Greedy Modality Selection via Approximate Submodular Maximization (Supplementary material)

Runxiang Cheng[*1] **Gargi Balasubramaniam**[*1] **Yifei He**[*1] **Yao-Hung Hubert Tsai**[2] **Han Zhao**[1]

[1]University of Illinois Urbana-Champaign, Illinois, USA
[2]Carnegie Mellon University, Pennsylvania, USA

1 PRELIMINARY FOR MISSING PROOFS

Proposition 1.1. *Let $X, Y \in \{0,1\}$ be random variables, $\mathcal{H}$ be the class of functions of X such that $\forall h \in \mathcal{H}, h(X) \in [0,1]$, and $\ell(\cdot,\cdot)$ be the cross-entropy loss. We have:*

$$\inf_{h\in\mathcal{H}} \mathbb{E}[\ell(Y, h(X))] = H(Y \mid X) \tag{1}$$

Proof. Let $x, \hat{y}$ be the instantiation of $X, \hat{Y}$ respectively, where $\hat{Y} := h(X)$. $\mathbb{1}(\cdot)$ denotes the indicator function, and $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ denotes the Kullback–Leibler divergence.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\ell(Y, h(X))] &= \mathbb{E}_{X,Y}[-\mathbb{1}(Y=1)\log \hat{Y} - \mathbb{1}(Y=0)\log(1-\hat{Y})] && (2)\\
&= -\mathbb{E}_X[\mathbb{E}_{Y|x}[\mathbb{1}(Y=1)\log \hat{y} + \mathbb{1}(Y=0)\log(1-\hat{y})]] && (3)\\
&= -\mathbb{E}_X[\Pr(Y=1\mid x)\log \hat{y} + \Pr(Y=0\mid x)\log(1-\hat{y})] && (4)\\
&= \mathbb{E}_X[\Pr(Y=1\mid x)\log\frac{1}{\hat{y}} + \Pr(Y=0\mid x)\log\frac{1}{1-\hat{y}}] && (5)\\
&= \mathbb{E}_X[\Pr(Y=1\mid x)\log\frac{\Pr(Y=1\mid x)}{\hat{y}} + \Pr(Y=0\mid x)\log\frac{\Pr(Y=0\mid x)}{1-\hat{y}}] && (6)\\
&\quad + \mathbb{E}_X[-\Pr(Y=1\mid x)\log\Pr(Y=1\mid x) - \Pr(Y=0\mid x)\log\Pr(Y=0\mid x)] && (7)\\
&= \mathbb{E}_X[D_{\mathrm{KL}}(\Pr(Y\mid x)\parallel h(x))] + \mathbb{E}_X[H(Y\mid x)] && (8)\\
&= D_{\mathrm{KL}}(\Pr(Y\mid X)\parallel h(X)) + H(Y\mid X) && (9)
\end{aligned}$$

Since $H(Y \mid X) \geq 0$ and is unrelated to $h(X)$, $\mathbb{E}_{\mathcal{D}}[\ell(Y, h(X))]$ is minimum when $h(X) = \Pr(Y \mid X)$. ■

2 MISSING PROOFS

Proposition 2.1. *Given $Y \in \{0,1\}$ and $\ell(Y,\hat{Y}) := \mathbb{1}(Y=1)\log\hat{Y} + \mathbb{1}(Y=0)\log(1-\hat{Y})$, $f_u(S) = I(S;Y)$.*

Proof. By Definition 3.1 and Proposition 1.1, we have:

$$\begin{aligned}
f_u(S) &= \inf_{h\in\mathcal{H}} \mathbb{E}[\ell(Y,c)] - \inf_{h\in\mathcal{H}} \mathbb{E}[\ell(Y,h(S))] && (10)\\
&= H(Y\mid c) - H(Y\mid S) && (11)\\
&= H(Y) - H(Y\mid S) && (12)\\
&= I(S;Y) && (13)
\end{aligned}$$

[*]Equal contribution.

Accepted for the 38[th] Conference on Uncertainty in Artificial Intelligence (UAI 2022).

