A Appendices

A.1 Basic facts about total variation distance

Over a countable space X and its discrete σ-algebra $\mathcal{P}(X)$, TVD is related to the ℓ_1 metric, $d(p,q) = \frac{1}{2}\sum_{x \in X} |p(x) - q(x)| = \frac{1}{2}||p-q||_1$. A few useful basic facts to recall are

- TVD is a metric, thus it obeys the triangle inequality.

- TVD is upper bounded by Kullback-Leibler (KL) divergence via Pinsker's inequality $d(p,q) \leq \sqrt{\frac{\ln 2}{2} D_{KL}(p||q)}$ where $D_{KL}(p||q) := \mathbb{E}_x \left[\log_2 \frac{p(x)}{q(x)}\right]$ is the KL divergence measured in bits.

- TVD is sub-additive over product measures $d(p_1q_1, p_2q_2) \leq d(p_1,p_1) + d(q_1,q_2)$, and relatedly, KL divergence is additive $D_{KL}(p_1q_1||p_2q_2) = D_{KL}(p_1||p_2) + D_{KL}(q_1||q_2)$.

As an alternative to the upper bound due to KL divergence, we can also bound TVD via its sub-additivity under product measures

$$d(\ell, \mathfrak{A}[\ell]) \leq \sum_{t=1}^{\infty} d\left(\mathbb{P}[\cdot|s_{<t};\ell], \mathbb{P}[\,\cdot\,|s_{<t};\mathfrak{A}[\ell]]\right).$$

In fact, this can cover more general cases, such as the analogous analysis of `FLC` (Yang et al., 2018) which zeros out everything except for the most likely tokens. We omit it due to page limit.

A.2 GPT-2 Language Model

The GPT-2 language model we used is a general purpose language model from OpenAI trained on WebText (Radford et al., 2019), which contains millions of web pages covering diverse topics. Citing concerns of malicious use, OpenAI only publicly released a small trained model with 117 million parameters. And that is the particular language model we use for empirical study in this work, `GPT-2-117M`.

We choose to use GPT-2 as the base language model in our work for several reasons. First, GPT-2 is trained on a large amount of data that we do not have access to. Second, it empirically achieves state-of-the-art performance across seven challenging semantics tasks, which includes question answering, reading comprehension, summarization and translation. Third, its architecture contains many late innovations such as transformer (Vaswani et al., 2017), instead of a recurrent neural network, and byte pair encoding for its vocabulary (Sennrich et al., 2016).

A.3 Derivation of Sec. 3.1

The effective LM is equal to

$$\mathbb{P}[s|h;\mathtt{Bins}[\ell]]$$

By definition of `Bins`

$$= \mathbb{E}_a\left[\mathbb{P}[s|s \in B_a, h;\ell]\right]$$

a are uniformly distributed with probability $1/2^k$

$$= \sum_a \frac{1}{2^k}\mathbb{P}[s|s \in B_a, h;\ell]$$

The bins are disjoint, s is only in B^s

$$= \frac{1}{2^k}\mathbb{P}[s|s \in B^s, h;\ell]$$

By definition of the marginal distribution

$$= \frac{1}{2^k}\frac{\mathbb{P}[s, s \in B^s|h;\ell]}{\sum_{s' \in B^s}\mathbb{P}[s', s' \in B^s|h;\ell]}$$

$$= \frac{1}{2^k}\frac{\mathbb{P}[s|h;\ell]}{\mathbb{P}[B^s|h;\ell]}. \tag{1}$$

The KL divergence follows as

$$D_{KL}(\mathbb{P}[\cdot|h;\ell]||\mathbb{P}[\cdot|h;\mathtt{Bins}[\ell]])$$

$$= \sum_s \mathbb{P}[s|h;\ell]\log_2\frac{\mathbb{P}[s|h;\ell]}{\mathbb{P}[s|h;\mathtt{Bins}[\ell]]}$$

Substituting in (1)

$$= \sum_s \mathbb{P}[s|h;\ell]\log_2\left(2^k\,\mathbb{P}[B^s|h;\ell]\right)$$

$$= k + \sum_{s \in \Sigma}\mathbb{P}[s|h;\ell]\log_2\mathbb{P}[B^s|h;\ell]$$

Σ is partitioned by $B = \{B_1, \cdots, B_{2^k}\}$

$$= k + \sum_a \sum_{s \in B_a}\mathbb{P}[s|h;\ell]\log_2\mathbb{P}[B_a|h;\ell]$$

$$= k + \sum_a \mathbb{P}[B_a|h;\ell]\log_2\mathbb{P}[B_a|h;\ell]$$

$$= k - H(B)$$

where $H(B)$ is the entropy of the partition $\{B_1, \cdots, B_{2^k}\}$.