Supplementary Material - Boosted Negative Sampling by Quadratically Constrained Entropy Maximization

Taygun Kekeç\textsuperscript{a}, David Mimno\textsuperscript{b}, David M. J. Tax\textsuperscript{a}

\textsuperscript{a}Pattern Recognition and Bioinformatics Laboratory
Delft University of Technology
Mekelweg 4 2628CD, Delft, The Netherlands
\textsuperscript{b}Information Sciences Department, Cornell University
Ithaca, NY 14853, New York

ABSTRACT

This is the supplementary material for the paper titled “Boosted Negative Sampling by Quadratically Constrained Entropy Maximization”. © 2019 Elsevier Ltd. All rights reserved.

Appendix A. Negative Sampling Objective

The negative sampling objective is given as follows:

\[ J(\theta) = E_{p_\theta} \left[ \ln p_\theta(x; \theta) \right] + E_{p_\theta} \left[ \ln(1 - \sigma(y; \theta)) \right] \]  \hspace{1cm} (A.1)

Here, \( \sigma \) is the sigmoid function:

\[ \sigma(u; \theta) = \frac{1}{1 + \exp[-G(u; \theta)]} \]

where \( G \) is the difference between the log likelihood of the sample under the model and the negative sampling distribution:

\[ G(u; \theta) = \ln p_\theta(u) - \ln p_\theta(u) \]

Substitution of \( \sigma \) and \( G \) functions gives us the following:

\[ J_T(\theta) = E_{p_\theta} \left[ \ln \frac{p_\theta^\theta(x)}{p_\theta^\theta(x) + p_\theta(x)} \right] + E_{p_\theta} \left[ \ln \frac{p_\theta(y)}{p_\theta(y) + p_\theta(y)} \right] \]

Using logarithmic properties and expectation additivity, we decompose this objective into:

\[ J(\theta, p_n) = E_{p_\theta} \left[ \ln p_\theta^\theta(x) \right] - E_{p_\theta} \left[ \ln(p_\theta^\theta(x) + p_\theta(x)) \right] \]

\[ -E_{p_n} \left[ \ln(p_\theta^\theta(y)) \right] + p_n(y) + E_{p_n} \left[ \ln p_n(y) \right] \]

where fourth term is constant in \( \theta \). \( \square \)

Appendix B. Smoothing the distribution

Assume we have a probability mass function, with ordered entries:

\[ p_1 \geq p_2 \geq ... \geq p_n > 0 \]  \hspace{1cm} (B.1)

with \( \sum_{i=1}^{n} p_i = 1 \). We smooth PMF \( p \) slightly, by modifying two neighbouring probabilities with a small probability \( \Delta_i \). This defines a new PMF \( \tilde{p} \), with \( \tilde{p}_i = p_i - \Delta_i, \tilde{p}_{i+1} = p_{i+1} + \Delta_i \), and all other probabilities remain the same. The entropy change: \( H(\tilde{p}) - H(p) \) can be stated as:

\[ = -(p_i - \Delta_i) \log(p_i - \Delta_i) - (p_{i+1} + \Delta_i) \log(p_{i+1} + \Delta_i) \]
\[ + p_i \log p_i + p_{i+1} \log p_{i+1} \]
\[ = -p_i \log(p_i - \Delta_i) - p_{i+1} \log(p_{i+1} + \Delta_i) \]
\[ + \Delta_i \log(p_i - \Delta_i) - \Delta_i \log(p_{i+1} + \Delta_i) \]
\[ = -p_i \log(1 - \frac{\Delta_i}{p_i}) - p_{i+1} \log(1 + \frac{\Delta_i}{p_{i+1}}) \]
\[ + \Delta_i \log(p_i (1 - \frac{\Delta_i}{p_i})) - \Delta_i \log(p_{i+1} (1 + \frac{\Delta_i}{p_{i+1}})) \]

The logarithms are of the form \( \log(1 + x) \) for which the Taylor expansion around \( x = 0 \) can be used:

\[ \log(1 + x) = 0 + x + O(x^2) \]  \hspace{1cm} (B.2)
Therefore, the substitution gives:

\[
H(\tilde{p}) - H(p) = \frac{\Delta_i}{p_i} - \frac{\Delta_{i+1}}{p_{i+1}} = \Delta_i \log p_i - \Delta_{i+1} \log p_{i+1} + O(\Delta_i)
\]

\[
= +\Delta_i \log p_i - \Delta_i \log p_{i+1} + O(\Delta_i^2)
\]

\[
= \Delta_i \log \frac{p_i}{p_{i+1}} + O(\Delta_i^2) > 0
\]

The first two terms cancel, the forth and the sixth are of order \(O(\Delta_i^2)\), and only the third and fifth term remain. Because \(p_i > p_{i+1}\), this difference between the entropies \(H(\tilde{p}) - H(p)\) is larger than 0. □

**Appendix C. Powering the distribution**

Assuming we have a probability mass function, as defined in Equation (B.1). We define a power \(\lambda, 0 < \lambda < 1\), and rescale the PMF:

\[
\tilde{p}_i = \frac{p_i^\lambda}{\sum_j p_j^\lambda}
\]

This new distribution is more smooth when

\[
\hat{\lambda}_i \leq \Delta_i
\]

where \(\Delta_i = p_i - p_{i+1}\). That would mean:

\[
\frac{p_i^\lambda - p_{i+1}^\lambda}{\sum_j p_j^\lambda} \leq p_i - p_{i+1}
\]

\[
0 < \lambda < 1
\]

\[
\sum_j p_j^\lambda \leq (\sum_j p_j^\lambda)(p_i - p_{i+1})
\]

\[
p_i^\lambda - p_{i+1}^\lambda \leq C(p_i - p_{i+1})
\]

This is actually the definition of Lipschitz continuity (Mohri et al., 2012). Unfortunately, for \(f(x) = x^\lambda\) where \(x \in [0, 1]\) and \(0 < \lambda < 1\) function \(f\) is not Lipschitz continuous, because for very small values of \(x\) the derivative goes to infinity.

If we now assume that \(\gamma < p_i < 1\), our purpose is to derive a lower bound \(\lambda\) for \(p_i\) such that (C.3) actually holds. First, we define the function \(f\):

\[
f(x) = x^\lambda, \quad x \in (0, 1), \quad \gamma < \lambda < 1
\]

\[
f'(x) = \lambda x^{\lambda-1} \quad \text{is always positive}
\]

\[
f''(x) = \lambda (\lambda-1) x^{\lambda-2} \quad \text{is always negative}
\]

in other words: the derivative is always positive, but each derivative becomes smaller and smaller. Because we have that \(x > \gamma\), and for \(h > 0\):

\[
f'(\gamma) > f'(x) = \lim_{h \downarrow 0} \frac{f(x+h) - f(x)}{(x+h) - x} > \frac{f(x+h) - f(x)}{(x+h) - x}
\]

\[
\text{Using } f''(x) = \lambda x^{\lambda-1}, \text{ and rewriting gives:}
\]

\[
f(x+h) - f(x) < \lambda y^{\lambda-1}(x+h) - x
\]

Substitution of \(x+h = p_i\) and \(x = p_{i+1}\) and solving \(\gamma\) reads:

\[
p_i^\lambda - p_{i+1}^\lambda < \lambda y^{\lambda-1}(p_i - p_{i+1})
\]

\[
\gamma\text{ gamma is lower bounded as such, powering the distribution acts as a smoother. □}
\]

**References**