■

Proposition 2.2. $\forall M \subseteq N \subseteq V$, $I(N;Y) - I(M;Y) = I(N \setminus M; Y \mid M) \geq 0$.

Proof. Let $N \coloneqq \{X_1, ..., X_n\}$, $M \coloneqq \{X_1, ..., X_m\}$, $n \geq m$.

$$I(N;Y) - I(M;Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, ..., X_1) - \sum_{i=1}^{m} I(X_i; Y \mid X_{i-1}, ..., X_1) \tag{14}$$

$$= \sum_{i=m+1}^{n} I(X_i; Y \mid X_{i-1}, ..., X_1) \tag{15}$$

$$= I(N \setminus M; Y \mid M) \tag{16}$$

$$\geq 0 \tag{17}$$

■

Proposition 2.3. *Under Assumption 2.1, $I(S;Y)$ is ϵ-approximately submodular, i.e., $\forall A \subseteq B \subseteq V$, $e \in V \setminus B$, $I(A \cup \{e\}; Y) - I(A;Y) + \epsilon \geq I(B \cup \{e\}; Y) - I(B;Y)$.*

Proof. For subset A, we have:

$$I(A \cup \{e\}; Y) - I(A;Y) = I(\{e\}; Y \mid A) \tag{18}$$

$$= I(\{e\}; Y, A) - I(\{e\}; A) \tag{19}$$

$$= I(\{e\}; Y) + I(\{e\}; A \mid Y) - I(\{e\}; A) \tag{20}$$

Similarly, $I(B \cup \{e\}; Y) - I(B;Y) = I(\{e\}; Y) + I(\{e\}; B \mid Y) - I(\{e\}; B)$. Given Assumption 2.1 holds, we denote $I(\{e\}; A \mid Y) = \epsilon_A$ and $I(\{e\}; B \mid Y) = \epsilon_B$ where $\epsilon_A, \epsilon_B \leq \epsilon$. In the worst case where $\epsilon_A = 0$, absolute submodularity is still satisfied if $\epsilon_B \leq I(\{e\}; B) - I(\{e\}; A)$, i.e.,

$$I(B \cup \{e\}; Y) - I(B;Y) = I(\{e\}; Y) + I(\{e\}; B \mid Y) - I(\{e\}; B) \tag{21}$$

$$= I(\{e\}; Y) - I(\{e\}; B) + \epsilon_B \tag{22}$$

$$\leq I(\{e\}; Y) - I(\{e\}; B) + I(\{e\}; B) - I(\{e\}; A) = I(A \cup \{e\}; Y) - I(A;Y) \tag{23}$$

But if $\epsilon_B > I(\{e\}; B) - I(\{e\}; A)$, the submodularity above will not hold. However, because $\epsilon_B \leq \epsilon$, we can define approximate submodularity characterized by the constant $\epsilon \geq 0$. Specifically:

$$I(B \cup \{e\}; Y) - I(B;Y) = I(\{e\}; Y) + I(\{e\}; B \mid Y) - I(\{e\}; B) \tag{24}$$

$$= I(\{e\}; Y) - I(\{e\}; B) + \epsilon_B \tag{25}$$

$$\leq I(\{e\}; Y) - I(\{e\}; B) + \epsilon \tag{26}$$

$$\leq I(\{e\}; Y) - I(\{e\}; A) + \epsilon \tag{27}$$

$$\leq I(\{e\}; Y) - I(\{e\}; A) + \epsilon_A + \epsilon \tag{28}$$

$$\leq I(A \cup \{e\}; Y) - I(A;Y) + \epsilon \tag{29}$$

■

Theorem 2.1. *Under Assumption 2.1, let $q \in \mathbb{Z}^+$, and S_p be the solution from Algorithm 1 at iteration p, we have:*

$$I(S_p; Y) \geq (1 - e^{-\frac{p}{q}}) \max_{S:|S| \leq q} I(S;Y) - q\epsilon \tag{30}$$

