A. Formal Version of Theorem 1

Theorem 2. In one epoch of Proc. 1, if the ToM model is $\epsilon$-optimal, i.e.

$$
L_{\text{pred}} = E_s KL[P_{\text{ToM}}(a \mid m, s; \theta) \parallel P_{i}(a \mid o, m)] < \epsilon
$$

where states $s = (i, k, D_{\text{supp}}, o, m, g)$ and instructions $m$ are sampled as Proc. 1, and for almost all states $s$ speaker gives a $\delta$-optimal instruction candidates pool $M$, i.e.

$$
\sum_{m \in M} P_{\text{ToM}}(a^g \mid m, s; \theta) \geq \delta
$$

then expected KL-divergence

$$
E_s KL[Q_{\text{ToM}}(m \mid s) \parallel Q(m \mid s; \theta)]
$$

between the instruction distribution calculated from ToM model

$$
Q_{\text{ToM}}(m \mid s; \theta) \overset{\Delta}{=} \sum_{m' \in M} P_{\text{ToM}}(a^g \mid m, s; \theta)\frac{P_{i}(a^g \mid o, m)}{\sum_{m'' \in M} P_{i}(a^g \mid o, m'''}
$$

and the target instruction distribution

$$
Q(m \mid s) \overset{\Delta}{=} \sum_{m'' \in M} P_{i}(a^g \mid o, m'')
$$

upper-bounded by

$$
N_M \sqrt{\frac{2\epsilon}{2(1-\sigma)}} + W_0(\epsilon)
$$

where $N_M$ is the size of largest pool of instruction candidates produced by the speaker, and $W_0$ is the principle branch of Lambert’s W function.

Proof. Applying Pinsker inequality,

$$
L_{\text{pred}} = E_s KL[P_{\text{ToM}}(a \mid m, s; \theta) \parallel P_{i}(a \mid o, m)]
\geq E_s, m, o, g P_i(a^g \mid o, m) - P_{\text{ToM}}(a^g \mid m, s; \theta)^2
\geq E_s, m, a P_i(a \mid o, m) - P_{\text{ToM}}(a \mid m, s; \theta)^2
\geq E_s, m, 2|P_i(a^g \mid o, m) - P_{\text{ToM}}(a^g \mid m, s; \theta)^2
\geq E_s, m, 2|\Delta(s, m)|^2
\geq 2(1 - \sigma)E_s E_{m \sim U(M)}|\Delta(s, m)|^2
$$

where $\Delta(s, m) = P_i(a^g \mid o, m) - P_{\text{ToM}}(a^g \mid m, s; \theta)$.

By processing the target expectation

$$
E_s KL[Q_{\text{ToM}}(m \mid s; \theta) \parallel Q(m \mid s)] = E_s \log \sum_{m' \in M} P_{i}(a^g \mid m', s; \theta)
\geq E_s \log \sum_{m' \in M} P_{\text{ToM}}(a^g \mid m', s; \theta)
\geq E_s \log \sum_{m' \in M} P_{\text{ToM}}(a^g \mid m', s; \theta)
\leq \frac{N_M}{\delta} E_s E_m \Delta(s, m')
\geq \frac{N_M}{\delta} E_s \sum_{m' \in M} W_0(KL[P_{\text{ToM}}(a \mid m, s; \theta) \parallel P_{i}(a \mid o, m)])
$$

$$
= \frac{N_M}{\delta} \sqrt{\frac{2\epsilon}{2(1-\sigma)}} + W_0(\epsilon)
$$

(20)

B. Training Time and space

All of our models can be trained on a 32 Gb V100. A model (speaker, listener, or ToM model) for referential game trains for about 20 hours, while a model (speaker, listener, or ToM model) for language navigation trains for 72 about hours. Tab. 1 and Fig. 3 reports the average of three runs, Fig. 2 reports data from 20 testing listeners.

C. Hyper-parameter Tuning

We only tuned the inner and outer learning rates of MAML among $10^i, i = -1, -2, -3, -4, -5$. A few influential hyperparameters are shown in Tab. 2. Other parameters are all kept same as previous work: Lowe et al. (2019a) for referential game, and Shridhar et al. (2021) for language navigation.