta) = H(p_X, q_X^\theta)$$
$$\lim_{\sigma\to 0} \widehat{H}_\sigma(S, q_X^\theta) = \widehat{H}(S, q_X^\theta)$$

An L_2 PAC-Bayesian generalization bound (McAllester, 2013) gives that for any parameterized class of models and any bounded notion of loss, and any $\lambda > 1/2$ and $\sigma > 0$, with probability at least $1-\delta$ over the draw of S from p_X^N we have the following simultaneously for all parameter vectors θ.

$$H_\sigma(p_X, q_X^\theta) \leq \frac{1}{1-\frac{1}{2\lambda}}\left(\widehat{H}_\sigma(S, q_X^\theta) + \frac{\lambda F_{\max}}{N}\left(\frac{||\theta||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right)$$

It is instructive to set $\lambda = 5$ in which case the bound becomes.

$$H_\sigma(p_X, q_X^\theta) \leq \frac{10}{9}\left(\widehat{H}_\sigma(S, q_X^\theta) + \frac{5F_{\max}}{N}\left(\frac{||\theta||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right)$$

While this bound is linear in $1/N$, and tighter in practice than square root bounds, note that there is a small residual gap when holding λ fixed at 5 while taking $N \to \infty$. In practice the regularization parameter λ can be tuned on holdout data. One point worth noting is the form of the dependence of the regularization coefficient on $F_{\max}$, N and the basin parameter σ.

It is also worth noting that the bound can be given in terms of "distance traveled" in parameter space from an initial (random) parameter setting θ_0.

$$H_\sigma(p_X, q_X^\theta) \leq \frac{10}{9}\left(\widehat{H}_\sigma(S, q_X^\theta) + \frac{5F_{\max}}{N}\left(\frac{||\theta-\theta_0||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right)$$

Evidence is presented in Dziugaite and Roy (2017) that the distance traveled bounds are tighter in practice than traditional L_2 generalization bounds.

4 EXPERIMENT DETAILS

Article pairs. We take pairs from the Who-Did-What dataset (Onishi *et al.*, 2016). The pairs in this dataset were constructed by drawing articles from the LDC Gigaword newswire corpus. A first article is drawn at random and then a list of candidate second articles is drawn using the first sentence of the first article as an information retrieval query. A second article is selected from the candidates using criteria described in Onishi *et al.* (2016), the most significant of which is that the second article must have occurred within a two week time interval of the first. The training statistics of this dataset after preprocessing is given in Table 1.

	train (tgt)	train (src)
# articles	68348	68348
vocab size	100001	87941
# words	20271664	19072167
avg length	296	279
max length	400	400
min length	10	12

Table 1: Training statistics of the article pairs

	train (tgt)	train (src)
# sentences	160239	160239
vocab size	24726	35445
# words	3275729	3100720
avg length	20	19
max length	175	172
min length	2	2

Table 2: Training statistics of the translation pairs

Translation pairs. Our translation pairs consists of English-German sentence pairs extracted from the IWSLT 2014 dataset. The training statistics of this dataset after preprocessing is given in Table 2.

Model. We train an LSTM encoder-decoder model where the decoder doubles as both the decoder of a translation model and a language model. The decoder is a left-to-right 2-layer LSTM in which a single word embedding matrix is used for both input embeddings and the softmax predictions. When this model is trained as a language model on PTB using standard hyperparameter values it achieves test perplexity of 72.26. The encoder is a separate left-to-right 2-layer LSTM using the same word embeddings as the decoder. We use the input-feeding attention archietecture of Luong *et al.* (2015).

The model is trained using SGD and batch size 10 with no BPTT-style truncation. The dimension of the input/hidden states is 900 (thus 1800 for the input-feeding decoder). We use step-wise dropout with rate 0.65 on word embeddings and hidden states. The model is trained for 40 epochs and the model that achieves the best validation perplexity is selected. The sequence-level cross entropy is estimated as $\mathrm{SQXENT} = \frac{1}{M}\mathrm{NLL}$ where NLL is the negative log likelihood of the corpus and M is the total number of sequences

in the corpus.

Mutual information is estimated by taking the difference in SQXENT between the language model and the translation model (17).For article pairs, we obtain

$$\widehat{I}(X,Y;p_{XY}) = 1131.74 - 1048.33 = 83.41$$

in nats which translates to 120.34 bits. For translation pairs, we obtain

$$\widehat{I}(X,Y;p_{XY}) = 81.73 - 43.80 = 37.9$$

in nats which translates to 54.72 bits.

References

Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

McAllester, D. (2013). A pac-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307:2118*.

Onishi, T., Wang, H., Bansal, M., Gimpel, K., and McAllester, D. (2016). Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235.