Proof. Let $S^* := \max_{S:|S|\le q} I(S;Y)$ be the optimal subset with cardinality at most q. By Proposition 3.2, $|S^*| = q$. We order S^* as $\{X_1^*,\ ...,\ X_q^*\}$. Then for all positive integer $i \le p$,

$$
\begin{aligned}
I(S^*;Y) &\le I(S^* \cup S_i;Y) && (31)\\
&= I(S_i;Y) + \sum_{j=1}^{q} I(X_j^*;Y \mid S_i \cup \{X_{j-1}^*, ..., X_1^*\}) && (32)\\
&= I(S_i;Y) + \sum_{j=1}^{q} (I(\{X_j^*, ..., X_1^*\} \cup S_i;Y) - I(\{X_{j-1}^*, ..., X_1^*\} \cup S_i;Y)) && (33)\\
&\le I(S_i;Y) + \sum_{j=1}^{q} (I(\{X_j^*\} \cup S_i;Y) - I(S_i;Y) + \epsilon) && (34)\\
&\le I(S_i;Y) + \sum_{j=1}^{q} (I(S_{i+1};Y) - I(S_i;Y) + \epsilon) && (35)\\
&\le I(S_i;Y) + q(I(S_{i+1}) - I(S_i;Y) + \epsilon) && (36)
\end{aligned}
$$

Eq. (31) is from Proposition 3.2, Eq. (32) and Eq. (33) are by the chain rule of mutual information, Eq. (34) is from Proposition 3.3, Eq. (35) is by the definition of Algorithm 1 that $I(S_{i+1};Y) - I(S_i;Y)$ is maximized in each iteration i. Let $\delta_i := I(S^*;Y) - I(S_i;Y)$, we can rewrite Eq. (36) into $\delta_i \le q(\delta_i - \delta_{i+1} + \epsilon)$, which can be rearranged into $\delta_{i+1} \le (1 - \frac{1}{q})\delta_i + \epsilon$.

Let $\delta_0 = I(S^*;Y) - I(S_0;Y)$. Since $S_0 = \emptyset$, we have $\delta_0 = I(S^*;Y)$. By the previous results, we can upper bound the quantity $\delta_p = I(S^*;Y) - I(S_p;Y)$ as follows:

$$
\begin{aligned}
\delta_p &\le (1 - \frac{1}{q})\delta_{p-1} + \epsilon && (37)\\
&\le (1 - \frac{1}{q})((1 - \frac{1}{q})\delta_{p-2} + \epsilon) + \epsilon && (38)\\
&\le (1 - \frac{1}{q})^p\delta_0 + (1 + (1 - \frac{1}{q}) + ... + (1 - \frac{1}{q})^{p-1})\epsilon && (39)\\
&= (1 - \frac{1}{q})^p\delta_0 + (\frac{1 - (1 - \frac{1}{q})^{p-1+1}}{1 - (1 - \frac{1}{q})})\epsilon && (40)\\
&= (1 - \frac{1}{q})^p\delta_0 + (q - q(1 - \frac{1}{q})^p)\epsilon && (41)\\
&\le (1 - \frac{1}{q})^p\delta_0 + q\epsilon && (42)\\
&\le e^{-\frac{p}{q}}\delta_0 + q\epsilon && (43)
\end{aligned}
$$

Eq. (39) to Eq. (41) is through the summation of the geometric series $1 + (1 - \frac{1}{q}) + ... + (1 - \frac{1}{q})^{p-1}$. Eq. (43) is by the inequality $1 - x \le e^{-x}$ for all $x \in \mathbb{R}$. Substitute the definitions of δ_p and δ_0 into Eq. (43) completes the proof. ■

Corollary 2.1. *Assume conditions in Theorem 3.1 hold, there exists optimal predictor $h^*(S_p) = \Pr(Y \mid S_p)$ such that*

$$
\begin{aligned}
\mathbb{E}[\ell_{01}(Y, h^*(S_p))] &\le \mathbb{E}[\ell_{ce}(Y, h^*(S_p))]\\
&\le H(Y) - (1 - e^{-\frac{p}{q}})I(S^*;Y) + q\epsilon && (44)
\end{aligned}
$$

