Semi-Supervised Zero-Shot Classification with Label Representation Learning: Supplementary Material

Xin Li Yuhong Guo
Department of Computer and
Information Sciences, Temple University
Philadelphia, PA 19122, USA

Dale Schuurmans
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada

1. Proofs of Propositions

Proposition 1 *For any vector $\mathbf{z} \in \mathbb{R}^n$, we have that*

$$\max_i \mathbf{z}_i \;\leq\; \tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}) \;\leq\; \max_i \mathbf{z}_i + \tau \log n \quad (1)$$

for $\tau > 0$. The middle expression therefore provides a smooth approximation of the maximum function that becomes arbitrarily tight as $\tau \to 0$.

Proof: First, to prove the left inequality in (1), note that it is easy to verify the following

$$\max_i \mathbf{z}_i = \max_{\mathbf{p}\in\Delta} \mathbf{p}^\top \mathbf{z} \leq \max_{\mathbf{p}\in\Delta} \mathbf{p}^\top \mathbf{z} - \tau F^*(\mathbf{p}) \quad (2)$$

where $\Delta = \{\mathbf{p} \in \mathbb{R}^n : \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1\}$ is the probability simplex and $F^*(\mathbf{p}) = \mathbf{p}^\top \log(\mathbf{p})$ is the negative entropy function; the inequality in (2) follows simply because $\tau > 0$ and $F^*(\cdot)$ takes only non-positive values over its domain Δ. Next observe that the maximization problem on the right hand side of (2) corresponds to the definition of the Fenchel conjugate of $\tau F^*(\mathbf{p})$, which is given by $\tau F(\mathbf{z}/\tau)$ for the log-sum-exp function $F(\mathbf{z}/\tau)$ [1]; therefore we have

$$\max_{\mathbf{p}\in\Delta} \mathbf{p}^\top \mathbf{z} - \tau F^*(\mathbf{p}) \;= \tau F(\mathbf{z}/\tau) \;= \tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}), \quad (3)$$

establishing the left inequality in (1).

To prove the right inequality in (1), first note that

$$\tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}) = c + \tau \log(\sum_{i=1}^{n} e^{(\mathbf{z}_i - c)/\tau}) \quad (4)$$

for any $c \in \mathbb{R}$, via simple algebra, hence

$$\tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}) = \max_i \mathbf{z}_i + \tau \log(\sum_{i=1}^{n} e^{(\mathbf{z}_i - \max_i \mathbf{z}_i)/\tau}).$$

The inequality then follows because $\mathbf{z}_i - \max_i \mathbf{z}_i \leq 0$ for all i, hence $e^{(\mathbf{z}_i - \max_i \mathbf{z}_i)/\tau} \leq 1$ for all i as long as $\tau > 0$. $\square$

Proposition 2 *For any scalar $x \in \mathbb{R}$, we have the bounds $(x)_+ \leq \varphi_\tau(x) \leq (x)_+ + \frac{\tau}{4}$ for any $\tau > 0$, where*

$$\varphi_\tau(x) = \begin{cases} 0 & \text{if} \quad -\tau \geq x \\ \frac{(x+\tau)^2}{4\tau} & \text{if} \quad -\tau < x < \tau \\ x & \text{if} \quad x \geq \tau \end{cases} \quad (5)$$

Therefore $\varphi_\tau(\cdot)$ provides a smooth approximation of the capped-operator $(\cdot)_+$ that becomes tight as as $\tau \to 0$.

Proof: Recall that the capped-operator $(x)_+ = \max(0, x)$ by definition, and note that the following holds

$$\max(0, x) \;=\; \max_{0\leq p\leq 1} px \;\leq\; \max_{0\leq p\leq 1} px - \tau F^*(p) \quad (6)$$

for a convex regularizer

$$F^*(p) = p^2 - p = -p(1-p);$$

in particular, the inequality in (6) follows because $\tau > 0$ and $F^*(\cdot)$ only takes non-positive values on the domain $0 \leq p \leq 1$. Next observe that $px - \tau F^*(p) = px + \tau p(1-p)$ is a quadratic concave function of p, hence the maximizer, $\arg\max_{0\leq p\leq 1} px - \tau F^*(p)$, can be easily recovered as

$$p = \begin{cases} 0 & \text{if} \quad -\tau \geq x \\ \frac{(x+\tau)}{2\tau} & \text{if} \quad -\tau < x < \tau \\ 1 & \text{if} \quad x \geq \tau \end{cases} \quad (7)$$

Plugging this solution back to the maximization objective yields the $\varphi_\tau(\cdot)$ function defined in (5):

$$\max_{0\leq p\leq 1} px - \tau F^*(p) = \begin{cases} 0 & \text{if} \quad -\tau \geq x \\ \frac{(x+\tau)^2}{4\tau} & \text{if} \quad -\tau < x < \tau \\ x & \text{if} \quad x \geq \tau \end{cases} \quad (8)$$

Equations (6) and (8) and the definition (5) establish that $(x)_+ \leq \varphi_\tau(x)$ for all $x \in \mathbb{R}$ and $\tau > 0$.

To show that $\varphi_\tau(x) \leq (x)_+ + \frac{\tau}{4}$ for all x, note that $(x)_+ = \varphi_\tau(x)$ for $x \leq -\tau$ and $x \geq \tau$ so it remains only to show that the inequality holds for $-\tau < x < \tau$. In this interval, we have that $\varphi_\tau(x) = \frac{(x+\tau)^2}{4\tau}$, so we need to upper bound the gap $g(x) = \varphi_\tau(x) - (x)_+ = \frac{(x+\tau)^2}{4\tau} - (x)_+$. First consider the subinterval $-\tau < x \leq 0$, where the gap is given by $g(x) = \varphi_\tau(x) = \frac{(x+\tau)^2}{4\tau}$. On this subinterval we have $g'(x) > 0$ so the gap value is strictly increasing in x, hence its maximum value is obtained at the rightmost point $x = 0$, yielding $\max_{-\tau<x\leq 0} g(x) = g(0) = \frac{\tau}{4}$. Similarly, on the subinterval $0 \leq x < \tau$ the gap is given by $g(x) = \varphi_\tau(x) - x = \frac{(x-\tau)^2}{4\tau}$. On this subinterval we have $g'(x) < 0$ so the gap value is strictly decreasing in x, hence its maximum value is obtained at the leftmost point $x = 0$, which again yields $\max_{0\leq x<\tau} g(x) = g(0) = \frac{\tau}{4}$. Thus, $\varphi_\tau(x) - (x)_+ \leq \frac{\tau}{4}$ for all x. □

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, England, 2004. 1