Proof. Denote the quantity $(1 - e^{-\frac{p}{q}}) \max_{S:|S|\le q} I(S;Y) - q\epsilon$ from Theorem 3.1 as letter b. By the definition of mutual information, we have $H(Y \mid S_p) \le H(Y) - b$. Following Proposition 1.1, $\inf_{h:S_p\to[0,1]} \mathbb{E}[\ell_{ce}(Y, h(S_p))] \le H(Y) - b$. In other words, $\exists h^* = \Pr(Y \mid S_p)\ s.t.\ \mathbb{E}[\ell_{ce}(Y, h^*(S_p))] \le H(Y) - b$.

When the predictor is probabilistic (i.e., $h(X) = 0$ if and only if $h(X) \le 0.5$), $\ell_{01}(Y, \hat{Y}) = \mathbb{1}(Y \ne \hat{Y})$ naturally extends to $Y\mathbb{1}(\hat{Y} \le 0.5) + (1 - Y)\mathbb{1}(\hat{Y} > 0.5)$, which is upper bounded by $\ell_{ce}(Y, \hat{Y})$ for all $(Y, \hat{Y})$. Therefore, for the same h^* as above, we have:

$$\mathbb{E}[\ell_{01}(Y, h^*(S_p))] \le \mathbb{E}[\ell_{ce}(Y, h^*(S_p))] \le H(Y) - b \tag{45}$$

■

Corollary 2.2. *Assume conditions in Theorem 3.1 hold. There exists optimal predictors $h_1^* = \Pr(Y \mid S_p)$, $h_2^* = \Pr(Y \mid S^*)$ such that*

$$\begin{aligned}\mathbb{E}[\ell_{ce}(Y, h_1^*(S_p))] - \mathbb{E}[\ell_{ce}(Y, h_2^*(S^*))] \\ \le e^{-\frac{p}{q}} I(S^*; Y) + q\epsilon\end{aligned} \tag{46}$$

Proof. Following Theorem 3.1, and denote $\arg\max_{S:|S|\le q} I(S;Y)$ as S^*, we have:

$$I(S_p; Y) \ge (1 - e^{-\frac{p}{q}}) \max_{S:|S|\le q} I(S;Y) - q\epsilon \tag{47}$$

$$\implies H(Y) - H(Y \mid S_p) \ge (1 - e^{-\frac{p}{q}})(H(Y) - H(Y \mid S^*)) - q\epsilon \tag{48}$$

$$\implies H(Y \mid S_p) - H(Y \mid S^*) \le e^{-\frac{p}{q}}(H(Y) - H(Y \mid S^*)) + q\epsilon \tag{49}$$

$$\implies H(Y \mid S_p) - H(Y \mid S^*) \le e^{-\frac{p}{q}}(I(S^*; Y)) + q\epsilon \tag{50}$$

Using Proposition 1.1 completes the proof. ■

Proposition 2.4. *Under Assumption 2.1, $I(S;Y)$ is ϵ-approximately sub-additive for any $S \subseteq V$, i.e., $I(S \cup S'; Y) \le I(S;Y) + I(S';Y) + \epsilon$.*

Proof.

$$I(S \cup S'; Y) = I(S;Y) + I(S'; Y \mid S) \tag{51}$$

$$= I(S;Y) + I(S \cup Y; S') - I(S;S') \tag{52}$$

$$= I(S;Y) + I(S';Y) + I(S;S' \mid Y) - I(S;S') \tag{53}$$

$$\le I(S;Y) + I(S';Y) + \epsilon \tag{54}$$

Eq. (53) to Eq. (54) because $I(S;S' \mid Y) \le \epsilon$ by Assumption 2.1, and $I(S;S')$ is always non-negative. ■

Proposition 2.5. *Under Assumption 3.1, $I(S;Y)$ is ϵ-approximately super-additive for any $S \subseteq V$, i.e., $I(S \cup S'; Y) \ge I(S;Y) + I(S';Y) - \epsilon$.*

Proof. Similarly to the proof of Proposition 3.4, we have:

$$I(S \cup S'; Y) = I(S;Y) + I(S';Y) + I(S;S' \mid Y) - I(S;S') \tag{55}$$

$$\ge I(S;Y) + I(S';Y) - \epsilon \tag{56}$$

Eq. (55) to Eq. (56) because $I(S;S') \le \epsilon$ by Assumption 3.1, and $I(S;S' \mid Y)$ is non-negative. ■

Proposition 2.6. *If conditions in Proposition 3.4 and Proposition 3.5 hold, we have $I(X_i;Y) - \epsilon \le \phi_{I,X_i} \le I(X_i;Y) + \epsilon$ for any $X_i \in V$.*

Proof. By Proposition 3.4 and Proposition 3.5, for any $X_i \in V$ and $S \subseteq V$, we have:

$$I(X_i;Y) - \epsilon \le I(S \cup \{X_i\}; Y) - I(S;Y) \le I(X_i;Y) + \epsilon \tag{57}$$

Let's first apply the right inequality in Eq. (57) to Definition 2.2. Because $I(X_i;Y)+\epsilon$ is independent of S, we can simplify the calculation of the upper bound of ϕ_{I,X_i} as follows.

$$\phi_{I,X_i} = \sum_{S\subseteq V\setminus\{X_i\}} \frac{|S|!(|V|-|S|-1)!}{|V|!}(I(S\cup\{i\};Y)-I(S;Y)) \tag{58}$$

$$\leq \sum_{S\subseteq V\setminus\{i\}} \frac{|S|!(|V|-|S|-1)!}{|V|!}(I(X_i;Y)+\epsilon) \tag{59}$$

$$= \sum_{|S|=0}^{|V|-1} \binom{|V|-1}{|S|} \frac{|S|!(|V|-|S|-1)!}{|V|!}(I(X_i;Y)+\epsilon) \tag{60}$$

$$= \sum_{|S|=0}^{|V|-1} \frac{(|V|-1)!}{|S|(|F|-1-|S|)!} \frac{|S|!(|V|-|S|-1)!}{|V|!}(I(X_i;Y)+\epsilon) \tag{61}$$

$$= \sum_{|S|=0}^{|V|-1} \frac{1}{|V|}(I(X_i;Y)+\epsilon) \tag{62}$$

$$= I(X_i;Y)+\epsilon \tag{63}$$

Applying the same procedure to the left inequality in Eq. (57) to Definition 2.2, we have $\phi_{I,X_i} \geq I(X_i;Y)-\epsilon$. Combining both results completes the proof. ■

Proposition 2.7. *Under Assumption 2.1,* $\forall X_i \in V$, *we have* $I(X_i;Y) \leq \phi^{mci}_{I,X_i} \leq I(X_i;Y)+\epsilon$.

Proof. By Proposition 3.3, $I(\cdot;Y)$ would be approximately submodular under Assumption 2.1, thus:

$$I(X_i;Y)+\epsilon = I(\emptyset\cup X_i;Y) - I(\emptyset;Y)+\epsilon \tag{64}$$

$$\geq \max_{S\subseteq V} I(S\cup X_i;Y)-I(S;Y) = \phi^{mci}_{I,X_i} \tag{65}$$

On the other hand, if $\arg\max_{S\subseteq V} I(S\cup X_i;Y)-I(S;Y)=\emptyset$, we have $\phi^{mci}_{I,X_i} = I(\emptyset\cup X_i;Y)-I(\emptyset;Y) = I(X_i;Y)$. If $\arg\max_{S\subseteq V} I(S\cup X_i;Y)-I(S;Y)$ is some non-empty subset A, we have $\phi^{mci}_{I,X_i} = I(A\cup X_i;Y)-I(A;Y) \geq I(\emptyset\cup X_i;Y)-I(\emptyset;Y)$. In this case, $\phi^{mci}_{I,X_i} \geq I(X_i;Y)$. Combining both inequalities completes the proof. ■